Prediction by Categorical Features: Generalization Properties and Application to Feature Ranking

Sivan Sabato[1] and Shai Shalev-Shwartz[2,1]

[1] IBM Research Laboratory in Haifa, Haifa 31905, Israel
[2] School of Computer Sci. & Eng., The Hebrew University, Jerusalem 91904, Israel

Abstract. We describe and analyze a new approach for feature ranking in the presence of categorical features with a large number of possible values. It is shown that popular ranking criteria, such as the Gini index and the misclassification error, can be interpreted as the training error of a predictor that is deduced from the training set. It is then argued that using the *generalization* error is a more adequate ranking criterion. We propose a modification of the Gini index criterion, based on a robust estimation of the generalization error of a predictor associated with the Gini index. The properties of this new estimator are analyzed, showing that for most training sets, it produces an accurate estimation of the true generalization error. We then 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.

1 Introduction

Filter methods for supervised feature selection rank a given set of features according to their relevance for predicting the label. As in other supervised learning methods, the ranking of the features is generated based on an input training set. Examples of widely used filter ranking criteria are the Gini index, the misclassification error, and the cross-entropy [6]. In this paper we address the problem of 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 features that can take many values. We propose an adaptation of existing filter criteria that copes with these difficulties.

Many feature ranking methods can be viewed as a 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. The training set is used both for constructing each predictor and for evaluating its error. Most current filters use the error over the training set as the ranking criterion. In contrast, we argue that the *generalization* error of each predictor is a more adequate ranking criterion. When dealing with binary features, the training error is likely to be close to the generalization error, and therefore the ranking generated by current filters 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. 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 most 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.

Filter methods have been extensively studied in the context of decision trees [10, 7, 12]. The failure of existing filter ranking criteria in the presence of categorical features with a large number of possible values has been previously discussed in [12, 11]. Quinlan suggested the Information Gain Ratio as a correction to the cross-entropy (a.k.a. Information Gain) criterion. In a broader context, information-theoretic measures are commonly used for feature ranking (see for example [14] 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 [14]. However, obtaining estimators for the entropy or mutual information seems to be difficult in the general case [2]. Another ranking criterion designed to address the above difficulty is a distance-based measure introduced by [3].

The problem we address shares some similarities with the problem of estimating the missing mass of a sample, typically encountered in language modeling [5, 8, 4]. 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 [8, 4]. 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.

2 Problem Setting

In this section we establish the notation used throughout the paper and formally describe our problem setting. In the supervised feature selection 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 by X_i the i'th feature and by V_i the set of values X_i can take. We make no assumptions on the identity of V_i 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 selection is to find a subset of the features that can be used later for constructing an accurate classification rule. We focus on the filter approach in which we rank *individual* features according to their "relevance" to the label. Different filters employ different criteria for assessing the relevance of a feature to the label. Since we are dealing with individual features, 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 (and denote by V the set of values X can take). To simplify our notation we denote $p_v \overset{\Delta}{=} \Pr[X = v]$ and $q_v \overset{\Delta}{=} \Pr[Y = 1|X = v]$.

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:

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

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

Two popular filters used for feature selection [6] are the misclassification error

$$\textstyle\sum_{v \in V} \hat{p}_v \, \min\{\hat{q}_v, (1 - \hat{q}_v)\} \; , \tag{2}$$

and the Gini index

$$2 \, \textstyle\sum_{v \in V} \hat{p}_v \, \hat{q}_v (1 - \hat{q}_v) \; . \tag{3}$$

In these filters, 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 also [11]), 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 filter 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. (2) and Eq. (3). The aforementioned problem was previously underscored in the context of the Information Gain filter [12, 3, 11]. In that context, Quinlan [12] 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 filter's 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 $h : V \to [0, 1]$, where $h(v)$ is the probability to predict the label 1 given the value v. The generalization error of h is the probability to wrongly predict the label,

$$\ell(h) \overset{\Delta}{=} \sum_{v \in V} p_v \; \left(q_v \left(1 - h(v)\right) + \left(1 - q_v\right) h(v)\right) \;\; . \tag{4}$$

