

Ranking Categorical Features Using Generalization Properties*

Sivan Sabato

*IBM Haifa Research Lab
Haifa University Campus
Haifa 31905, Israel*

SIVANS@IL.IBM.COM

Shai Shalev-Shwartz

*Toyota Technological Institute at Chicago
University Press Building
1427 East 60th Street, Second Floor
Chicago, Illinois 60637, USA*

SHAI@TTI-C.ORG

Editor: Yoav Freund

Abstract

Feature ranking is a fundamental machine learning task with various applications, including feature selection and decision tree learning. We describe and analyze a new feature ranking method that supports categorical features with a large number of possible values. We show that existing ranking criteria rank a feature according to the *training* error of a predictor based on the feature. This approach can fail when ranking categorical features with many values. We propose the Ginger ranking criterion, that estimates the *generalization* error of the predictor associated with the Gini index. We show that for almost all training sets, the Ginger criterion produces an accurate estimation of the true generalization error, regardless of the number of values in a categorical feature. We also address the question of finding the optimal predictor that is based on a single categorical feature. It is shown that the predictor associated with the misclassification error criterion has the minimal expected generalization error. We bound the bias of this predictor with respect to the generalization error of the Bayes optimal predictor, and analyze its concentration properties. We demonstrate the efficiency of our approach for feature selection and for learning decision trees in a series of experiments with synthetic and natural data sets.

Keywords: feature ranking, categorical features, generalization bounds, Gini index, decision trees

1. Introduction

In this paper we address the problem of supervised feature ranking in the presence of categorical features. Feature ranking mechanisms have various applications; For instance, they can be used to define a filter for feature selection or as a splitting criterion for growing decision trees. In the feature ranking task we order a given set of features according to their relevance for predicting a target label. As in other supervised learning tasks, the ranking of the features is generated based on an input training set. Examples of widely used feature ranking criteria are the Gini index, the misclassification error, and Information Gain, also termed ‘cross-entropy’ (Hastie et al., 2001). The focus of this paper is feature ranking in the presence of *categorical* features. We show that a direct application of existing ranking criteria might lead to poor results in the presence of categorical

*. A preliminary version of this paper appeared at the 20th Annual Conference on Learning Theory under the title “Prediction by Categorical Features: Generalization Properties and Application to Feature Ranking”

features that can take many values. We propose an adaptation of existing ranking criteria that copes with these difficulties.

Many feature ranking methods are equivalent to the following two-phase process: First, each individual feature is used to construct a predictor of the label. Then, the features are ranked based on the errors of these predictors. Most current approaches use the same training set both for constructing the predictor and for evaluating its error. When dealing with binary features, the training error is likely to be close to the generalization error, and therefore the ranking generated by current methods works rather well. However, this is not the case when dealing with categorical features that can take a large number of values. To illustrate this fact, consider the problem of predicting whether someone is unemployed, based on their social security number (SSN). A predictor constructed using any finite training set would have zero error on the training set but a large generalization error. Therefore, a ranking criterion that supports categorical features should employ a more robust estimation of the generalization error.

The first contribution of this paper is an estimator for the generalization error of the predictor associated with the Gini index. This estimator can be calculated from the training set and we propose to use it instead of the original Gini index criterion in the presence of categorical features. We prove that regardless of the underlying distribution, our estimation is close to the true value of the generalization error for almost all training sets.

Based on our perspective of ranking criteria as estimators of the generalization error of a certain predictor, a natural question that arises is which predictor to use. Among all predictors that are based on a single feature, we ultimately would like to use the one whose generalization error is minimal. We prove that the best predictor in this sense is the predictor associated with the misclassification error criterion. We analyze the difference between the expected generalization error of this predictor and the error of the Bayes optimal hypothesis. Finally, we show a concentration result for the generalization error of this predictor.

Feature ranking criteria have been extensively studied in the context of decision trees (Mingers, 1989; Kearns and Mansour, 1996; Quinlan, 1993). The failure of existing feature ranking criteria in the presence of categorical features with a large number of possible values has been previously discussed in Quinlan (1993) and Mitchell (1997). Quinlan suggested the Information Gain Ratio as a correction to the Information Gain criterion. In a broader context, information-theoretic measures are commonly used for feature ranking (see for example Torkkola, 2006, and the references therein). One justification for their use is the existence of bounds on the Bayes optimal error that are based on these measures (Torkkola, 2006). However, obtaining estimators for the entropy or mutual information seems to be difficult in the general case (Antos and Kontoyiannis, 2001). Another ranking criterion designed to address the above difficulty is a distance-based measure introduced by de Mantaras (1991).

The problem we address shares some similarities with the problem of estimating the missing mass of a sample, typically encountered in language modeling (Good, 1953; McAllester and Schapire, 2000; Drukh and Mansour, 2005). The missing mass of a sample is the total probability mass of the values not occurring in the sample. Indeed, in the aforementioned example of the SSN feature, the value of the missing mass will be close to one. In some of our proofs we borrow ideas from McAllester and Schapire (2000) and Drukh and Mansour (2005). However, our problem is more involved, as even for a value that we do observe in the sample, if it appears only a small number of times then the training error is likely to diverge from the generalization error. Finally, we would like to note that classical VC theory for bounding the difference between the training error

and the generalization error is not applicable here. This is because the VC dimension grows with the number of values a categorical feature may take, and in our framework this number is unbounded.

This paper is organized as follows. In Sec. 2 we formally describe our problem setting. We introduce our main results in Sec. 3 and prove them in Sec. 4. We present experimental results in Sec. 5 and concluding remarks are given in Sec. 6.

2. Problem Setting

In this section we establish the notation used throughout the paper and formally describe our problem setting. In the supervised feature ranking setting we are provided with k categorical features and with a label. Each categorical feature is a random variable that takes values from a finite set. We denote a feature by X and the set of values X can take by V . We make no assumptions on the identity of V for each X nor on its size. The label is a binary random variable, denoted Y , that takes values from $\{0, 1\}$.

Generally speaking, the goal of supervised feature ranking is to rank the features based on their merit in constructing an accurate classification rule. The features are ranked according to their “relevance” to the label. Different criteria exist for assessing the relevance of a feature to the label. Since relevance is assessed for each feature separately, let us ignore the fact that we have k features and from now on focus on defining a relevance measure for a single feature X . We denote by V the set of values that X can take. To simplify our notation we denote

$$p_v \triangleq \Pr[X = v] \quad \text{and} \quad q_v \triangleq \Pr[Y = 1|X = v].$$

In practice, the probabilities $\{p_v\}$ and $\{q_v\}$ are unknown. Instead, it is assumed that we have a training set $S = \{(x_i, y_i)\}_{i=1}^m$, which is sampled i.i.d. according to the joint probability distribution $\Pr[X, Y]$. Based on S , the probabilities $\{p_v\}$ and $\{q_v\}$ are usually estimated as follows. Let $c_v = |\{i : x_i = v\}|$ be the number of examples in S for which the feature takes the value v and let $c_v^+ = |\{i : x_i = v \wedge y_i = 1\}|$ be the number of examples in which the value of the feature is v and the label is 1. Then $\{p_v\}$ and $\{q_v\}$ are estimated as follows:

$$\hat{p}_v \triangleq \frac{c_v}{m} \quad \text{and} \quad \hat{q}_v \triangleq \begin{cases} \frac{c_v^+}{c_v} & c_v > 0 \\ \frac{1}{2} & c_v = 0. \end{cases}$$

Note that \hat{p}_v and \hat{q}_v are implicit functions of the training set S .

Two popular relevance criteria (Hastie et al., 2001) are the misclassification error

$$\sum_{v \in V} \hat{p}_v \min\{\hat{q}_v, (1 - \hat{q}_v)\}, \quad (1)$$

and the Gini index

$$2 \sum_{v \in V} \hat{p}_v \hat{q}_v (1 - \hat{q}_v). \quad (2)$$

In these criteria, smaller values indicate more relevant features.

Both the misclassification error and the Gini index were found to work rather well in practice when $|V|$ is small. However, for categorical features with a large number of possible values, we might end up with a poor feature ranking criterion. As an example (see Mitchell, 1997), suppose that Y indicates whether a person is unemployed and we have two features: X_1 is the person’s SSN

and X_2 is 1 if the person has a mortgage and 0 otherwise. For the first feature, V is the set of all the SSNs. Because the SSN alone determines the target label, we have that \hat{q}_v is either 0 or 1 for any v such that $\hat{p}_v > 0$. Thus, both the misclassification error and the Gini index are zero for this feature. For the second feature, it can be shown that with high probability over the choice of the training set, the two criteria mentioned above take positive values. Therefore, both criteria prefer the first feature over the second. In contrast, for our purposes X_2 is much better than X_1 . This is because X_2 can be used later for learning a reasonable classification rule based on a finite training set, while X_1 will suffer from over-fitting.

It would have been natural to attribute the failure of the relevance criteria to the fact that we use estimated probabilities instead of the true (unknown) probabilities. However, note that in the above example, the same problem would arise even if we used $\{p_v\}$ and $\{q_v\}$ in Eq. (1) and Eq. (2). The aforementioned problem was previously underscored in the context of the Information Gain criterion (Quinlan, 1993; de Mantaras, 1991; Mitchell, 1997). In that context, Quinlan (1993) suggested an adaptation of the Information Gain, called Information Gain Ratio, which was found rather effective in practice.

In this paper we take a different approach, and propose to interpret a feature ranking criterion as the generalization error of a classification rule that can be inferred from the training set. To do so, let us first introduce some additional notation. A probabilistic hypothesis is a function $h : V \rightarrow [0, 1]$, where $h(v)$ is the probability to predict the label 1 given the value v . The generalization error of h is the probability to incorrectly predict the label,

$$\ell(h) \triangleq \sum_{v \in V} p_v (q_v (1 - h(v)) + (1 - q_v) h(v)) \quad . \quad (3)$$

We now define two hypotheses based on the training set S . The first one is

$$h_S^{\text{Gini}}(v) = \hat{q}_v \quad . \quad (4)$$

As its name indicates, h_S^{Gini} is closely related to the Gini index filter given in Eq. (2). To see this, we note that the generalization error of h_S^{Gini} is

$$\ell(h_S^{\text{Gini}}) = \sum_{v \in V} p_v (q_v (1 - \hat{q}_v) + (1 - q_v) \hat{q}_v) \quad .$$

If the estimated probabilities $\{\hat{p}_v\}$ and $\{\hat{q}_v\}$ coincide with the true probabilities $\{p_v\}$ and $\{q_v\}$, then $\ell(h_S^{\text{Gini}})$ is identical to the Gini index defined in Eq. (2). This will be approximately true, for example, when $m \gg |V|$. In other words, the Gini index is the training error of h_S^{Gini} . When the training set is small, using $\ell(h_S^{\text{Gini}})$ is preferable to using the Gini index given in Eq. (2), because $\ell(h_S^{\text{Gini}})$ takes into account the fact that the estimated probabilities might be skewed.

The second hypothesis we define is

$$h_S^{\text{Bayes}}(v) = \begin{cases} 1 & \hat{q}_v > \frac{1}{2} \\ 0 & \hat{q}_v < \frac{1}{2} \\ \frac{1}{2} & \hat{q}_v = \frac{1}{2} \end{cases} \quad . \quad (5)$$

Note that if $\{\hat{q}_v\}$ coincide with $\{q_v\}$ then h_S^{Bayes} is the Bayes optimal classifier, which we denote by h_∞^{Bayes} . If in addition $\{\hat{p}_v\}$ and $\{p_v\}$ are the same, then $\ell(h_S^{\text{Bayes}})$ is identical to the misclassification

error defined in Eq. (1). Here again, the misclassification error might differ from $\ell(h_S^{\text{Bayes}})$ for small training sets.

To illustrate the advantage of $\ell(h_S^{\text{Gini}})$ and $\ell(h_S^{\text{Bayes}})$ over their counterparts given in Eq. (2) and Eq. (1), we return to the example mentioned above. For X_1 , the SSN feature we have $\ell(h_S^{\text{Gini}}) = \ell(h_S^{\text{Bayes}}) = \frac{1}{2}M_0$, where $M_0 \triangleq \sum_{v:c_v=0} p_v$. In general, we denote

$$M_k \triangleq \sum_{v:c_v=k} p_v . \quad (6)$$

The quantity M_0 is known as the missing mass (Good, 1953; McAllester and Schapire, 2000) and for the SSN feature, $M_0 \geq (|V| - m)/|V|$. Therefore, the generalization error of both h_S^{Gini} and h_S^{Bayes} would be close to 1 for a reasonable m . On the other hand, for X_2 , the feature of having a mortgage, it can be verified that both $\ell(h_S^{\text{Bayes}})$ and $\ell(h_S^{\text{Gini}})$ are likely to be small. Therefore, using $\ell(h_S^{\text{Gini}})$ or $\ell(h_S^{\text{Bayes}})$ yields a correct ranking for this naive example.

We have proposed a modification of the Gini index and the misclassification error that uses the generalization error and therefore is suitable even when m is smaller than $|V|$. In practice, however, we cannot directly use the generalization error criterion since it depends on the unknown probabilities $\{p_v\}$ and $\{q_v\}$. To overcome this obstacle, we must derive estimators for the generalization error that can be calculated from the training set. In the next section we discuss the problem of estimating $\ell(h_S^{\text{Gini}})$ and $\ell(h_S^{\text{Bayes}})$ based on the training set. Additionally, we analyze the difference between $\ell(h_S^{\text{Bayes}})$ and the error of the Bayes optimal hypothesis.

3. Main Results

We start this section with a derivation of an estimator for $\ell(h_S^{\text{Gini}})$, which can serve as a new feature ranking criterion. We show that for most training sets, this estimator will be close to the true value of $\ell(h_S^{\text{Gini}})$. We then shift our attention to $\ell(h_S^{\text{Bayes}})$. First, we prove that among all predictors with no prior knowledge on the distribution $\Pr[X, Y]$, the generalization error of h_S^{Bayes} is smallest in expectation. Next, we bound the difference between the generalization error of h_S^{Bayes} and the error of the Bayes optimal hypothesis. Finally, we prove a concentration bound for $\ell(h_S^{\text{Bayes}})$. Regretfully, we could not find a good estimator for $\ell(h_S^{\text{Bayes}})$. Nevertheless, we believe that our concentration results can be used for finding such an estimator. This task is left for future research.