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

$$h_S^{\text{Gini}}(v) = \hat{q}_v \;\; . \tag{5}$$

As its name indicates, h_S^{Gini} is closely related to the Gini index filter given in Eq. (3). To see this, we note that the generalization error of h_S^{Gini} is

$$\ell(h_S^{\text{Gini}}) = \sum_{v \in V} p_v \; \left(q_v \left(1 - \hat{q}_v\right) + \left(1 - q_v\right) \hat{q}_v\right) \;\; . \tag{6}$$

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. (3). This will be approximately true, for example, when $m \gg |V|$. In contrast, when the training set is small, using $\ell(h_S^{\text{Gini}})$ is preferable to using the Gini index given in Eq. (3), because $\ell(h_S^{\text{Gini}})$ takes into account the fact that the estimated probabilities might be skewed.

The second hypothesis we define is

$$h_S^{\text{Bayes}}(v) \;=\; \begin{cases} 1 & \hat{q}_v > \frac{1}{2} \\ 0 & \hat{q}_v < \frac{1}{2} \\ \frac{1}{2} & \hat{q}_v = \frac{1}{2} \end{cases} \;\; . \tag{7}$$

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. (2). 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. (3) and Eq. (2), we return to the example mentioned above. For the SSN feature we have $\ell(h_S^{\text{Gini}}) = \ell(h_S^{\text{Bayes}}) = \frac{1}{2} M_0$, where $M_0 \overset{\Delta}{=} \sum_{v : c_v = 0} p_v$. In general, we denote

$$M_k \;\overset{\Delta}{=}\; \sum_{v : c_v = k} p_v \;\; . \tag{8}$$

The quantity M_0 is known as the missing mass [5, 8] 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 the second feature (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 utilized 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}:

$$\hat{\ell} \triangleq \frac{|\{v : c_v = 1\}|}{2m} + \sum_{v:c_v>1} \frac{2c_v}{c_v - 1} \hat{p}_v \hat{q}_v (1 - \hat{q}_v) \ . \tag{9}$$

In the next section, we derive this estimator based on a conditional cross-validation technique. We suggest to use the estimation of $\ell(h_S^{\text{Gini}})$ given in Eq. (9) rather than the original Gini index given in Eq. (3) 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. (9) and Eq. (3) 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 [5, 8]).

We now turn to analyze the quality of the proposed estimator. We first show (Thm. 1 below) 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 $O(\frac{\ln(m/\delta)}{\sqrt{m}})$.

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. (5) and let $\ell(h_S^{\text{Gini}})$ be* *the generalization error of h_S^{Gini}, where ℓ is as defined in Eq. (4). Let $\hat{\ell}$ be the estimation of $\ell(h_S^{\text{Gini}})$ as given in Eq. (9). Then, $\left|\mathbb{E}[\ell(h_S^{\text{Gini}})] - \mathbb{E}[\hat{\ell}]\right| \leq \frac{1}{2m}$, where expectation is taken over all sets 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*

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

Based on the above theorem, $\hat{\ell}$ can be used as a filter 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. (7) is optimal in this respect, but that its concentration is 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. This symmetry requirement limits the examined hypotheses to those that do not exploit prior knowledge on the underlying distribution $\Pr[X,Y]$. Formally, let H_S be the set of all hypotheses that can be written as $h(v) = f_h(c_v(S), c_v^+(S))$ where $f_h : \mathbb{N} \times \mathbb{N} \to [0,1]$ is a function such that $f_h(n_1, n_2) = 1 - f_h(n_1, n_1 - n_2)$ for all $n_1, n_2 \in \mathbb{N}$. 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. (4). Let h_S^{Bayes} be the hypothesis given in Eq. (7) and let h_∞^{Bayes} be the Bayes optimal hypothesis. Let H_S be the set of symmetric hypotheses. Then $\mathbb{E}[\ell(h_S^{\text{Bayes}})] = \min_{h \in H_S} \mathbb{E}[\ell(h)]$, and*

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

where M_k is as defined in Eq. (8).

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*

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

The concentration bound for $\ell(h_S^{\text{Bayes}})$ is worse 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 filter 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. [1] 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 the results presented in the previous section are proved. Due to the lack of space, some of the proofs are omitted and can be found in [13].

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. (9) using a conditional cross-validation technique, and then utilize this interpretation of $\hat{\ell}$ to prove Thm. 1 and Thm. 2.