We propose the following estimator for the generalization error of h_S^{Gini} :

$$\hat{\ell} \triangleq \frac{|\{v : c_v = 1\}|}{2m} + \sum_{v:c_v>1} \frac{2c_v}{c_v - 1} \hat{p}_v \hat{q}_v (1 - \hat{q}_v) . \quad (7)$$

This estimator can be derived using a leave-one-out technique (see for instance Wasserman, 2004). In the next section we show a different derivation, based on a conditional cross-validation technique. We suggest to use the estimation of $\ell(h_S^{\text{Gini}})$ given in Eq. (7) rather than the original Gini index given in Eq. (2) as a feature ranking criterion. Let us compare these two criteria: First, for values v that appear many times in the training set we have that $\frac{c_v}{c_v - 1} \approx 1$. If for all $v \in V$ we have that the size of the training set is much larger than $1/p_v$, then all values in V are likely to appear many times in the training set and thus the definitions in Eq. (7) and Eq. (2) consolidate. The two definitions differ when there are values that appear rarely in the training set. For such values, the correction term is larger than 1. Special consideration is given to values that appear exactly once in the training

set. For such values we estimate the generalization error to be $\frac{1}{2}$, which is the highest possible error. Intuitively, since one example provides us with no information as to the variance of the label Y given $X = v$, we cannot have a more accurate estimation for the contribution of this value to the total generalization error. Furthermore, the fraction of values that appear exactly once in the training set is an estimator for the probability mass of those values that do not appear at all in the training set (see also Good, 1953; McAllester and Schapire, 2000).

We now turn to analyze the quality of the proposed estimator. We first show in Thm. 1 that the bias of this estimator is small. Then, in Thm. 2, we prove a concentration bound for the estimator, which holds for any joint distribution of $\Pr[X, Y]$ and does not depend on the size of V . Specifically, we show that for any $\delta \in (0, 1)$, in a fraction of at least $1 - \delta$ of the training sets the error of the estimator is $O\left(\frac{\ln(m/\delta)}{\sqrt{m}}\right)$.

Theorem 1 *Let S be a set of m examples sampled i.i.d. according to the probability measure $\Pr[X, Y]$. Let h_S^{Gini} be the Gini hypothesis given in Eq. (4) and let $\ell(h_S^{\text{Gini}})$ be the generalization error of h_S^{Gini} , where ℓ is as defined in Eq. (3). Let $\hat{\ell}$ be the estimation of $\ell(h_S^{\text{Gini}})$ as given in Eq. (7). Then, $|\mathbb{E}[\ell(h_S^{\text{Gini}})] - \mathbb{E}[\hat{\ell}]| \leq \frac{1}{2m}$, where expectation is taken over all samples S of m examples.*

The next theorem shows that for most training sets, our estimator is close to the true generalization error of h_S^{Gini} .

Theorem 2 *Under the same assumptions as in Thm. 1, let δ be an arbitrary scalar in $(0, 1)$. Then, with probability of at least $1 - \delta$ over the choice of S , we have*

$$|\ell(h_S^{\text{Gini}}) - \hat{\ell}| \leq O\left(\frac{\ln(m/\delta)\sqrt{\ln(1/\delta)}}{\sqrt{m}}\right).$$

Based on the above theorem, $\hat{\ell}$ can be used as a ranking criterion. The convergence rate shown can be used to establish confidence intervals on the true Gini generalization error. The proofs of Thm. 1 and Thm. 2 are given in the next section.

So far we have derived an estimator for the generalization error of the Gini hypothesis and shown that it is close to the true Gini error. The Gini hypothesis has the advantage of being highly concentrated around its mean. This is important especially when the sample size is fairly small. However, the Gini hypothesis does not produce the lowest generalization error in expectation. We now turn to show that the hypothesis h_S^{Bayes} defined in Eq. (5) is optimal in this respect, but that its concentration might be weaker. These two facts are characteristic of the well known bias-variance tradeoff commonly found in estimation and prediction tasks.

Had we known the underlying distribution of our data, we could have used the Bayes optimal hypothesis, h_∞^{Bayes} , that achieves the smallest possible generalization error. When the underlying distribution is unknown, the training set is used to construct the hypothesis. Thm. 3 below shows that among all hypotheses that can be learned from a finite training set, h_S^{Bayes} achieves the smallest generalization error in expectation. More precisely, h_S^{Bayes} is optimal among all the hypotheses that are symmetric with respect to both $|V|$ and the label values. Clearly, symmetric hypotheses cannot exploit prior knowledge on the underlying distribution $\Pr[X, Y]$. Formally, let \mathcal{F} be the set of all symmetric functions over $\mathbb{N} \times \mathbb{N}$, that is,

$$\mathcal{F} = \{f : \mathbb{N} \times \mathbb{N} \rightarrow [0, 1] \mid \forall n_1, n_2 \in \mathbb{N}, f(n_1, n_2) = 1 - f(n_1, n_1 - n_2)\}$$

and let H be the following set of mappings from samples of size m to hypotheses:

$$H = \left\{ h : (V \times \{0, 1\})^m \rightarrow V^{[0,1]} \mid \exists f \in \mathcal{F} \text{ s.t. } \forall S \in (V \times \{0, 1\})^m, \forall v \in V, \quad h[S](v) = f(c_v(S), c_v^+(S)) \right\}. \quad (8)$$

That is, H is the set of mappings that given a sample, generate a hypothesis based solely on the sample. Thus, hypotheses that rely on any prior knowledge on $\Pr[X, Y]$ are excluded.

The following theorem establishes the optimality of h_S^{Bayes} and bounds the difference between the Bayes optimal error and the error achieved by h_S^{Bayes} .

Theorem 3 *Let S be a set of m examples sampled i.i.d. according to the probability measure $\Pr[X, Y]$. For any hypothesis h , let $\ell(h)$ be the generalization error of h , as defined in Eq. (3). Let h_S^{Bayes} be the hypothesis given in Eq. (5), let h_∞^{Bayes} be the Bayes optimal hypothesis, and let H be the set of hypothesis mappings defined in Eq. (8). Then*

$$\mathbb{E}[\ell(h_S^{\text{Bayes}})] = \min_{h \in H} \mathbb{E}[\ell(h[S])], \quad (9)$$

and

$$\mathbb{E}[\ell(h_S^{\text{Bayes}})] - \ell(h_\infty^{\text{Bayes}}) \leq \frac{1}{2} \mathbb{E}[M_0] + \frac{1}{8} \mathbb{E}[M_1] + \frac{1}{8} \mathbb{E}[M_2] + \sum_{k=3}^m \frac{1}{\sqrt{ek}} \mathbb{E}[M_k], \quad (10)$$

where M_k is as defined in Eq. (6). Furthermore,

$$\lim_{m \rightarrow \infty} \left(\frac{1}{2} \mathbb{E}[M_0] + \frac{1}{8} \mathbb{E}[M_1] + \frac{1}{8} \mathbb{E}[M_2] + \sum_{k=3}^m \frac{1}{\sqrt{ek}} \mathbb{E}[M_k] \right) = 0. \quad (11)$$

Note that the first term in the difference between $\mathbb{E}[\ell(h_S^{\text{Bayes}})]$ and $\ell(h_\infty^{\text{Bayes}})$ is exactly half the expectation of the missing mass. This is expected, because we cannot improve our prediction over the baseline error of $\frac{1}{2}$ for values not seen in the training set, as exemplified in the SSN example described in the previous section. Subsequent terms in the bound can be attributed to the fact that even for values observed in the training set, a wrong prediction might be generated if there is a small number of examples.

We have shown that h_S^{Bayes} has the smallest generalization error in expectation, but this does not guarantee a small generalization error on a given sample. Thm. 4 below bounds the concentration of $\ell(h_S^{\text{Bayes}})$. This concentration along with Thm. 3 provides us with a bound on the difference between h_S^{Bayes} and the Bayes optimal error that is true for most samples.

Theorem 4 *Under the same assumptions of Thm. 3, assume that $m \geq 8$ and let δ be an arbitrary scalar in $(0, 1)$. Then, with probability of at least $1 - \delta$ over the choice of S , we have*

$$|\ell(h_S^{\text{Bayes}}) - \mathbb{E}[\ell(h_S^{\text{Bayes}})]| \leq O\left(\frac{\ln(m/\delta) \sqrt{\ln(1/\delta)}}{m^{1/6}}\right).$$

The concentration bound for $\ell(h_S^{\text{Bayes}})$ is weaker than the concentration bound for $\ell(h_S^{\text{Gini}})$, suggesting that indeed the choice between h_S^{Gini} and h_S^{Bayes} is not trivial. To use $\ell(h_S^{\text{Bayes}})$ as a ranking criterion, an estimator for this quantity is needed. However, at this point we cannot provide such an estimator. We conjecture that based on Thm. 4 an estimator with a small bias but a weak concentration can be constructed. We leave this task to further work. Finally, we would like to note that Antos et al. (1999) have shown that the Bayes optimal error cannot be estimated based on a finite training set. Finding an estimator for $\ell(h_S^{\text{Bayes}})$ would allow us to approximate the Bayes optimal error up to the bias term quantified in Thm. 3.

4. Proofs of Main Results

In this section we provide the full proofs of the theorems presented above.

4.1 Proof of Thm. 1

In the previous section, an estimator for the generalization error of the Gini hypothesis was presented. We stated that for most training sets this estimation is reliable. In this section, we first derive the estimator $\hat{\ell}$ given in Eq. (7) using a conditional cross-validation technique, and then use this interpretation of $\hat{\ell}$ to prove Thm. 1 and Thm. 2.

To derive the estimator given in Eq. (7), let us first rewrite $\ell(h_S^{\text{Gini}})$ as the sum $\sum_v \ell_v(h_S^{\text{Gini}})$, where $\ell_v(h_S^{\text{Gini}})$ is the amount of error due to value v and is formally defined as

$$\ell_v(h) \triangleq \Pr[X = v] \Pr[h(X) \neq Y \mid X = v] = p_v (q_v(1 - h(v)) + (1 - q_v)h(v)) .$$

We now estimate the two factors $\Pr[X = v]$ and $\Pr[h_S^{\text{Gini}}(X) \neq Y \mid X = v]$ independently. Later on we multiply the two estimations. The resulting local estimator of $\ell_v(h)$ is denoted $\hat{\ell}_v$ and our global estimator is $\hat{\ell} \triangleq \sum_v \hat{\ell}_v$.

To estimate $\Pr[X = v]$, we use the straightforward estimator \hat{p}_v . Turning to the estimation of $\Pr[h_S^{\text{Gini}}(X) \neq Y \mid X = v]$, recall that h_S^{Gini} , defined in Eq. (4), is a probabilistic hypothesis where \hat{q}_v is the probability to return the label 1 given that the value of X is v . Equivalently, we can think of the label that $h_S^{\text{Gini}}(v)$ returns as being generated based on the following process: Let $S(v)$ be the set of those indices in the training set in which the feature takes the value v , namely, $S(v) = \{i : x_i = v\}$. Then, to set the label $h_S^{\text{Gini}}(v)$ we randomly choose an index $i \in S(v)$ and return the label y_i . Based on this interpretation, a natural path for estimating $\Pr[h_S^{\text{Gini}}(X) \neq Y \mid X = v]$ is through cross-validation: Select an $i \in S(v)$ to determine $h_S^{\text{Gini}}(v)$, and estimate the generalization error to be the fraction of the examples whose label is different from the label of the selected example. That is, the estimation is $\frac{1}{c_v - 1} \sum_{j \in S(v): j \neq i} \mathbf{1}_{y_i \neq y_j}$. Obviously, this procedure cannot be used if $c_v = 1$. We handle this case separately later on. To reduce the variance of this estimation, this process can be repeated, selecting each single example from $S(v)$ in turn and validating each time using the rest of the examples in $S(v)$. It is then possible to average over all the choices of the examples. The resulting estimation therefore becomes

$$\sum_{i \in S(v)} \frac{1}{c_v} \left(\frac{1}{c_v - 1} \sum_{j \in S(v): j \neq i} \mathbf{1}_{y_i \neq y_j} \right) = \frac{1}{c_v(c_v - 1)} \sum_{i, j \in S(v): i \neq j} \mathbf{1}_{y_i \neq y_j} .$$

Thus, we estimate $\Pr[h_S^{\text{Gini}}(X) \neq Y \mid X = v]$ based on the fraction of differently-labeled pairs of examples in $S(v)$. Multiplying this estimator by \hat{p}_v we obtain the following estimator for $\ell_v(h_S^{\text{Gini}})$,

$$\begin{aligned} \hat{\ell}_v &= \hat{p}_v \frac{1}{c_v(c_v - 1)} \sum_{i, j \in S(v), i \neq j} \mathbf{1}_{y_i \neq y_j} \\ &= \hat{p}_v \frac{2c_v^+(c_v - c_v^+)}{c_v(c_v - 1)} = \hat{p}_v \frac{2c_v^2 \hat{q}_v(1 - \hat{q}_v)}{c_v(c_v - 1)} = \hat{p}_v \cdot \frac{2c_v}{c_v - 1} \hat{q}_v(1 - \hat{q}_v). \end{aligned} \tag{12}$$

Finally, for values v that appear only once in the training set, the above cross-validation procedure cannot be applied, and we therefore estimate their generalization error to be $\frac{1}{2}$, the highest possible

error. The full definition of $\hat{\ell}_v$ is thus:

$$\hat{\ell}_v = \begin{cases} \hat{p}_v \cdot \frac{1}{2} & c_v \leq 1 \\ \hat{p}_v \cdot \frac{2c_v}{c_v-1} \hat{q}_v (1 - \hat{q}_v) & c_v \geq 2. \end{cases} \quad (13)$$

The resulting estimator $\hat{\ell}$ defined in Eq. (7) is exactly the sum $\sum_v \hat{\ell}_v$.

Based on the above derivation of $\hat{\ell}_v$, we now turn to prove Thm. 1, in which it is shown that the expectations of our estimator and of the true generalization error of the Gini hypothesis are close. To do so, we first inspect each of these expectations separately, starting with $\mathbb{E}[\hat{\ell}_v]$. The following lemma calculates the expectation of $\hat{\ell}_v$ over those training sets with exactly k appearances of the value v .

Lemma 5 For k such that $1 < k \leq m$, $\mathbb{E}[\hat{\ell}_v | c_v(S) = k] = \frac{k}{m} \cdot 2q_v(1 - q_v)$.