To derive the estimator given in Eq. (9), 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

$$\ell_v(h) \overset{\Delta}{=} \Pr[X=v]\Pr[h(X) \neq Y \mid X=v] = p_v\left(q_v\,(1-h(v)) + (1-q_v)\,h(v)\right).$$

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 $\hat{\ell} \overset{\Delta}{=} \sum_v \hat{\ell}_v$.

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. (5), 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 $\ell_v(h_S^{\text{Gini}})$,

$$\begin{aligned} \hat{\ell}_v &= \hat{p}_v \frac{1}{c_v(c_v - 1)} \sum_{i,j \in S(v), i \neq j} \mathbf{1}_{y_i \neq y_j} \qquad (10) \\ &= \hat{p}_v \frac{2c_v^+(c_v - c_v^+)}{c_v(c_v - 1)} \;=\; \hat{p}_v \frac{2c_v^2 \hat{q}_v(1 - \hat{q}_v)}{c_v(c_v - 1)} \;=\; \hat{p}_v \cdot \frac{2c_v}{c_v - 1} \hat{q}_v(1 - \hat{q}_v). \end{aligned}$$

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:

$$\hat{\ell}_v = \begin{cases} \hat{p}_v \cdot \frac{1}{2} & c_v \leq 1 \\ \hat{p}_v \cdot \frac{2c_v}{c_v - 1} \hat{q}_v(1 - \hat{q}_v) & c_v \geq 2 \end{cases} \qquad (11)$$

The resulting estimator $\hat{\ell}$ defined in Eq. (9) 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 1. *For k such that $1 < k \leq m$, $\mathbb{E}[\hat{\ell}_v \mid c_v(S) = k] = \frac{k}{m} \cdot 2q_v(1 - q_v)$.*

Proof. If $c_v = k$, then $\hat{p}_v = \frac{k}{m}$. Therefore, based on Eq. (10), we have

$$\mathbb{E}[\hat{\ell}_v \mid c_v(S) = k] = \frac{k}{m}\frac{1}{k(k-1)}\,\mathbb{E}\,\Big[\sum_{i,j \in S(v), i \neq j} \mathbf{1}_{y_i \neq y_j} \mid c_v(S) = k\,\Big] \ . \tag{12}$$

Let $Z_1, \ldots, Z_k$ 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. (12) equals to

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

Combining the above with Eq. (12) concludes the proof. □

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

$$\mathbb{E}[\hat{\ell}_v] = \sum_S \Pr[S]\,\mathbb{E}[\hat{\ell}_v] = \sum_{k=0}^{m} \sum_{S : c_v(S) = k} \Pr[S] \cdot \mathbb{E}[\hat{\ell}_v \mid c_v(S) = k]. \tag{13}$$

From the definition of $\hat{\ell}$, we have $\mathbb{E}[\hat{\ell}_v \mid c_v(S) = 1] = \frac{1}{2m}$ and $\mathbb{E}[\hat{\ell}_v \mid c_v(S) = 0] = 0$. Combining this with Lemma 1 and Eq. (13), we get

$$\begin{aligned} E[\hat{\ell}_v] &= \Pr[c_v = 1] \cdot \frac{1}{2m} + \sum_{k=2}^{m} \Pr[c_v = k] \cdot \frac{k}{m} \cdot 2q_v(1 - q_v) \\ &= \frac{1}{m}\,(\frac{1}{2} - 2q_v(1 - q_v))\ \Pr[c_v = 1] + 2q_v(1 - q_v) \sum_{k=0}^{m} \Pr[c_v = k] \cdot \frac{k}{m} \\ &= \frac{1}{m}\,(\frac{1}{2} - 2q_v(1 - q_v))\ \Pr[c_v = 1] + p_v \cdot 2q_v(1 - q_v)\ , \end{aligned} \tag{14}$$

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 $\ell_v(h_S^{\text{Gini}})$. The proof of the following lemma can be found in [13].