Proof If $c_v = k$, then $\hat{p}_v = \frac{k}{m}$. Therefore, based on Eq. (12), we have

$$\mathbb{E}[\hat{\ell}_v | c_v(S) = k] = \frac{k}{m} \frac{1}{k(k-1)} \mathbb{E} \left[\sum_{i,j \in S(v), i \neq j} \mathbf{1}_{y_i \neq y_j} | c_v(S) = k \right]. \quad (14)$$

Let Z_1, \dots, Z_k be independent binary random variables with $\Pr[Z_i = 1] = q_v$ for all $i \in [k]$. The conditional expectation on the right-hand side of Eq. (14) equals to

$$\mathbb{E} \left[\sum_{i \neq j} \mathbf{1}_{Z_i \neq Z_j} \right] = \sum_{i \neq j} \mathbb{E}[\mathbf{1}_{Z_i \neq Z_j}] = \sum_{i \neq j} 2q_v(1 - q_v) = k(k-1) \cdot 2q_v(1 - q_v).$$

Combining the above with Eq. (14) concludes the proof. ■

Based on the above lemma, we are now ready to calculate $\mathbb{E}[\hat{\ell}_v]$. We have

$$\mathbb{E}[\hat{\ell}_v] = \sum_S \Pr[S] \mathbb{E}[\hat{\ell}_v] = \sum_{k=0}^m \sum_{S: c_v(S)=k} \Pr[S] \cdot \mathbb{E}[\hat{\ell}_v | c_v(S) = k]. \quad (15)$$

From the definition of $\hat{\ell}$, we have $\mathbb{E}[\hat{\ell}_v | c_v(S) = 1] = \frac{1}{2m}$ and $\mathbb{E}[\hat{\ell}_v | c_v(S) = 0] = 0$. Combining this with Lemma 5 and Eq. (15), we get

$$\begin{aligned} E[\hat{\ell}_v] &= \Pr[c_v = 1] \cdot \frac{1}{2m} + \sum_{k=2}^m \Pr[c_v = k] \cdot \frac{k}{m} \cdot 2q_v(1 - q_v) \\ &= \frac{1}{m} \left(\frac{1}{2} - 2q_v(1 - q_v) \right) \Pr[c_v = 1] + 2q_v(1 - q_v) \sum_{k=0}^m \Pr[c_v = k] \cdot \frac{k}{m} \\ &= \frac{1}{m} \left(\frac{1}{2} - 2q_v(1 - q_v) \right) \Pr[c_v = 1] + p_v \cdot 2q_v(1 - q_v), \end{aligned} \quad (16)$$

where the last equality follows from the fact that $\sum_{k=0}^m \Pr[c_v = k] \frac{k}{m} = \mathbb{E}[\hat{p}_v] = p_v$. Having calculated the expectation of $\hat{\ell}_v$ we now calculate the expectation of $\ell_v(h_S^{\text{Gini}})$.

Lemma 6 $\mathbb{E}[\ell_v(h_S^{\text{Gini}})] = p_v \left(\frac{1}{2} - 2q_v(1 - q_v) \right) \Pr[c_v = 0] + p_v \cdot 2q_v(1 - q_v)$.

Proof From the definition of $\ell_v(h_S^{\text{Gini}})$, we have that

$$\begin{aligned} \mathbb{E}[\ell_v(h_S^{\text{Gini}})] &= \mathbb{E}[p_v(q_v(1 - h_S^{\text{Gini}}(v)) + (1 - q_v)h_S^{\text{Gini}}(v))] \\ &= p_v(q_v(1 - \mathbb{E}[h_S^{\text{Gini}}(v)]) + (1 - q_v)\mathbb{E}[h_S^{\text{Gini}}(v)]) \\ &= p_v(q_v + (1 - 2q_v)\mathbb{E}[h_S^{\text{Gini}}(v)]) . \end{aligned} \tag{17}$$

Next, we calculate $\mathbb{E}[h_S^{\text{Gini}}(v)]$ as follows

$$\begin{aligned} \mathbb{E}[h_S^{\text{Gini}}(v)] &= \sum_S \Pr[S]h_S^{\text{Gini}}(v) \\ &= \Pr[c_v(S) = 0] \cdot \frac{1}{2} + \sum_{k=1}^m \sum_{i=0}^k \Pr[c_v(S) = k \text{ and } c_v^+(S) = i] \frac{i}{k} \\ &= \Pr[c_v(S) = 0] \cdot \frac{1}{2} + \sum_{k=1}^m \Pr[c_v(S) = k] \sum_{i=0}^k \Pr[c_v^+(S) = i \mid c_v(S) = k] \frac{i}{k} \\ &= \Pr[c_v(S) = 0] \cdot \frac{1}{2} + \sum_{k=1}^m \Pr[c_v(S) = k] \cdot q_v \\ &= \Pr[c_v(S) = 0] \cdot \frac{1}{2} + \Pr[c_v(S) > 0] \cdot q_v \\ &= q_v + \frac{1}{2}(1 - 2q_v)\Pr[c_v(S) = 0] . \end{aligned} \tag{18}$$

Plugging Eq. (18) into Eq. (17) and rearranging terms we conclude our proof. ■

Equipped with the expectation of $\hat{\ell}_v$ given in Eq. (16) and the expectation of $\ell_v(h_S^{\text{Gini}})$ given in Lemma 6, we are now ready to prove Thm. 1.

Proof [of Thm. 1] Using the definitions of $\ell(h_S^{\text{Gini}})$ and $\hat{\ell}$ we have that

$$\mathbb{E}[\hat{\ell}] - \mathbb{E}[\ell(h_S^{\text{Gini}})] = \mathbb{E}[\sum_v \hat{\ell}_v] - \mathbb{E}[\sum_v \ell_v(h_S^{\text{Gini}})] = \sum_v (\mathbb{E}[\hat{\ell}_v] - \mathbb{E}[\ell_v(h_S^{\text{Gini}})]) . \tag{19}$$

Fix some $v \in V$. From Eq. (16) and Lemma 6 we have

$$\mathbb{E}[\hat{\ell}_v] - \mathbb{E}[\ell_v(h_S^{\text{Gini}})] = \left(\frac{1}{2} - 2q_v(1 - q_v)\right) \left(\frac{1}{m} \Pr[c_v = 1] - p_v \Pr[c_v = 0]\right) . \tag{20}$$

Also, it is easy to see that

$$\begin{aligned} \frac{1}{m} \Pr[c_v = 1] - p_v \Pr[c_v = 0] &= p_v(1 - p_v)^{m-1} - p_v(1 - p_v)^m \\ &= p_v^2(1 - p_v)^{m-1} = \frac{p_v}{m} \Pr[c_v = 1] . \end{aligned}$$

Plugging this into Eq. (20) we obtain:

$$\mathbb{E}[\hat{\ell}_v] - \mathbb{E}[\ell_v(h_S^{\text{Gini}})] = \left(\frac{1}{2} - 2q_v(1 - q_v)\right) \frac{1}{m} p_v \Pr[c_v = 1] .$$

For any q_v we have that $0 \leq 2q_v(1 - q_v) \leq \frac{1}{2}$, which implies the following inequality:

$$0 \leq \mathbb{E}[\hat{\ell}_v] - \mathbb{E}[\ell_v(h_S^{\text{Gini}})] \leq \frac{1}{2m} p_v \Pr[c_v = 1] \leq \frac{p_v}{2m} .$$

Summing this over v and using Eq. (19) we conclude that

$$0 \leq \mathbb{E}[\hat{\ell}] - \mathbb{E}[\ell(h_S^{\text{Gini}})] \leq \sum_v \frac{p_v}{2m} = \frac{1}{2m} .$$

■

4.2 Proof of Thm. 2

We now turn to prove Thm. 2 in which we argue that with high confidence on the choice of S , the value of our estimator is close to the actual generalization error of h_S^{Gini} . To do this, we show that both our estimator and the true generalization error of h_S^{Gini} are concentrated around their mean. The proof of Thm. 2 will then follow from Thm. 1.

We start by showing that our estimator $\hat{\ell}$ is concentrated around its expectation. The concentration of $\hat{\ell}$ follows relatively easily by application of McDiarmid's Theorem (McDiarmid, 1989):

Theorem 7 (McDiarmid) *Let X_1, \dots, X_m be independent random variables taking values in a set V and let $f : V^m \rightarrow \mathbb{R}$ be such that for every $1 \leq i \leq m$*

$$\sup |f(x_1, \dots, x_m) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_m)| \leq c_i$$

where the supremum is taken over all $x_1, \dots, x_m, x'_i \in V$. Then with probability at least $1 - \delta$

$$f(X_1, \dots, X_m) \leq \mathbb{E}[f(X_1, \dots, X_m)] + \sqrt{\frac{1}{2} \ln\left(\frac{1}{\delta}\right) \sum_{i=1}^m c_i}$$

and with probability at least $1 - \delta$

$$f(X_1, \dots, X_m) \geq \mathbb{E}[f(X_1, \dots, X_m)] - \sqrt{\frac{1}{2} \ln\left(\frac{1}{\delta}\right) \sum_{i=1}^m c_i} .$$

To simplify our notation, we will henceforth use the shorthand $\forall^\delta S \quad \pi[S, \delta]$ to indicate that the predicate $\pi[S, \delta]$ holds with probability of at least $1 - \delta$ over the choice of S .

Lemma 8 *Let $\delta \in (0, 1)$. Then, $\forall^\delta S \quad |\hat{\ell} - \mathbb{E}[\hat{\ell}]| \leq 12\sqrt{\frac{\ln(\frac{2}{\delta})}{2m}}$.*

Proof We prove the lemma using McDiarmid's theorem. To do so, we need to show that $\hat{\ell}$ has the bounded differences property; namely, we shall find an upper bound for the effect of any change of a single example in S on $\hat{\ell}$. Changing example (x_i, y_i) in S to (x'_i, y'_i) is tantamount to first removing (x_i, y_i) and then adding (x'_i, y'_i) . Since the effect of adding is simply the opposite of the effect of removing, it is sufficient to find an upper bound for the effect a single removal of example can have. Then the effect of a change on the sample would be no larger than twice the effect of the removal.

Let $S^{\setminus i}$ denote the set $S \setminus \{(x_i, y_i)\}$. We therefore need to bound $|\hat{\ell}(S) - \hat{\ell}(S^{\setminus i})|$. Assume, without loss of generality, that $x_i = v$ and $y_i = 0$. Then, using the definition of $\hat{\ell}_v$ we have that

$$|\hat{\ell}(S) - \hat{\ell}(S^{\setminus i})| = |\hat{\ell}_v(S) - \hat{\ell}_v(S^{\setminus i})| .$$

Based on the definitions of $\hat{p}_v = c_v/m$ and $\hat{q}_v = c_v^+/c_v$, we can rewrite Eq. (13) as

$$\hat{\ell}_v(S) = \begin{cases} \frac{1}{2m} & c_v = 1 \\ \frac{2c_v^+(c_v - c_v^+)}{m(c_v - 1)} & c_v \geq 2. \end{cases}$$

Therefore, if $c_v \geq 3$,

$$\begin{aligned} |\hat{\ell}_v(S) - \hat{\ell}_v(S^{\setminus i})| &= \frac{2c_v^+}{m} \left(\frac{c_v - c_v^+}{c_v - 1} - \frac{c_v - c_v^+ - 1}{c_v - 2} \right) = \frac{2c_v^+(c_v^+ - 1)}{m(c_v - 1)(c_v - 2)} \\ &\leq \frac{2c_v(c_v - 1)}{m(c_v - 1)(c_v - 2)} = \frac{2c_v}{m(c_v - 2)} \leq \frac{6}{m}, \end{aligned}$$

while if $c_v = 2$ then

$$|\hat{\ell}_v(S) - \hat{\ell}_v(S^{\setminus i})| = \frac{2c_v^+(2 - c_v^+)}{m} - \frac{1}{2m} \leq \frac{2}{m}.$$

Lastly, if $c_v = 1$ then $|\hat{\ell}_v(S) - \hat{\ell}_v(S^{\setminus i})| = \frac{1}{2m}$. Therefore for any sample S

$$|\hat{\ell}_v(S) - \hat{\ell}_v(S^{\setminus i})| \leq \frac{6}{m},$$

and thus the effect of a single change in S is no larger than $\frac{12}{m}$. We can now apply McDiarmid's theorem to get that with probability of at least $1 - \delta$:

$$|\hat{\ell} - \mathbb{E}[\hat{\ell}]| \leq \sqrt{\frac{1}{2} \ln\left(\frac{2}{\delta}\right) m \left(\frac{12}{m}\right)^2} = 12\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2m}}.$$

■

We now turn to show a concentration bound on the true generalization error $\ell(h_S^{\text{Gini}})$. Here we cannot directly use McDiarmid's Theorem since the bounded differences property does not hold for $\ell(h_S^{\text{Gini}})$. To see this, suppose that $V = \{0, 1\}$, $p_0 = p_1 = \frac{1}{2}$, $q_0 = 0.99$ and $q_1 = 1$. Assume in addition that $|S(0)| = 1$; namely, there is only a single example in S for which the feature takes the value 0, an unlikely but possible scenario. In this case, if the single example in $S(0)$ is labeled 1, then $\ell(h_S^{\text{Gini}}) = 0.01$, but if this example is labeled 0, then $\ell(h_S^{\text{Gini}}) = 0.99$. That is, a change of a single example might have a dramatic effect on $\ell(h_S^{\text{Gini}})$. This problem can intuitively be attributed to the fact that S is an atypical sample of the underlying distribution $\{p_v\}$. To circumvent this obstacle, we use the following lemma. Note that a similar result can be derived from the results in Kutin (2002), albeit with much larger constants. The lemma below provides tighter bounds for a more restricted case.

Lemma 9 *Let S be a sample with m examples drawn i.i.d from the distribution $\text{Pr}[X, Y]$. Let δ be a confidence parameter. For two samples S_1 and S_2 with m examples, we say that $d(S_1, S_2) \leq 1$ if there is at most one example that is different between the two samples. Let f be a real function of the sample. If there exists a function of the sample g and real numbers c, b such that the following*

conditions hold:

$$\forall S_1, S_2 \text{ s.t. } d(S_1, S_2) \leq 1 \quad |g(S_1) - g(S_2)| \leq \frac{c}{m} \quad (21)$$

$$\forall^\delta S \quad f(S) = g(S) \quad (22)$$

$$|\mathbb{E}[f(S)] - \mathbb{E}[g(S)]| \leq \frac{b}{\sqrt{m}} \quad , \quad (23)$$

then

$$\forall^{2\delta} S \quad |f(S) - \mathbb{E}[f(S)]| \leq \frac{c\sqrt{\ln(\frac{2}{\delta})} + b\sqrt{2}}{\sqrt{2m}} \quad .$$

Proof From Eq. (21) and McDiarmid's theorem we have

$$\forall^\delta S \quad |g(S) - \mathbb{E}[g(S)]| \leq \frac{c\sqrt{\ln(\frac{2}{\delta})}}{\sqrt{2m}} \quad .$$

In addition,

$$|f(S) - \mathbb{E}[f(S)]| \leq |f(S) - g(S)| + |g(S) - \mathbb{E}[g(S)]| + |\mathbb{E}[f(S)] - \mathbb{E}[g(S)]| \quad .$$

Therefore, using Eq. (22) and Eq. (23) and applying a union bound, we have

$$\forall^{2\delta} S |f(S) - \mathbb{E}[f(S)]| \leq 0 + \frac{c\sqrt{\ln(\frac{2}{\delta})}}{\sqrt{2m}} + \frac{b}{\sqrt{m}} = \frac{c\sqrt{\ln(\frac{2}{\delta})} + b\sqrt{2}}{\sqrt{2m}} \quad .$$

■

To use Lemma 9 we define a new hypothesis h_S^δ that depends both on the sample S and on the desired confidence parameter δ . This hypothesis would 'compensate' for atypical samples. We let $f \triangleq \ell(h_S^{\text{gini}})$ and $g \triangleq \ell(h_S^\delta)$, and show that the conditions of the lemma hold.

We construct a hypothesis h_S^δ such that g satisfies the three requirements given in Eqs. (21-23) based on Lemma 10 below. This lemma states that except for values with small probabilities, we can assure that with high confidence, $c_v(S)$ grows with p_v . This means that as long as p_v is not too small, a change of a single example in $c_v(S)$ does not change $h_S^\delta(v)$ too much. On the other hand, if p_v is small then the value v has little effect on the error to begin with. Therefore, regardless of the probability p_v , the error $\ell(h_S^\delta)$ cannot be changed too much by a single change of example in S . This would allow us to prove a concentration bound on $\ell(h_S^\delta)$ using McDiarmid's theorem. Let us first introduce a new notation. Given a confidence parameter $\delta > 0$, a probability $p \in [0, 1]$, and a sample size m , we define

$$\rho(\delta, p, m) \triangleq mp - \sqrt{mp \cdot 3 \ln(2/\delta)} \quad .$$

Lemma 10 below states that $c_v(S)$ is likely to be at least $\rho(\delta/m, p_v, m)$ for all values with non-negligible probabilities.

Lemma 10 *Let $\delta \in (0, 1)$ be a confidence parameter. Then,*

$$\forall^\delta S \quad \forall v \in V : p_v \geq \frac{6 \ln(\frac{2m}{\delta})}{m} \quad \Rightarrow \quad c_v(S) \geq \rho(\delta/m, p_v, m) > 1 \quad .$$

Proof The proof is based on lemma 44 from Drukh and Mansour (2005). This lemma states that for all $v \in V$ such that $p_v \geq \frac{3 \ln(\frac{2}{\delta})}{m}$ we have that

$$\forall^\delta S \quad |p_v - \hat{p}_v| \leq \sqrt{\frac{p_v \cdot 3 \ln(\frac{2}{\delta})}{m}}. \quad (24)$$

Based on this lemma, we immediately get that for all v such that $p_v \geq 3 \ln(\frac{2}{\delta})/m$,

$$\forall^\delta S \quad c_v \geq \rho(\delta, p_v, m).$$

Note, however, that this bound is trivial for $p_v = 3 \ln(\frac{2}{\delta})/m$, because in this case $\rho(\delta, p_v, m) = 0$. We therefore use the bound for values in which $p_v \geq 6 \ln(\frac{2}{\delta})/m$. For these values it is easy to show that $\rho(\delta, p_v, m) > 1$ for any $\delta \in (0, 1)$. Trivially, there are at most m values v for which $p_v \geq \frac{6 \ln(2/\delta)}{m}$. Therefore, by substituting δ/m for δ and applying a union bound, the proof is concluded. \blacksquare

Based on the bound given in the above lemma, we define h_S^δ to be

$$h_S^\delta(v) \triangleq \begin{cases} h_S^{\text{Gini}}(v) & p_v < \frac{6 \ln(\frac{2m}{\delta})}{m} \text{ or } c_v \geq \rho(\frac{\delta}{m}, p_v, m) \\ \frac{c_v^+ + q_v(\lceil \rho(\frac{\delta}{m}, p_v, m) \rceil - c_v)}{\lceil \rho(\frac{\delta}{m}, p_v, m) \rceil} & \text{otherwise.} \end{cases}$$

That is, $h_S^\delta(v)$ is equal to $h_S^{\text{Gini}}(v)$ if either p_v is negligible or if there are enough representatives of v in the sample. If this is not the case, then S is not a typical sample and thus we “force” it to be typical by adding $\lceil \rho(\frac{\delta}{m}, p_v, m) \rceil - c_v$ ‘pseudo-examples’ to S with the value v and with labels that are distributed according to q_v . Therefore, except for values with negligible probability p_v , the hypothesis $h_S^\delta(v)$ is determined by at least $\lceil \rho(\frac{\delta}{m}, p_v, m) \rceil$ ‘examples’. As a direct result of this construction we obtain that a single example from S has a small effect on the value of $\ell(h_S^\delta)$.

We can now show that each of the properties in (21-23) hold. From the definition of h_S^δ and Lemma 10 it is clear that Eq. (22) holds. Let us now show that Eq. (23) holds, with b .

Lemma 11 $|\mathbb{E}[\ell(h_S^{\text{Gini}})] - \mathbb{E}[\ell(h_S^\delta)]| \leq \frac{1}{m}$.

Proof We have

$$\mathbb{E}[\ell(h_S^{\text{Gini}})] - \mathbb{E}[\ell(h_S^\delta)] = \sum_v \left(\mathbb{E}[\ell_v(h_S^{\text{Gini}}) - \ell_v(h_S^\delta)] \right). \quad (25)$$

We bound $\mathbb{E}[\ell_v(h_S^{\text{Gini}}) - \ell_v(h_S^\delta)]$ as follows. First, for values v such that $p_v < 6 \ln(\frac{2m}{\delta})/m$, we have that $h_S^{\text{Gini}}(v) = h_S^\delta(v)$. Thus $\mathbb{E}[\ell_v(h_S^{\text{Gini}}) - \ell_v(h_S^\delta)] = 0$. For the rest of the values, $p_v \geq 6 \ln(\frac{2m}{\delta})/m$ and thus the definition of $\ell_v(h_S^\delta)$ implies

$$\begin{aligned} \mathbb{E}[\ell_v(h_S^{\text{Gini}}) - \ell_v(h_S^\delta)] &= \\ \Pr[c_v < \rho(\delta/m, p_v, m)] \cdot \mathbb{E}[\ell_v(h_S^{\text{Gini}}) - \ell_v(h_S^\delta) \mid c_v < \rho(\delta/m, p_v, m)] &. \end{aligned} \quad (26)$$

Using Eq. (24) again, we obtain that $\Pr[c_v < \rho(\delta/m, p_v, m)] \leq \delta/m$. In addition, since both $\ell_v(h_S^{\text{Gini}})$ and $\ell_v(h_S^\delta)$ are in $[0, p_v]$ we have that

$$\left| \mathbb{E}[\ell_v(h_S^{\text{Gini}}) - \ell_v(h_S^\delta) \mid c_v < \rho(\delta/m, p_v, m)] \right| \leq p_v.$$

Combining the above two facts with Eq. (26) we get

$$\left| \mathbb{E}[\ell_v(h_S^{\text{Gini}}) - \ell_v(h_S^\delta)] \right| \leq \frac{P_v \delta}{m} \leq \frac{P_v}{m}.$$

Summing the above over v and using Eq. (25) we conclude that,

$$\left| \mathbb{E}[\ell(h_S^{\text{Gini}}) - \ell(h_S^\delta)] \right| \leq \sum_v \frac{P_v}{m} = \frac{1}{m}.$$

■

Finally, the following lemma shows that Eq. (21) also holds.

Lemma 12 For any $\delta > 0$, and for any two samples S_1 and S_2 with m examples such that $d(S_1, S_2) \leq 1$ with d defined as in Lemma 9,

$$\left| \ell(h_{S_1}^\delta) - \ell(h_{S_2}^\delta) \right| \leq \frac{12 \ln(\frac{2m}{\delta})}{m}.$$

The proof of this lemma is deferred to the appendix.

We have shown that the functions $g \triangleq \ell(h_S^\delta)$ and $f \triangleq \ell(h_S^{\text{Gini}})$ satisfy the three requirements given in Eqs. (21-23) and therefore Lemma 9 can be used to show that $\ell(h^{\text{Gini}})$ is concentrated.

Lemma 13 $\forall \delta > 0 \quad \forall^\delta S \quad \left| \ell(h_S^{\text{Gini}}) - \mathbb{E}[\ell(h_S^{\text{Gini}})] \right| \leq \frac{12 \ln(\frac{4m}{\delta}) \sqrt{\ln(\frac{4}{\delta})}}{\sqrt{2m}} + \frac{1}{m}.$

Proof In Lemma 9, let $f \triangleq \ell(h_S^{\text{Gini}})$ and let $g \triangleq \ell(h_S^\delta)$. Let $c \triangleq 12 \ln(\frac{2m}{\delta})$, and let $b \triangleq \frac{1}{\sqrt{m}}$. By Lemma 10, Eq. (22) holds. By Lemma 12, Eq. (21) holds, and by Lemma 11, Eq. (23) holds. Therefore, from Lemma 9 we have

$$\forall \delta > 0 \quad \forall^{2\delta} S \quad |f(S) - \mathbb{E}[f(S)]| \leq \frac{12 \ln(\frac{2m}{\delta}) \sqrt{\ln(\frac{2}{\delta})}}{\sqrt{2m}} + \frac{1}{m}.$$

The proof is concluded by substituting $\frac{\delta}{2}$ for δ . ■

Thm. 2 states that with high confidence, the estimator $\hat{\ell}$ is close to the true generalization error of the Gini hypothesis, $\ell(h_S^{\text{Gini}})$. We conclude the analysis of the Gini estimator by proving this theorem.

Proof [of Thm. 2] Substituting $\frac{\delta}{2}$ for δ and applying a union bound, we have that all three properties stated in Lemma 13, Thm. 1 and Lemma 8 hold with probability of at least $1 - \delta$. We therefore conclude that with probability of at least $1 - \delta$,

$$\begin{aligned} \left| \ell(h_S^{\text{Gini}}) - \hat{\ell} \right| &\leq \left| \ell(h_S^{\text{Gini}}) - \mathbb{E}[\ell(h_S^{\text{Gini}})] \right| + \left| \mathbb{E}[\ell(h_S^{\text{Gini}})] - \mathbb{E}[\hat{\ell}] \right| + \left| \mathbb{E}[\hat{\ell}] - \hat{\ell} \right| \\ &\leq \frac{2}{m} + \frac{12 \ln(\frac{8m}{\delta}) \sqrt{\ln(\frac{8}{\delta})}}{\sqrt{2m}} + 12 \sqrt{\frac{\ln(\frac{4}{\delta})}{2m}} = O\left(\frac{\ln(\frac{m}{\delta}) \sqrt{\ln(\frac{1}{\delta})}}{\sqrt{m}}\right). \end{aligned}$$

■

4.3 Proof of Thm. 3

Throughout this section we use the notation $S^{(m)}$ to denote a random training set of m examples. Before proving Thm. 3, we provide the following lemma, that shows that the expectation of M_k converges to 0 for any k .

Lemma 14 *For any natural k and a countable V ,*

$$\lim_{m \rightarrow \infty} \mathbb{E}[M_k(S^{(m)})] = 0$$

Proof Following McAllester and Schapire (2000) we have that for any m

$$\mathbb{E}[M_k(S^{(m)})] = \sum_{v \in V} p_v \Pr[|S_v^{(m)}| = k].$$

Since V is a countable set we can rewrite it as $V \stackrel{\Delta}{=} \{v_1, v_2, v_3, \dots\}$. Let $\varepsilon > 0$, and let N be a positive integer such that $\sum_{i=1}^N p_{v_i} > 1 - \frac{\varepsilon}{2}$. Since $\lim_{m \rightarrow \infty} \left(\Pr[|S_v^{(m)}| = k] \right) = 0$ for any natural k , there exists an m' such that for any $m > m'$, $\sum_{i=1}^N p_{v_i} \Pr[|S_{v_i}^{(m)}| = k] < \frac{\varepsilon}{2}$. In addition, $\sum_{i=N+1}^{|V|} p_{v_i} < \frac{\varepsilon}{2}$. Hence, for every $m > m'$,

$$\mathbb{E}[M_k(S^{(m)})] = \sum_{i=1}^N p_{v_i} \Pr[|S_{v_i}^{(m)}| = k] + \sum_{i=N+1}^{|V|} p_{v_i} \Pr[|S_{v_i}^{(m)}| = k] < \varepsilon.$$

■

Proof [of Thm. 3] To prove Eq. (9), we calculate the expectation of the generalization error $\mathbb{E}[\ell(h_S)]$ of an arbitrary hypothesis mapping $h \in H$ and show that this error is minimized when $h[S] = h_S^{\text{Bayes}}$. Let $f_h : \mathbb{N} \times \mathbb{N} \rightarrow [0, 1]$ be a function such that $f_h(n_1, n_2) = 1 - f_h(n_1, n_1 - n_2)$ and let h be a hypothesis mapping such that for all $v \in V$, $h[S](v) = f_h(c_v(S), c_v^+(S))$. Then,

$$\begin{aligned} \mathbb{E}[\ell(h[S])] &= \sum_v p_v \mathbb{E}[q_v(1 - f_h(c_v(S), c_v^+(S))) + (1 - q_v)f_h(c_v(S), c_v^+(S))] \\ &= \sum_v p_v (q_v + (1 - 2q_v)) \mathbb{E}[f_h(c_v(S), c_v^+(S))]. \end{aligned}$$

From the above expression it is clear that if $q_v < \frac{1}{2}$ then $\mathbb{E}[\ell(h[S])]$ is minimal when $\mathbb{E}[f_h(c_v(S), c_v^+(S))]$ is minimal, and if $q_v > \frac{1}{2}$ then $\mathbb{E}[\ell(h[S])]$ is minimal when $\mathbb{E}[f_h(c_v(S), c_v^+(S))]$ is maximal. If $q_v = \frac{1}{2}$ the expectation equals $\frac{1}{2}$ regardless of the choice of f_h . We have

$$\begin{aligned} \mathbb{E}[f_h(c_v(S), c_v^+(S))] &= \sum_S \Pr[S] f_h(c_v(S), c_v^+(S)) \\ &= \sum_{k=0}^m \Pr[c_v(S) = k] \sum_{i=0}^k \Pr[c_v^+(S) = i \mid c_v(S) = k] f_h(k, i) \end{aligned}$$

Consider the summation on i for a single k from the above sum. If k is odd, then

$$\begin{aligned}
 & \sum_{i=0}^k \Pr[c_v^+ = i \mid c_v = k] f_h(k, i) \\
 &= \sum_{i=0}^{\frac{k-1}{2}} \Pr[c_v^+ = i \mid c_v = k] f_h(k, i) + \sum_{i=\frac{k+1}{2}}^k \Pr[c_v^+ = i \mid c_v = k] (1 - f_h(k, k - i)) \\
 &= \sum_{i=0}^{\frac{k-1}{2}} \Pr[c_v^+ = i \mid c_v = k] f_h(k, i) + \sum_{i=0}^{\frac{k-1}{2}} \Pr[c_v^+ = k - i \mid c_v = k] (1 - f_h(k, i)) \\
 &= C + \sum_{i=0}^{\frac{k-1}{2}} (\Pr[c_v^+ = i \mid c_v = k] - \Pr[c_v^+ = k - i \mid c_v = k]) f_h(k, i)
 \end{aligned}$$

where C is a constant that does not depend on f_h . In the above expression, note that if $q_v < \frac{1}{2}$ then for each $i \leq \frac{k-1}{2}$, $\Pr[c_v^+ = i \mid c_v = k] - \Pr[c_v^+ = k - i \mid c_v = k]$ is positive, and that if $q_v > \frac{1}{2}$ then this expression is negative. This means that in both cases, to minimize $\mathbb{E}[\ell(h_S)]$, we need to maximize $f_h(k, i)$ for $i \leq \frac{k-1}{2}$. For an even k the analysis is similar, except that we have the special case of $i = \frac{k}{2}$ that does not pair with another summand. However, from the symmetry constraint on f_h it follows that $f_h(k, \frac{k}{2}) = \frac{1}{2}$. Therefore no maximization or minimization is allowed for this value of i . Based on the above analysis, the function f_h that minimizes $\mathbb{E}[\ell(h_S)]$ is:

$$f_h(n_1, n_2) = \begin{cases} 1 & n_2 \leq \frac{n_1-1}{2} \\ 0 & n_2 \geq \frac{n_1+1}{2} \\ \frac{1}{2} & n_2 = \frac{n_1}{2} \end{cases}$$

Setting $h_S(v) = f_h(c_v(S), c_v^+(S))$ we have that $h_S(v) = h_S^{\text{Bayes}}(v)$ for all values v in V .