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

Equipped with the expectation of $\hat{\ell}_v$ given in Eq. (14) and the expectation of $\ell_v(h_S^{\text{Gini}})$ given in Lemma 2, 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

$$\mathbb{E}[\hat{\ell}] - \mathbb{E}[\ell(h_S^{\text{Gini}})] \ = \ \mathbb{E}[\sum_v \hat{\ell}_v] - \mathbb{E}[\sum_v \ell_v(h_S^{\text{Gini}})] \ = \ \sum_v (\mathbb{E}[\hat{\ell}_v] - \mathbb{E}[\ell_v(h_S^{\text{Gini}})]) \ . \tag{15}$$

Fix some $v \in V$. From Eq. (14) and Lemma 2 we have

$$\mathbb{E}[\hat{\ell}_v] - \mathbb{E}[\ell_v(h_S^{\text{Gini}})] = (\frac{1}{2} - 2q_v(1 - q_v))(\frac{1}{m}\Pr[c_v = 1] - p_v \Pr[c_v = 0]) \ . \tag{16}$$

Also, it is easy to see that $\frac{1}{m}\Pr[c_v = 1] - p_v \Pr[c_v = 0] = \frac{p_v}{m}\Pr[c_v = 1]$. Plugging this into Eq. (16) we obtain: $\mathbb{E}[\hat{\ell}_v] - \mathbb{E}[\ell_v(h_S^{\text{Gini}})] = (\frac{1}{2} - 2q_v(1-q_v))\frac{1}{m}p_v \Pr[c_v = 1]$. For any q_v we have that $0 \le 2q_v(1-q_v) \le \frac{1}{2}$, which implies the following inequality: $0 \le \mathbb{E}[\hat{\ell}_v] - \mathbb{E}[\ell_v(h_S^{\text{Gini}})] \le \frac{1}{2m}p_v \Pr[c_v = 1] \le \frac{p_v}{2m}$. Summing this over v and using Eq. (15) we conclude that $0 \le \mathbb{E}[\hat{\ell}] - \mathbb{E}[\ell(h_S^{\text{Gini}})] \le \sum_v \frac{p_v}{2m} = \frac{1}{2m}$. □

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. Then, based on Thm. 1, we can easily prove Thm. 2.

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

The proof of this lemma can be found in [13]. 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 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. For h_S^δ we show that the following properties hold:

$$\forall^\delta S \quad \ell(h_S^\delta) = \ell(h_S^{\text{Gini}}) \tag{17}$$

$$\left|\mathbb{E}[\ell(h_S^\delta)] - \mathbb{E}[\ell(h_S^{\text{Gini}})]\right| \le 1/m \tag{18}$$

$$\forall^\delta S \quad \left|\ell(h_S^\delta) - \mathbb{E}[\ell(h_S^\delta)]\right| \le O\left(\ln(m/\delta)/\sqrt{m}\right) . \tag{19}$$

Eq. (17) states that with high confidence, the generalization error of the new hypothesis h_S^δ is exactly equal to the error of h_S^{Gini}. Eq. (18) states that the expectations of the generalization errors of the two hypotheses are close. Finally, Eq. (19) states that the generalization error of h_S^δ is concentrated around its expectation. Combining these three properties and using the triangle inequality, we will be able to bound $|\ell(h_S^{\text{Gini}}) - \mathbb{E}[\ell(h_S^{\text{Gini}})]|$ with high confidence.

We construct a hypothesis h_S^δ that satisfies the three requirements given in Eqs. (17-19) based on Lemma 4 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 McDiardmid'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

$$\rho(\delta, p, m) \ \triangleq\ mp - \sqrt{mp \cdot 3\ln(2/\delta)}. \tag{20}$$

Lemma 4 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 4. *Let $\delta \in (0, 1)$ be a confidence parameter. Then,*

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

The proof is based on lemma 44 from [4] and can be found in [13]. 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 (17-19) hold. From the definition of h_S^δ and Lemma 4 it is clear that Eq. (17) holds. Lemma 5 and Lemma 6 below state that Eq. (18) and Eq. (19) hold. Lemma 7 that follows bounds the concentration of $\ell(h_S^{\text{Gini}})$ using the three properties. The proofs of these three lemmas can be found in [13].

Lemma 5. $\left|\mathbb{E}[\ell(h_S^{Gini})] - \mathbb{E}[\ell(h_S^\delta)]\right| \leq \frac{1}{m}$.

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

Lemma 7. *For all $\delta > 0$ we have $\forall^\delta S$* $\quad |\ell(h_S^{Gini}) - \mathbb{E}[\ell(h_S^{Gini})]| \leq \frac{1}{m} + \frac{12\ln\left(\frac{4m}{\delta}\right)\sqrt{\ln\left(\frac{4}{\delta}\right)}}{\sqrt{2m}}$.

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 7, Thm. 1 and Lemma 3 hold with probability of at least $1-\delta$. We therefore conclude that with probability of at least $1-\delta$,

$$\left|\ell(h_S^{\text{Gini}}) - \hat{\ell}\right| \le |\ell(h_S^{\text{Gini}}) - \mathbb{E}[\ell(h_S^{\text{Gini}})]| + \left|\mathbb{E}[\ell(h_S^{\text{Gini}})] - \mathbb{E}[\hat{\ell}]\right| + \left|\mathbb{E}[\hat{\ell}] - \hat{\ell}\right|$$

$$\le \frac{2}{m} + \frac{12\ln\left(\frac{8m}{\delta}\right)\sqrt{\ln\left(\frac{8}{\delta}\right)}}{\sqrt{2m}} + 12\sqrt{\frac{\ln(\frac{4}{\delta})}{2m}} \;=\; O\left(\frac{\ln(\frac{m}{\delta})\sqrt{\ln(\frac{1}{\delta})}}{\sqrt{m}}\right) .$$

□

Due to lack of space, we omit the proof of Thm. 3 and refer the reader to [13]. 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_v:

$$V_1^\delta = \{v \mid p_v \le 6\ln(2m/\delta)\, m^{-\frac{2}{3}}\}$$
$$V_2^\delta = \{v \mid 6\ln(2m/\delta)\, m^{-\frac{2}{3}} < p_v \le 6\ln(2m/\delta)\, m^{-\frac{1}{2}}\}$$
$$V_3^\delta = \{v \mid 6\ln(2m/\delta)\, m^{-\frac{1}{2}} < p_v\}$$

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

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:

$$|\ell(h_S^{\text{Bayes}}) - \mathbb{E}[\ell(h_S^{\text{Bayes}})]| \;=\; |\sum_{i=1}^{3}\left(\ell_i^\delta(S) - \mathbb{E}[\ell_i^\delta]\right)| \;\le\; A_1 + A_2 + A_3 + A_4 \quad \text{, where}$$

$$\begin{aligned} A_1 &= \left|\; \ell_1^\delta(S) - \mathbb{E}[\ell_1^\delta] \;\right| \\ A_2 &= \left|\; \ell_2^\delta(S) - \mathbb{E}[\ell_2^\delta(S') \mid \kappa(S,S')] \;\right| \\ A_3 &= \left|\; \ell_3^\delta(S) - \mathbb{E}[\ell_3^\delta(S') \mid \kappa(S,S')] \;\right| \\ A_4 &= \left|\; \mathbb{E}[\ell_2^\delta(S') + \ell_3^\delta(S') \mid \kappa(S,S')] - \mathbb{E}[\ell_2^\delta + \ell_3^\delta] \;\right| \quad . \end{aligned} \tag{21}$$

To prove Thm. 4 we bound each of the above terms as follows: First, to bound A_1 (Lemma 8 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 A_2 (Lemma 9 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 A_3 (Lemma 12 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 A_4 (Lemma 13 below). The proof of the first lemma, stated below, is omitted.

Lemma 8. $\forall \delta > 0 \quad \forall^{\delta} S \quad |\ell_1^{\delta}(S) - \mathbb{E}[\ell_1^{\delta}]| \leq \frac{12 \ln\left(\frac{2m}{\delta}\right)}{m^{1/6}} \sqrt{\frac{1}{2} \ln\left(\frac{2}{\delta}\right)}$.

Lemma 9. $\forall \delta > 0 \quad \forall^{\delta} S \quad |\ell_2^{\delta}(S) - \mathbb{E}[\ell_2^{\delta}(S') \mid \kappa(S, S')]| \leq \frac{\sqrt{3 \ln(2m/\delta) \ln(2/\delta)}}{m^{1/4}}$.