To prove Eq. (10), we first calculate the difference between $\ell_v(h_S^{\text{Bayes}})$ and the expectation of $\ell_v(h_S^{\text{Bayes}})$. Assume without loss of generality that $q_v > \frac{1}{2}$. Then $\ell_v(h_S^{\text{Bayes}}) = p_v(1 - q_v)$, and

$$\mathbb{E}[\ell_v(h_S^{\text{Bayes}})] = p_v(q_v \Pr[\hat{q}_v < \frac{1}{2}] + (1 - q_v)(1 - \Pr[\hat{q}_v < \frac{1}{2}]) + \frac{1}{2} \Pr[\hat{q}_v = \frac{1}{2}]).$$

Subtracting, we have

$$\begin{aligned}
 \mathbb{E}[\ell_v(h_S^{\text{Bayes}})] - \ell_v(h_S^{\text{Bayes}}) &= p_v(2q_v - 1)(\Pr[\hat{q}_v < \frac{1}{2}] + \frac{1}{2} \Pr[\hat{q}_v = \frac{1}{2}]) \\
 &\leq p_v(2q_v - 1) \Pr[c_v = 0] \cdot \frac{1}{2} + p_v \sum_{k=1}^m \Pr[c_v = k] (2q_v - 1) \Pr[\hat{q}_v \leq \frac{1}{2} \mid c_v = k].
 \end{aligned}$$

We use Lemma 17 below to bound $(2q_v - 1) \Pr[\hat{q}_v \leq \frac{1}{2} \mid c_v = k]$ for $k \geq 3$. For $k = 0, 1, 2$ we maximize this term individually for each k . This leads us to the following bound:

$$\begin{aligned}
 & \mathbb{E}[\ell_v(h_S^{\text{Bayes}})] - \ell_v(h_S^{\text{Bayes}}) \\
 & \leq \frac{1}{2} p_v \Pr[c_v = 0] + \frac{1}{8} p_v \Pr[c_v = 1] + \frac{1}{8} p_v \Pr[c_v = 2] + \sum_{k=3}^m \frac{1}{\sqrt{ek}} p_v \Pr[c_v = k].
 \end{aligned}$$

Recall that M_k is the probability mass of the values seen k times in the sample. Following McAllester and Schapire (2000) we have that for $k \geq 0$, $\mathbb{E}[M_k] = \sum_\nu p_\nu \Pr[c_\nu = k]$. Hence, summing over all the values ν , we have

$$\begin{aligned} \mathbb{E}[\ell(h_S^{\text{Bayes}})] - \ell(h_\infty^{\text{Bayes}}) &= \sum_\nu (\mathbb{E}[\ell_\nu(h_S^{\text{Bayes}})] - \ell_\nu(h_\infty^{\text{Bayes}})) \\ &\leq \frac{1}{2} \mathbb{E}[M_0] + \frac{1}{8} \mathbb{E}[M_1] + \frac{1}{8} \mathbb{E}[M_2] + \sum_{k=3}^m \frac{1}{\sqrt{ek}} \mathbb{E}[M_k]. \end{aligned}$$

To prove Eq. (11), denote by $S^{(m)}$ a sample of m examples. Let $\varepsilon > 0$ be a scalar. Then there exists an integer t such that $\frac{1}{\sqrt{et}} < \frac{\varepsilon}{2}$. Since $\sum_{k=1}^m \mathbb{E}[M_k(S^{(m)})] = 1$, we have

$$\sum_{k=t}^m \frac{1}{\sqrt{ek}} \mathbb{E}[M_k(S^{(m)})] < \frac{\varepsilon}{2}. \tag{27}$$

Now, by Lemma 14, for every $k < t$, $\lim_{m \rightarrow \infty} \mathbb{E}[M_k(S^{(m)})] = 0$. Hence, there exists an m' such that for every $m > m'$,

$$\frac{1}{2} \mathbb{E}[M_0(S^{(m)})] + \frac{1}{8} \mathbb{E}[M_1(S^{(m)})] + \frac{1}{8} \mathbb{E}[M_2(S^{(m)})] + \sum_{k=3}^t \frac{1}{\sqrt{ek}} \mathbb{E}[M_k(S^{(m)})] < \frac{\varepsilon}{2}. \tag{28}$$

Combining Eq. (27) and Eq. (28), we have that for every $m > m'$,

$$\frac{1}{2} \mathbb{E}[M_0] + \frac{1}{8} \mathbb{E}[M_1] + \frac{1}{8} \mathbb{E}[M_2] + \sum_{k=3}^m \frac{1}{\sqrt{ek}} \mathbb{E}[M_k] < \varepsilon.$$

Hence the limit of this expression when $m \rightarrow \infty$ is 0. ■

4.4 Proof of Thm. 4

To prove Thm. 4, we first introduce some additional notation. Let $\delta \in (0, 1)$ be a confidence parameter. Let V_1^δ , V_2^δ , and V_3^δ be three sets that partition V according to the values of the probabilities p_ν :

$$\begin{aligned} V_1^\delta &= \{\nu \mid p_\nu \leq 6 \ln \left(\frac{2m}{\delta} \right) m^{-\frac{2}{3}}\} \\ V_2^\delta &= \{\nu \mid 6 \ln \left(\frac{2m}{\delta} \right) m^{-\frac{2}{3}} < p_\nu \leq 6 \ln \left(\frac{2m}{\delta} \right) m^{-\frac{1}{2}}\} \\ V_3^\delta &= \{\nu \mid 6 \ln \left(\frac{2m}{\delta} \right) m^{-\frac{1}{2}} < p_\nu\} \end{aligned}$$

We denote the contribution of each set to $\ell(h_S^{\text{Bayes}})$ by $\ell_i^\delta(S) \triangleq \sum_{\nu \in V_i^\delta} \ell_\nu(h_S^{\text{Bayes}})$. Additionally, given two samples S and S' , let $\kappa(S, S')$ be the predicate that gets the value “true” if for all $\nu \in V$ we have $c_\nu(S) = c_\nu(S')$.

Using the above definitions and the triangle inequality, we can bound $|\ell(h_S^{\text{Bayes}}) - \mathbb{E}[\ell(h_S^{\text{Bayes}})]|$ as follows:

$$\begin{aligned} |\ell(h_S^{\text{Bayes}}) - \mathbb{E}[\ell(h_S^{\text{Bayes}})]| &= \left| \sum_{i=1}^3 \left(\ell_i^\delta(S) - \mathbb{E}[\ell_i^\delta] \right) \right| \\ &\leq \left| \ell_1^\delta(S) - \mathbb{E}[\ell_1^\delta] \right| + \left| \ell_2^\delta(S) - \mathbb{E}[\ell_2^\delta(S') \mid \kappa(S, S')] \right| + \\ &\quad \left| \ell_3^\delta(S) - \mathbb{E}[\ell_3^\delta(S') \mid \kappa(S, S')] \right| + \left| \mathbb{E}[\ell_2^\delta(S') + \ell_3^\delta(S') \mid \kappa(S, S')] - \mathbb{E}[\ell_2^\delta + \ell_3^\delta] \right|. \end{aligned}$$

To prove Thm. 4 we bound each of the above terms as follows: First, to bound $|\ell_1^\delta(S) - \mathbb{E}[\ell_1^\delta]|$ (Lemma 15 below), we use the fact that for each $v \in V_1^\delta$ the probability p_v is small. Thus, a single change of an example in S has a moderate effect on the error and we can use McDiarmid's theorem. To bound $|\ell_2^\delta(S) - \mathbb{E}[\ell_2^\delta(S') \mid \kappa(S, S')]|$ (Lemma 16 below) we note that the expectation is taken with respect to those samples S' in which $c_v(S') = c_v(S)$ for all v . Therefore, the variables $\ell_v(h_S^{\text{Bayes}})$ are independent. We show in addition that each of these variables is bounded in $[0, p_v]$ and thus we can apply Hoeffding's bound. Next, to bound $|\ell_3^\delta(S) - \mathbb{E}[\ell_3^\delta(S') \mid \kappa(S, S')]|$ (Lemma 19 below), we use the fact that in a typical sample, $c_v(S)$ is large for all $v \in V_3^\delta$. Thus, we bound the difference between $\ell_v(h_S^{\text{Bayes}})$ and $\mathbb{E}[\ell_v(S') \mid \kappa(S, S')]$ for each value in V_3^δ separately. Then, we apply a union bound to show that for all of these values the above difference is small. Finally, we use the same technique to bound $|\mathbb{E}[\ell_2^\delta(S') + \ell_3^\delta(S') \mid \kappa(S, S')] - \mathbb{E}[\ell_2^\delta + \ell_3^\delta]|$ (Lemma 20 below). The proof of the first lemma, stated below, is omitted.

Lemma 15 $\forall \delta > 0 \quad \forall^\delta S \quad |\ell_1^\delta(S) - \mathbb{E}[\ell_1^\delta]| \leq \frac{12 \ln(\frac{2m}{\delta})}{m^{1/6}} \sqrt{\frac{1}{2} \ln\left(\frac{2}{\delta}\right)}$.

Proof We prove the lemma using McDiarmid's theorem. To do so, we examine the effect a removal of a single example (x_i, y_i) from S can have on $\ell_1^\delta(h_S^{\text{Bayes}})$. The largest effect occurs if $x_i \in V_1^\delta$ and the removal of y_i changes the value of $h^{\text{Bayes}}(x_i)$. In this case,

$$|\ell_1^\delta(S) - \ell_1^\delta(S \setminus i)| = |\ell_{x_i}(h_S^{\text{Bayes}}) - \ell_{x_i}(h_{S \setminus i}^{\text{Bayes}})| \leq p_v \leq 6 \ln\left(\frac{2m}{\delta}\right) m^{-\frac{2}{3}}.$$

Applying McDiarmid's theorem, it follows that $|\ell_1^\delta(S) - \mathbb{E}[\ell_1^\delta]|$ is at most

$$\sqrt{\frac{1}{2} \ln\left(\frac{2}{\delta}\right) m \cdot \left(12 \ln\left(\frac{2m}{\delta}\right) m^{-\frac{2}{3}}\right)^2} = \frac{12 \ln\left(\frac{2m}{\delta}\right)}{m^{1/6}} \sqrt{\frac{1}{2} \ln\left(\frac{1}{\delta}\right)}.$$

■

Lemma 16 $\forall \delta > 0 \quad \forall^\delta S \quad |\ell_2^\delta(S) - \mathbb{E}[\ell_2^\delta(S') \mid \kappa(S, S')]| \leq \frac{\sqrt{3 \ln\left(\frac{2m}{\delta}\right) \ln\left(\frac{2}{\delta}\right)}}{m^{1/4}}$.

Proof Since the expectation is taken over samples S' for which $c_v(S') = c_v(S)$ for each $v \in V$, we get that the value of the random variable $\ell_v(h_S^{\text{Bayes}})$ for each v depends only on the assignment of label for each example. Therefore the random variables $\ell_v(h_S^{\text{Bayes}})$ are all independent of each other when conditioned on $\kappa(S, S')$, and $\ell_2^\delta(S) = \sum_{v \in V_2^\delta} \ell_v(h_S^{\text{Bayes}})$ is a sum of independent random variables. The

expectation of this sum is $\mathbb{E}[\ell_2^\delta(S') \mid \kappa(S, S')]$. In addition, it is trivial to show that $\ell_v(h_S^{\text{Bayes}}) \in [0, p_v]$ for all v . Thus, by Hoeffding's inequality,

$$\Pr[|\ell_2^\delta(S) - \mathbb{E}[\ell_2^\delta(S') \mid \kappa(S, S')]| \geq t] \leq 2e^{-2t^2 / \sum_{v \in V_2^\delta} p_v^2}. \tag{29}$$

Using the fact that for v in V_2^δ , $p_v \leq 6 \ln\left(\frac{2m}{\delta}\right) / \sqrt{m}$ we obtain that

$$\sum_{v \in V_2^\delta} p_v^2 \leq \max_{v \in V_2^\delta} \{p_v\} \cdot \sum_{v \in V_2^\delta} p_v \leq 6 \ln\left(\frac{2m}{\delta}\right) / \sqrt{m}.$$

Plugging the above into Eq. (29) we get that

$$\Pr[|\ell_2^\delta(S) - \mathbb{E}[\ell_2^\delta(S') \mid \kappa(S, S')]| \geq t] \leq 2e^{-2t^2 \sqrt{m} / (6 \ln\left(\frac{2m}{\delta}\right))}.$$

Setting the right-hand side to δ and solving for t , we conclude our proof. ■

So far, we have bounded the terms $|\ell_1^\delta(S) - \mathbb{E}[\ell_1^\delta(S')]|$ and $|\ell_2^\delta(S) - \mathbb{E}[\ell_2^\delta(S') \mid \kappa(S, S')]|$. In both of these cases, we used the fact that p_v is small for all $v \in V_1^\delta \cup V_2^\delta$. We now turn to bound the term $|\ell_3^\delta(S) - \mathbb{E}[\ell_3^\delta(S') \mid \kappa(S, S')]|$. In this case, the probabilities p_v are no longer negligible. Therefore, we use a different technique whereby we analyze the probability of $h_S^{\text{Bayes}}(v)$ to be ‘wrong’, that is to return the less probable label. Since p_v is no longer small, we expect c_v to be relatively large. The following key lemma bounds the probability of $h_S^{\text{Bayes}}(v)$ to be wrong given that c_v is large. The resulting bound depends on the difference between q_v and $1/2$ and becomes vacuous whenever q_v is close to $1/2$. On the other hand, if q_v is close to $1/2$, the price we pay for a wrong prediction is small. In the second part of this lemma, we balance these two terms and end up with a bound that does not depend on q_v .

Lemma 17 *Let $\vec{Z} = (Z_1, \dots, Z_k)$ be a sequence of i.i.d. binary random variables such that $\Pr[Z_i = 1] = q$ for all i , and assume that $q \geq \frac{1}{2}$. Then,*

$$\Pr\left[\sum_i Z_i \leq k/2\right] \leq e^{-2(q-\frac{1}{2})^2 k} \quad \text{and} \quad (2q-1) \Pr\left[\sum_i Z_i \leq k/2\right] \leq \frac{1}{\sqrt{ek}}.$$

Proof The first inequality is a direct application of Hoeffding's inequality. Multiplying both sides by $2q-1$ we get that the left-hand side of the second inequality is bounded above by $(2q-1)e^{-2(q-\frac{1}{2})^2 k}$. We now let $x = q - \frac{1}{2}$ and use the inequality $2xe^{-2x^2 k} \leq 1/\sqrt{ek}$, which holds for all $x \geq 0$ and $k > 0$. ■

Based on the above lemma, we now bound $|\ell_3^\delta(S) - \mathbb{E}[\ell_3^\delta(S') \mid \kappa(S, S')]|$. First, we show that if $c_v(S)$ is large then $\ell_v(S)$ is likely to be close to the expectation of ℓ_v over samples S' in which $c_v(S) = c_v(S')$. This is equivalent to the claim of the following lemma.

Lemma 18 *Under the same assumptions of Lemma 17. Let $f(\vec{Z})$ be the function*

$$f(\vec{Z}) = \begin{cases} (1-q) & \text{if } \sum_i Z_i > k/2 \\ q & \text{if } \sum_i Z_i < k/2 \\ \frac{1}{2} & \text{if } \sum_i Z_i = k/2 \end{cases}.$$

Then, for all $\delta \in (0, e^{-1/2}]$ we have $\forall \delta \vec{Z} \quad |f(\vec{Z}) - \mathbb{E}[f]| \leq \sqrt{\frac{2 \ln\left(\frac{1}{\delta}\right)}{ek}}$.

Proof To simplify our notation, denote $\alpha = \Pr[\sum_i Z_i > k/2]$, $\beta = \Pr[\sum_i Z_i < k/2]$, and $\gamma = \Pr[\sum_i Z_i = k/2]$. A straightforward calculation shows that

$$|f(\bar{Z}) - \mathbb{E}[f(\bar{Z})]| = \begin{cases} (2q-1)(\beta + \gamma/2) & \text{with probability } \alpha \\ (2q-1)(\alpha + \gamma/2) & \text{with probability } \beta \\ (2q-1)(\alpha - \beta) & \text{with probability } \gamma \end{cases} .$$

Using the fact that (α, β, γ) is in the probability simplex we immediately obtain that

$$|f(\bar{z}) - \mathbb{E}[f(\bar{Z})]| \leq (2q-1) .$$

If $2q-1 \leq \sqrt{2 \ln(\frac{1}{\delta})}/k$ then the bound in the lemma clearly holds. Therefore, from now on we assume that $2q-1 > \sqrt{2 \ln(\frac{1}{\delta})}/k$. In this case, using the first inequality of Lemma 17 we have that $\beta + \gamma \leq e^{-2(q-\frac{1}{2})^2 k} \leq \delta$. Therefore, $1 - \delta < \alpha$, and so with probability of at least $1 - \delta$ we have that

$$|f(\bar{Z}) - \mathbb{E}[f(\bar{Z})]| = (2q-1)(\beta + \gamma/2) \leq (2q-1)(\beta + \gamma) .$$

Applying the second inequality of Lemma 17 on the right-hand side of the above inequality we get that $|f(\bar{Z}) - \mathbb{E}[f(\bar{Z})]| \leq \sqrt{1/ek} \leq \sqrt{2 \ln(1/\delta)/ek}$, where the last inequality holds since we assume that $\delta \leq e^{-1/2}$. \blacksquare

Equipped with the above lemma we are now ready to bound $|\ell_3^\delta(S) - \mathbb{E}[\ell_3^\delta(S') | \kappa(S, S')]|$.

Lemma 19 *If $m \geq 4$ then $\forall^{(2\delta)} S \quad |\ell_3^\delta(S) - \mathbb{E}[\ell_3^\delta(S') | \kappa(S, S')]| \leq 1/m^{\frac{1}{4}}$.*

Proof Recall that $\ell_3^\delta(S) = \sum_{v \in V_3^\delta} \ell_v(S)$. $m \geq 4$, hence $\delta/m \leq 1/m \leq e^{-1/2}$. Choose $v \in V_3^\delta$ and without loss of generality assume that $q_v \geq 1/2$. Thus, from Lemma 18 and the definition of $\ell_v(S)$ we get that with probability of at least $1 - \delta/m$ over the choice of the labels in $S(v)$:

$$|\ell_v(S) - \mathbb{E}[\ell_v(S') | \kappa(S, S')]| \leq p_v \sqrt{\frac{2 \ln(\frac{m}{\delta})}{e \cdot c_v(S)}} . \quad (30)$$

By the definition of V_3^δ and Lemma 10, $\forall^\delta S, \forall v \in V_3^\delta, c_v(S) \geq \rho(\delta/m, p_v, m)$. Using the fact that ρ is monotonically increasing with respect to p_v it is possible to show (see Lemma 21 in the appendix) that $\rho(\delta/m, p_v, m) \geq 2 \ln(\frac{m}{\delta}) m^{1/2}$ for all $v \in V_3^\delta$ for $m \geq 4$. Therefore, if indeed $c_v(S) \geq \rho(\delta/m, p_v, m)$ for any $v \in V_3^\delta$, we have that

$$\sqrt{\frac{2 \ln(\frac{m}{\delta})}{e \cdot c_v(S)}} \leq p_v m^{-1/4} .$$

Using a union bound to make sure that this condition holds and Eq. (30) holds for all $v \in V_3^\delta$ simultaneously, we obtain that $\forall^{(2\delta)} S \quad \forall v \in V_3^\delta \quad |\ell_v(S) - \mathbb{E}[\ell_v(S') | \kappa(S, S')]| \leq p_v m^{-1/4}$. Summing over $v \in V_3^\delta$, using the triangle inequality, and using the fact that $\sum_v p_v = 1$ we conclude the proof. \blacksquare

Lemma 20 For $m \geq 8$,

$$\forall \delta S \quad |\mathbb{E}[\ell_2^\delta(S') + \ell_3^\delta(S') \mid \kappa(S, S')] - \mathbb{E}[\ell_2^\delta(S') + \ell_3^\delta(S')]| \leq \frac{1}{m} + \frac{1}{m^{1/6}}.$$

Proof As in the proof of Lemma 19, we use the definitions of V_3^δ and V_2^δ along with Lemma 10 and Lemma 21 to get that for $m \geq 8$

$$\forall \delta S \quad \forall v \in V_2^\delta \cup V_3^\delta \quad c_v(S) \geq \rho(\delta/m, p_v, m) \geq 3 \ln(m/\delta) m^{1/3}. \quad (31)$$

To bound the difference between the conditional expectation and the unconditional expectation, let us first examine both these quantities for individual values v . To simplify our notation, denote $\alpha_1 = \Pr[\hat{q}_v(S') > 1/2 \mid c_v(S') = c_v(S)]$, $\beta_1 = \Pr[\hat{q}_v(S') < 1/2 \mid c_v(S') = c_v(S)]$, and $\gamma_1 = \Pr[\hat{q}_v(S') = 1/2 \mid c_v(S') = c_v(S)]$. Similarly, denote $\alpha_2 = \Pr[\hat{q}_v(S') > 1/2]$, $\beta_2 = \Pr[\hat{q}_v(S') < 1/2]$, and $\gamma_2 = \Pr[\hat{q}_v(S') = 1/2]$. Using the definition of ℓ_v we have that

$$\mathbb{E}[\ell_v(S') \mid c_v(S) = c_v(S')] = p_v \left((1 - q_v) \alpha_1 + q \beta_1 + \frac{1}{2} \gamma_1 \right).$$

Similarly, for the unconditional expectation:

$$\mathbb{E}[\ell_v(S')] = p_v \left((1 - q_v) \alpha_2 + q \beta_2 + \frac{1}{2} \gamma_2 \right). \quad (32)$$

Subtracting the above two equations and rearranging terms it can be shown that

$$\begin{aligned} \Delta &\triangleq |\mathbb{E}[\ell_v(S') \mid c_v(S) = c_v(S')] - \mathbb{E}[\ell_v(S')]| \\ &= p_v \left(q - \frac{1}{2} \right) | (\beta_1 + \gamma_1) - (\beta_2 + \gamma_2) + (\gamma_1 - \gamma_2) |. \end{aligned} \quad (33)$$

Let $Z_1, \dots, Z_{c_v(S)}$ be an i.i.d. sequence of random variables with $\Pr[Z_i = 1] = q_v$. Then we have $\beta_1 + \gamma_1 = \Pr[\sum_i Z_i \leq c_v(S)/2]$. In addition $c_v(S) \geq \lceil \rho(\delta/m, p_v, m) \rceil \triangleq \rho$. Assume without loss of generality that $q_v \geq 1/2$. Thus we have $\Pr[\sum_{i=1}^\rho Z_i \leq \rho/2] \geq \Pr[\sum_{i=1}^{c_v(S)} Z_i \leq c_v(S)/2]$. We clearly have that $0 \leq \beta_1 + \gamma_1 \leq \Pr[\sum_{i=1}^\rho Z_i \leq \rho/2]$. We now argue that

$$0 \leq \beta_2 + \gamma_2 \leq \frac{\delta}{m} + \Pr\left[\sum_{i=1}^\rho Z_i \leq \rho/2\right].$$

The left-hand side inequality is trivial. To prove the right-hand side inequality, we note that

$$\begin{aligned} \beta_2 + \gamma_2 &= \sum_{i=1}^m \Pr[c_v(S') = i] \Pr\left[\hat{q}_v(S') \leq \frac{1}{2} \mid c_v(S') = i\right] \\ &\leq \Pr[c_v(S') \leq \rho] + \Pr\left[\hat{q}_v(S') \leq \frac{1}{2} \mid c_v(S') = \rho\right] \\ &\leq \frac{\delta}{m} + \Pr\left[\sum_{i=1}^\rho Z_i \leq \rho/2\right]. \end{aligned}$$

Therefore,

$$|(\beta_1 + \gamma_1) - (\beta_2 + \gamma_2)| \leq \frac{\delta}{m} + \Pr\left[\sum_{i=1}^k Z_i \leq k/2\right]. \quad (34)$$

Similarly, since $0 \leq \gamma_1 \leq \beta_1 + \gamma_1$ and $0 \leq \gamma_2 \leq \beta_2 + \gamma_2$ we also have that

$$|\gamma_1 - \gamma_2| \leq \frac{\delta}{m} + \Pr\left[\sum_{i=1}^{\rho} Z_i \leq \rho/2\right]. \quad (35)$$

Combining Eq. (34) and Eq. (35) with Eq. (33) we get that

$$\Delta \leq p_v(2q-1) \left(\frac{\delta}{m} + \Pr\left[\sum_{i=1}^{\rho} Z_i \leq \rho/2\right] \right) \leq p_v \left(\frac{1}{m} + \frac{1}{\sqrt{e \cdot \rho(\frac{\delta}{m}, p_v, m)}} \right),$$

where the last inequality follows from Lemma 17. Finally, by summing over $v \in V_2^\delta \cup V_3^\delta$ and using Eq. (31) we conclude our proof. \blacksquare

5. Experiments

In this section we present experimental results that demonstrate the merits of our feature ranking criterion given in Eq. (7). Throughout this section we compare the following four feature ranking criteria:

1. **IG:** The Information Gain criterion (Quinlan, 1993; de Mantaras, 1991; Mitchell, 1997).
2. **IGR:** The Information Gain Ratio criterion (Quinlan, 1993).
3. **Gini:** The original Gini Index (Breiman et al., 1984), which is given in Eq. (2).
4. **Ginger:** Our modified Gini criterion that aims to minimize the generalization error, given in Eq. (7).

We first present experiments with synthetic data that exemplify the generalization properties of the different criteria. Next, we compare the performance of the different criteria on a natural data set from the UCI repository. Finally, we compare the use of the different ranking criteria for the task of growing a decision tree.

5.1 Synthetic Data

Three synthetic data sets were constructed to exemplify the generalization properties of the different ranking criteria in different scenarios. In all of the synthetic data sets the target label was first sampled according to the probability measure $\Pr[Y = 1] = \frac{1}{2}$. Synthetic data set I includes only binary features. The goal of data set I is to show that the Ginger criterion behaves similarly to the Gini criterion on binary features. 11 binary features were constructed. For each $i \in \{0, 1, \dots, 10\}$ the i th feature was sampled according to the probability measure $\Pr[X_i = Y|Y] = \frac{1+0.1i}{2}$. Thus, feature X_0 is completely uncorrelated with the label, while feature X_{10} perfectly predicts the label.

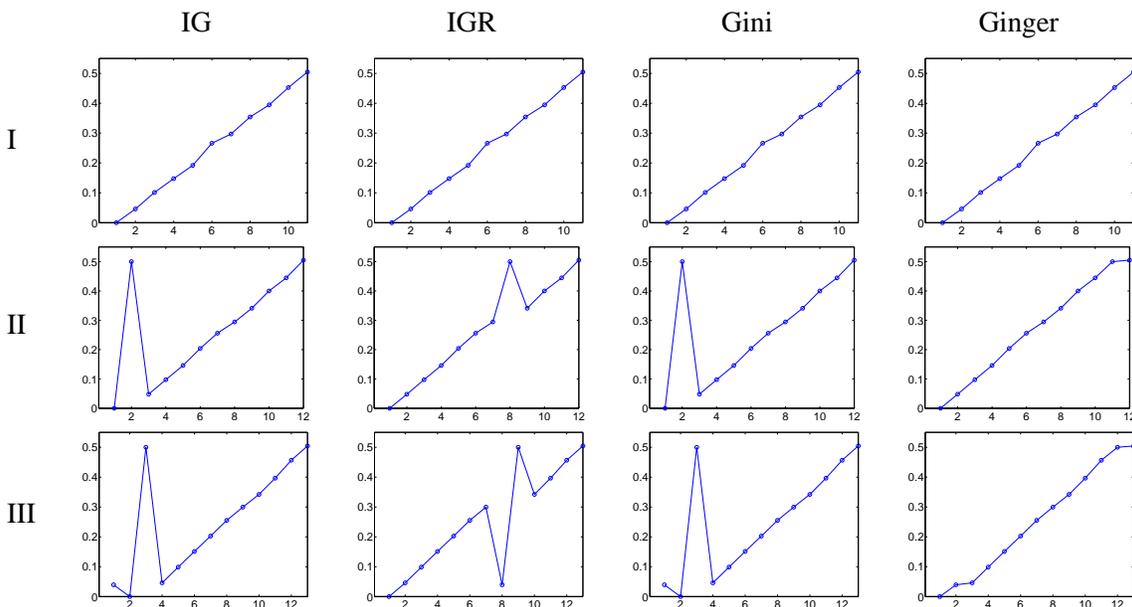


Figure 1: Each of the plots above show the generalization error of each feature (the y axis) against the ranking order of the feature in one of the ranking criteria (the x axis). Each column corresponds to a specific ranking criteria. Each row corresponds to a specific synthetic data set.

A training set of 5000 examples was generated, and the features were ranked using each of the four ranking criteria on the training set. The generalization errors of the 11 classification rules of each feature, defined as in Eq. (5), were measured on a fresh test set of 5000 examples. A plot of the generalization error of each feature against the ranking order of the feature is given for each of the ranking criteria on the top row of Fig. 1. This plot should be monotonically increasing for good feature ranking criteria. As the plots show, all four criteria perform well on this data set.

Data set II is identical to data set I, except that one more feature, indexed X_{11} , was added. X_{11} is simply the index of the example (this simulates an SSN-like feature as described in Sec. 2). Clearly, the generalization error of X_{11} is $\frac{1}{2}$ as no value of the feature that occurred in the training set would occur in a test set. The performance of the four feature ranking criteria on data set II is shown on the second row of Fig. 1. As expected, the Gini criterion and the IG criterion both suffer from overfitting and rank X_{11} very high. The IGR criterion, suggested by Quinlan (1993) attempts to fix the overfitting effect of the IG criterion by dividing IG by the entropy of the feature. As the plots show, this correction indeed causes IGR to rank X_{11} lower than do IG and Gini. However, the correction is not strong enough, as the new feature is still ranked 8th out of 12 features although its generalization error is the worst. Finally, it is clear from the plots that the new Ginger criterion produces a correct ranking of the features in this example.

Data set III is identical to data set II, except that one more feature indexed X_{12} was added. X_{12} was constructed according to the following probability measure:

$$\begin{aligned} \Pr[X = i | Y = 1] &= \begin{cases} \frac{1}{2000} & \text{if } i \in \{1, \dots, 2000\} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \\ \Pr[X = i | Y = -1] &= \begin{cases} \frac{1}{2000} & \text{if } i \in \{2001, \dots, 4000\} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

X_{12} is thus categorical with many values but it is still highly predictive of the label. The performance of the four feature ranking criteria on data set III is shown on the bottom row of Fig. 1. As the plots show, the rankings of the Gini criterion and of the IG criterion are not adversely affected by the addition of this feature, although they still fail on X_{11} , the SSN-like feature. IGR penalizes X_{12} because it has a large number of values, thus its ranking for this feature is too low. The new Ginger criterion is the only one to rank the features in accordance with their respective generalization error, as is apparent from its monotonically increasing plot.

5.2 Natural Data

To test the ranking criteria on natural data, we used the USCensus1990raw data set from the UCI Repository.¹ This data set contains person records, where each record has 125 features, such as age, salary, marital status etc. Several labeled data sets were constructed from USCensus1990raw by defining a binary target label based on one of the attributes, and using the rest of the attributes as features. For attributes that take more than two values, the binary label was set to 1 if the feature takes its most frequent value and -1 otherwise. Only cases where the probability of the label to be 1 was at least 0.1 and no more than 0.9 were used. This process resulted in 62 binary learning problems.

In Fig. 2, each of the rows corresponds to one learning problem. A plot is shown for each problem and each ranking criterion, depicting the generalization error of each feature against the ranking order of the features. Recall that good ranking criteria should produce monotonically increasing graphs. The plots clearly show that the Ginger criterion produces the most accurate feature ranking. Fig. 3 compares the Ginger criterion to each of the other ranking criteria. In each of the plots, each data point corresponds to one of the 62 learning problems and portrays the difference in generalization error between the feature that was top-ranked by Ginger and the feature that was top-ranked by the other criterion. Positive data points are cases where Ginger outperformed the other criterion. Again, it is apparent that the Ginger criterion outperforms the other criteria.

5.3 Decision Trees

Decision trees are a popular classification tool (see for instance Mitchell, 1997). The process of growing a decision tree is a greedy iterative procedure which is performed as follows: The procedure starts with a tree composed only of a root node. At each iteration, one of the leaves of the tree is turned into an inner node, whose children represent all the possible values of one feature. Choosing

1. The original census data set was used rather than the preprocessed data set. The preprocessed data set obtained from Meek, Thiesson, and Heckerman eliminates categorical attributes that have many values, exactly the type of attributes that this paper addresses. The data set used in our experiments is available through <http://kdd.ics.uci.edu/databases/census1990/USCensus1990raw.data.txt>.

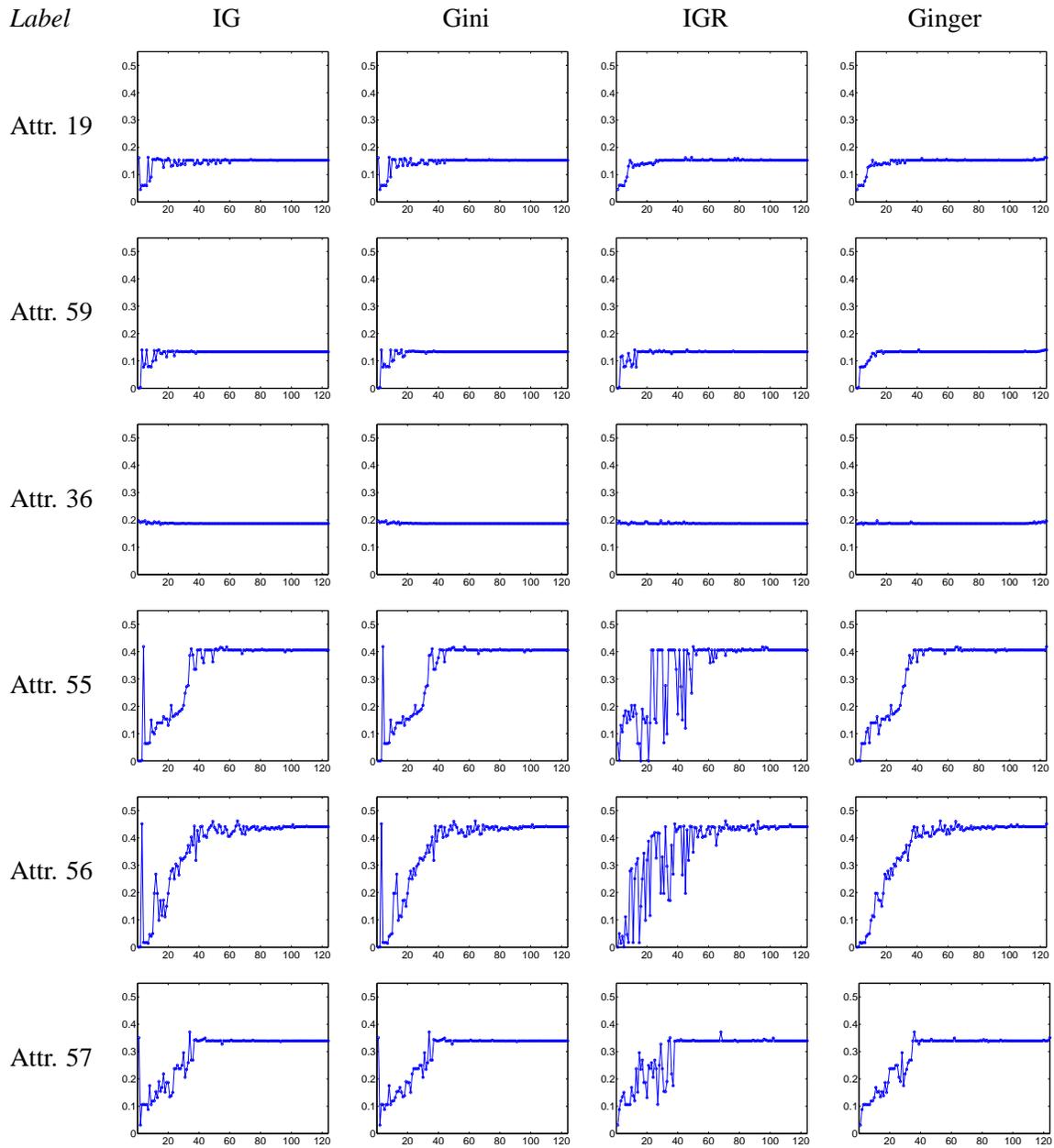


Figure 2: Each of the plots above show the generalization error of the features in a learning problem (the y axis) against the ranking order of the features in one of the ranking criteria (the x axis). Each column corresponds to a specific ranking criterion. Each row corresponds to a specific learning problem, generated from USCensus1990raw by setting the label to be the most common value of one of the attributes.

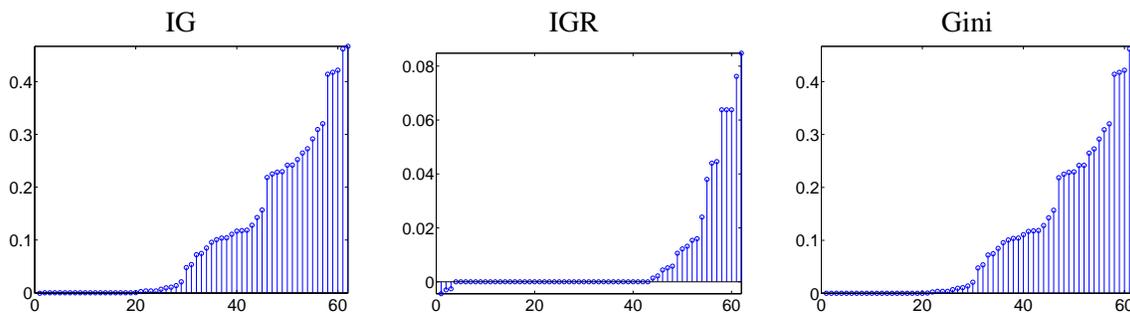


Figure 3: Each plot above portrays the difference in generalization error between the feature that was top-ranked by Ginger and the feature that was top-ranked by one of the other criteria, for each of the 62 learning problems obtained from USCensus1990raw.

which leaf to split and which feature to use for splitting can be based on feature ranking criteria such as the ones discussed in this paper. In our experiments, we compared decision tree learning with each of the four feature ranking criteria: IG, IGR, Gini, and Ginger. The experiments were performed on the 62 learning problems described in Sec. 5.2.

Usually, the iterative process of growing a decision tree continues until no further splits can be made. Then, as a post processing step, the tree is pruned, so as to improve the generalization error of the decision tree. Since this paper focuses on splitting criteria rather than on pruning methods, the experiments do not include tree post-pruning. Instead, the generalization error is measured as a function of the number of splits. Given a ranking criterion, the following procedure is used to choose which leaf to split and which feature to split by: Let m be the number of training examples. A decision tree T with k leaves is equivalent to a mapping $T : \{1, \dots, m\} \rightarrow \{1, \dots, k\}$. That is, each example is mapped to one of the leaves of the tree. We can think of the vector $(T(1), \dots, T(m))$ as the vector of values of a constructed feature. At each iteration of the decision tree learning process, a new tree needs to be generated from the current tree by splitting one of the current tree leaves based on one of the features. Each possible new tree induces a different new constructed feature as described above. To select the leaf to split and the feature to split by, we assess the quality of each new constructed feature based on the ranking criterion in use. The selected leaf and feature are those that correspond to the top-ranked constructed feature.

Fig. 4 shows the training error and generalization error of the Gini, IGR and Ginger splitting criteria as a function of the number of splits, for several learning problems. The IG criterion plot was omitted since its behavior was almost identical to that of the Gini criterion. As can be seen from the plots, the training error of the Gini criterion drops faster, but the resulting tree suffers from severe overfitting. In contrast, the generalization error of the Ginger criterion is much smaller and remains close to the training error, as long as the number of splits is not too large. As expected, after making a large number of splits all criteria exhibit an overfitting effect. Comparing the IGR and the Ginger criteria, we observe that both methods perform rather well, each showing an advantage on some of the learning problems.

Lastly, Fig. 5 compares the performance of the decision tree learning with the Ginger splitting criterion to decision tree learning with the other splitting criteria. In each of the plots, the data points correspond to the 62 learning problems, and portray the difference in the minimal generalization

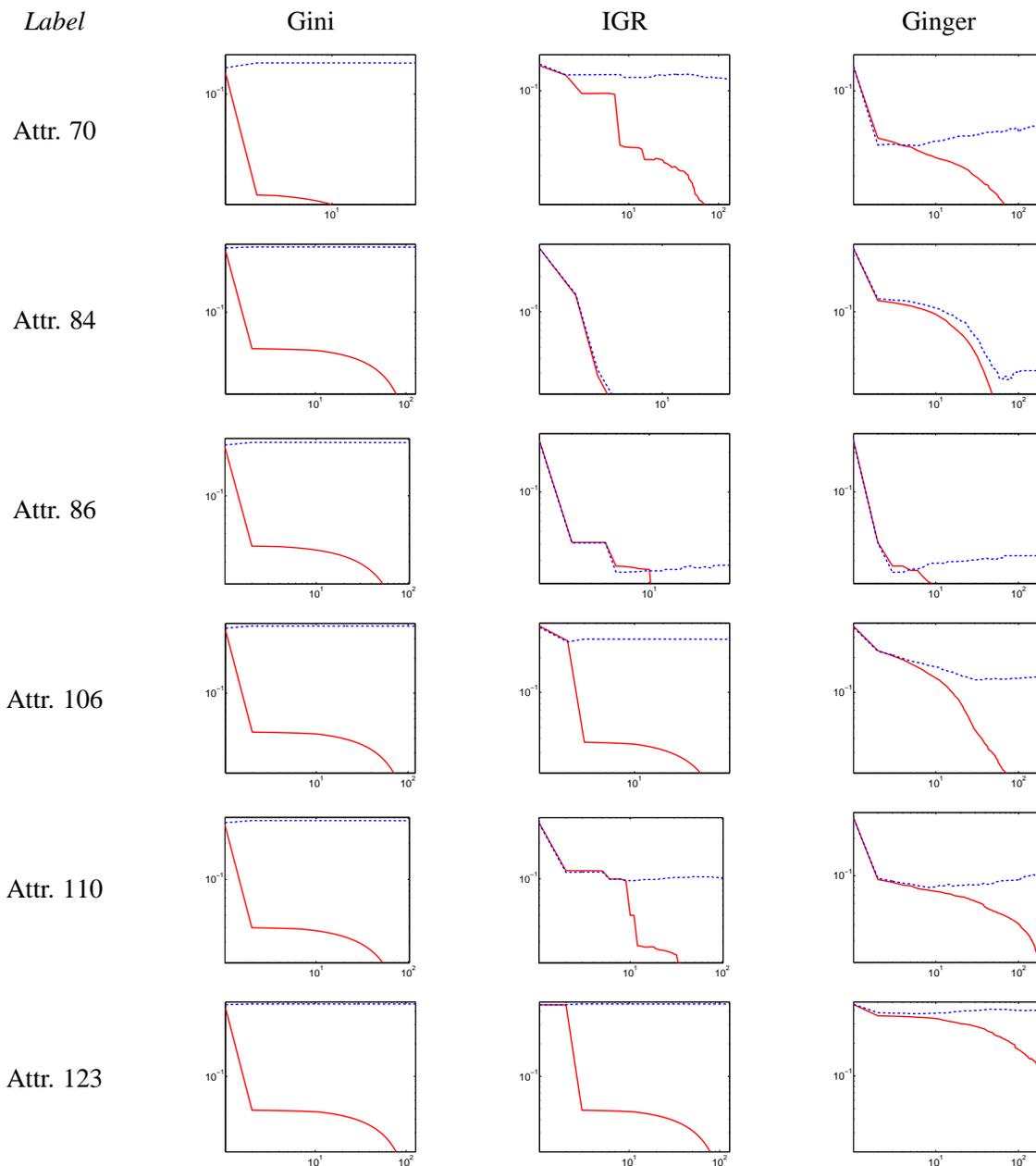


Figure 4: The training error (solid red line) and generalization error (dotted blue line) of decision trees grown according to the Gini, IGR, and Ginger splitting criteria, as a function of the number of splits. Each column corresponds to a specific splitting criterion. Each row corresponds to a specific learning problem, generated from USCensus1990raw by setting the label to be the most common value of one of the attributes.

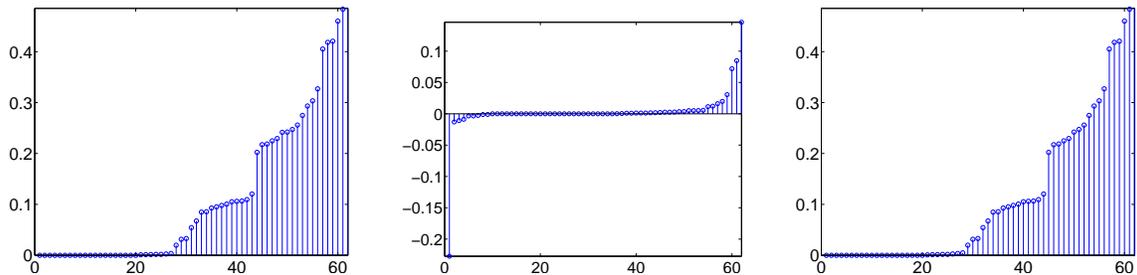


Figure 5: Left: The minimal generalization error of the IG criterion minus the minimal generalization error of the Ginger criterion for each of the labeled data sets. Middle: Same for IGR. Right: Same for Gini.

error achieved by the decision tree grown using Ginger and the one that was achieved using the other criterion. Positive data points are cases where Ginger outperformed the other criterion. The plots show that the Ginger criterion outperforms the IG and Gini criteria, and that in most cases the Ginger criterion outperforms the IGR criterion as well.

6. Discussion

In this paper, a new approach for feature ranking is proposed, based on a direct estimation of the true generalization error of predictors that are deduced from the training set. We focused on two specific predictors, namely h_S^{Gini} and h_S^{Bayes} . An estimator for the generalization error of h_S^{Gini} , termed the Ginger criterion, was proposed and its convergence was analyzed. Experimental evaluation suggests that the Ginger criterion outperforms existing feature ranking methods. We showed that the expected error of h_S^{Bayes} is optimal and proved a concentration bound for this error. Constructing an estimator for h_S^{Bayes} is left for future work.

There are various extensions for this work that we did not pursue. First, it is interesting to analyze the number of categorical features one can rank while avoiding overfitting. The experiments with decision trees suggest that the Ginger criterion has potential to improve the generalization error of decision trees. It may be possible to use the bounds for constructing a stopping criterion for growing the decision tree. Second, our view of a ranking criterion as an estimator for the generalization error of a predictor can be used for constructing new ranking criteria by defining other predictors. Finally, understanding the relationship between this view and information theoretic measures is also an interesting future direction.

Appendix A. Technical Proofs

Lemma 21 *Let c be a positive constant. Then, if $p_v > 6 \ln \left(\frac{2}{\delta} \right) m^{-c}$, and $m \geq 2^{\frac{1}{1-c}}$ we have*

$$\forall \delta > 0 \quad \rho(\delta, p_v, m) \geq 3 \ln \left(\frac{2}{\delta} \right) m^{1-c}.$$

Proof By the definition of ρ ,

$$\rho(\delta, p_v, m) = mp_v - \sqrt{mp_v \cdot 3 \ln \left(\frac{2}{\delta} \right)} = \sqrt{mp_v} \left(\sqrt{mp_v} - \sqrt{3 \ln \left(\frac{2}{\delta} \right)} \right).$$

Therefore, $\rho(\frac{\delta}{m}, p_v, m)$ is upward monotonic with p_v . Thus if $p_v > 6 \ln \left(\frac{2m}{\delta} \right) m^{-c}$,

$$\begin{aligned} \rho(\delta, p_v, m) &= mp_v - \sqrt{mp_v \cdot 3 \ln \left(\frac{2}{\delta} \right)} \\ &\geq 6 \ln \left(\frac{2}{\delta} \right) m^{1-c} - \sqrt{6 \ln \left(\frac{2}{\delta} \right) m^{1-c} \cdot 3 \ln \left(\frac{2}{\delta} \right)} \\ &= 3 \ln \left(\frac{2}{\delta} \right) m^{\frac{1-c}{2}} \left(2m^{\frac{1-c}{2}} - \sqrt{2} \right) \\ &= 3 \ln \left(\frac{2}{\delta} \right) m^{\frac{1-c}{2}} \left(m^{\frac{1-c}{2}} + m^{\frac{1-c}{2}} - \sqrt{2} \right) \\ &\geq 3 \ln \left(\frac{2}{\delta} \right) m^{1-c}. \end{aligned}$$

■

Proof [Lemma 12] Similarly to the proof of Lemma 8, we will bound the effect a single removal of an example from S can have on $\ell(h_S^\delta)$. The maximal effect of a single change in the sample is no larger than twice the maximal effect of a single removal. Assume without loss of generality that the removed example is $x_i = (v, 0)$, and denote the resulting sample by $S \setminus i$. The removal only affects $\ell_v(h_S^\delta)$. Therefore

$$\begin{aligned} |\ell(h_S^\delta) - \ell(h_{S \setminus i}^\delta)| &= |\ell_v(h_S^\delta) - \ell_v(h_{S \setminus i}^\delta)| \\ &= \left| p_v \left(q_v(1 - h_S^\delta(v)) + (1 - q_v)h_S^\delta(v) - p_v q_v(1 - h_{S \setminus i}^\delta(v)) + (1 - q_v)h_{S \setminus i}^\delta(v) \right) \right| \\ &= \left| p_v(1 - 2q_v)(h_S^\delta(v) - h_{S \setminus i}^\delta(v)) \right| \\ &\leq p_v \left| h_S^\delta(v) - h_{S \setminus i}^\delta(v) \right|. \end{aligned}$$

For v such that $p_v < \frac{6 \ln(\frac{2m}{\delta})}{m}$,

$$|\ell(h_S^\delta) - \ell(h_{S \setminus i}^\delta)| \leq p_v < \frac{6 \ln(\frac{2m}{\delta})}{m}. \tag{36}$$

For v such that $p_v \geq \frac{6 \ln(\frac{2m}{\delta})}{m}$, we distinguish between three cases by c_v , the number of examples of v in S :

1. $c_v < \rho(\frac{\delta}{m}, p_v, m)$,
2. $\rho(\frac{\delta}{m}, p_v, m) \leq c_v < \rho(\frac{\delta}{m}, p_v, m) + 1$,

$$3. \rho\left(\frac{\delta}{m}, p_v, m\right) + 1 \leq c_v.$$

In case 1,

$$h_S^\delta(v) = \frac{c_v^+ + q_v(\lceil \rho\left(\frac{\delta}{m}, p_v, m\right) \rceil - c_v)}{\lceil \rho\left(\frac{\delta}{m}, p_v, m\right) \rceil} \quad \text{and} \quad h_{S^i}^\delta(v) = \frac{c_v^+ + q_v(\lceil \rho\left(\frac{\delta}{m}, p_v, m\right) \rceil - (c_v - 1))}{\lceil \rho\left(\frac{\delta}{m}, p_v, m\right) \rceil},$$

hence

$$|h_S^\delta(v) - h_{S^i}^\delta(v)| = \frac{q_v}{\lceil \rho\left(\frac{\delta}{m}, p_v, m\right) \rceil}.$$

In case 2, $\lceil \rho\left(\frac{\delta}{m}, p_v, m\right) \rceil = c_v$, therefore

$$h_S^\delta(v) = h_S^{\text{Gini}}(v) = \frac{c_v^+}{c_v} \quad \text{and} \quad h_{S^i}^\delta(v) = \frac{c_v^+ + q_v(c_v - (c_v - 1))}{c_v},$$

hence

$$|h_S^\delta(v) - h_{S^i}^\delta(v)| = \frac{q_v}{c_v} = \frac{q_v}{\lceil \rho\left(\frac{\delta}{m}, p_v, m\right) \rceil}.$$

In case 3, since $\rho\left(\frac{\delta}{m}, p_v, m\right) > 1$ we have $c_v \geq 2$ and

$$h_S^\delta(v) = h_S^{\text{Gini}}(v) = \frac{c_v^+}{c_v} \quad \text{and} \quad h_{S^i}^\delta(v) = h_{S^i}^\delta(v) = \frac{c_v^+}{c_v - 1}$$

Hence

$$|h_S^\delta(v) - h_{S^i}^\delta(v)| = \frac{c_v^+}{c_v(c_v - 1)} \leq \frac{c_v}{c_v(c_v - 1)} = \frac{1}{c_v - 1} \leq \frac{1}{\lceil \rho\left(\frac{\delta}{m}, p_v, m\right) \rceil}.$$

Therefore, in all cases, for v such that $p_v \geq \frac{6 \ln\left(\frac{2m}{\delta}\right)}{m}$,

$$\begin{aligned} |\ell(h_S^\delta) - \ell(h_{S^i}^\delta)| &\leq p_v |h_S^\delta(v) - h_{S^i}^\delta(v)| \leq \frac{p_v}{\rho\left(\frac{\delta}{m}, p_v, m\right)} = \frac{p_v}{mp_v - \sqrt{mp_v \cdot 3 \ln\left(\frac{2m}{\delta}\right)}} \\ &= \frac{1}{m} \frac{\sqrt{p_v}}{\sqrt{p_v} - \sqrt{\frac{3 \ln\left(\frac{2m}{\delta}\right)}{m}}} \leq \frac{1}{m} \left(\frac{\sqrt{\frac{6 \ln\left(\frac{2m}{\delta}\right)}{m}}}{\sqrt{\frac{6 \ln\left(\frac{2m}{\delta}\right)}{m}} - \sqrt{\frac{3 \ln\left(\frac{2m}{\delta}\right)}{m}}} \right) = \frac{1}{m} \frac{2}{\sqrt{2} - 1} \leq \frac{4}{m}. \end{aligned}$$

Combining this with Eq. (36), we have

$$|\ell(h_S^\delta) - \ell(h_{S^i}^\delta)| \leq \max \left\{ \frac{4}{m}, \frac{6 \ln\left(\frac{2m}{\delta}\right)}{m} \right\} = \frac{6 \ln\left(\frac{2m}{\delta}\right)}{m}.$$

Hence, doubling the effect of a single removal, we have that for any two samples S_1 and S_2 such that $d(S_1, S_2) \leq 1$

$$|\ell(h_{S_1}^\delta) - \ell(h_{S_2}^\delta)| \leq \frac{12 \ln\left(\frac{2m}{\delta}\right)}{m}.$$

■

References

- A. Antos and I. Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Struct. Algorithms*, 19(3-4):163–193, 2001.
- A. Antos, L. Devroye, and L. Györfi. Lower bounds for Bayes error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(7):643–645, 1999.
- L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth & Brooks, 1984.
- R. Lopez de Mantaras. A distance-based attribute selection measure for decision tree induction. *Machine Learning*, 6(1):81–92, 1991.
- E. Drukh and Y. Mansour. Concentration bounds for unigrams language model. *Journal of Machine Learning Research*, 6:1231–1264, 2005.
- I.J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:237–264, 1953.
- T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.
- M. Kearns and Y. Mansour. On the boosting ability of top-down decision tree learning algorithms. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on the Theory of Computing*, pages 459–468, 1996.
- S. Kutin. Extensions to McDiarmid’s inequality when differences are bounded with high probability. Technical report, University of Chicago TR-2002-04, 2002.
- D.A. McAllester and R.E. Schapire. On the convergence rate of good-turing estimators. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, pages 1–6, 2000.
- C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, pages 148–188, 1989.
- J. Mingers. An empirical comparison of selection measures for decision-tree induction. *Machine Learning*, 3:319–342, 1989.
- T. M. Mitchell. *Machine Learning*. McGraw Hill, 1997.
- J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
- K. Torkkola. Information theoretic methods. In I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, editors, *Feature Extraction, Foundations and Applications*. Springer, 2006.
- L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004.