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 $\ell_2^{\delta}(S) = \sum_{v \in V_2^{\delta}} \ell_v(h_S^{\text{Bayes}})$ is a sum of independent random variables and 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,

$$\Pr[|\ell_2^{\delta}(S) - \mathbb{E}[\ell_2^{\delta}(S') \mid \kappa(S, S')]| \geq t] \;\leq\; 2e^{-2t^2 / \sum_{v \in V_2^{\delta}} p_v^2} . \tag{22}$$

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

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

Plugging the above into Eq. (22) we get that

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

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

So far, we have bounded the terms A_1 and A_2. In both of these cases, we utilized 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 A_3. 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', i.e. 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 10. *Let $\bar{Z} = (Z_1, \ldots, Z_k)$ be a sequence of i.i.d. binary random variables where $\Pr[Z_i = 1] = q$ for all i, and assume that $q \geq \frac{1}{2}$. Then,*

$$\Pr[\sum_i Z_i \leq k/2] \;\leq\; e^{-2(q - \frac{1}{2})^2 k} \quad \textit{and} \quad (2q - 1) \Pr[\sum_i Z_i \leq k/2] \;\leq\; \frac{1}{\sqrt{e\,k}} \; .$$

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 utilize the inequality $2xe^{-2x^2 k} \leq 1/\sqrt{e\,k}$, which holds for all $x \geq 0$ and $k > 0$. □

Based on the above lemma, we now bound A_3. 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 11. *Under the same assumptions of Lemma 10. Let $f(\bar{Z})$ be the function*

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

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

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

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

Using the fact that (α, β, γ) is in the probability simplex we immediately obtain that $|f(\bar{z}) - \mathbb{E}[f(\bar{Z})]| \le (2\,q-1)$. If $2\,q-1 \le \sqrt{2\,\ln(1/\delta)\,/k}$ then the bound in the lemma clearly holds. Therefore, from now on we assume that $2\,q-1 > \sqrt{2\,\ln(1/\delta)\,/k}$. In this case, using the first inequality of Lemma 10 we have that $\beta + \gamma \;\le\; e^{-2(q-\frac{1}{2})^2 k} \;\le\; \delta$. Therefore, $1 - \delta < \alpha$, and so with probability of at least $1-\delta$ we have that

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

Applying the second inequality of Lemma 10 on the right-hand side of the above inequality we get that $|f(\bar{Z}) - \mathbb{E}[f(\bar{Z})]| \;\le\; \sqrt{1/ek} \;\le\; \sqrt{2\ln(1/\delta)/ek}$, where the last inequality holds since we assume that $\delta \le e^{-1/2}$. □

Equipped with the above lemma we are now ready to bound A_3.

Lemma 12. *If $m \ge 4$ then* $\quad \forall^{(2\delta)} S \quad |\ell_3^\delta(S) - \mathbb{E}[\ell_3^\delta(S') \mid \kappa(S,S')]| \;\le\; 1/m^{\frac{1}{4}}$.

Proof. Recall that $\ell_3^\delta(S) \;=\; \sum_{v \in V_3^\delta} \ell_v(S)$. $m \;\ge\; 4$, hence $\delta/m \le 1/m \le e^{-1/2}$. Choose $v \in V_3^\delta$ and without loss of generality assume that $q_v \;\ge\; 1/2$. Thus, from Lemma 11 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 $S(v)$: $|\ell_v(S) - \mathbb{E}[\ell_v(S')|\kappa(S,S')]| \;\le\; p_v \sqrt{\frac{2\ln(m/\delta)}{e \cdot c_v(S)}}$. By the definition of V_3^δ and Lemma 4, $\forall^\delta S,\ \forall v \in V_3^\delta,\ c_v(S) \ge \rho(\delta/m, p_v, m)$. Using the fact that ρ is monotonically increasing with respect to p_v it is possible to show (see [13]) that $\rho(\delta/m, p_v, m) \;\ge\; 2\ln(m/\delta)\, m^{1/2}$ for all $v \in V_3^\delta$ for $m \ge 4$. Therefore, $|\ell_v(S) - \mathbb{E}[\ell_v(S')|\kappa(S,S')]| \;\le\; p_v\, m^{-1/4}$. Using a union bound, we obtain that $\forall^{(2\delta)} S \quad \forall v \in V_3^\delta \quad |\ell_v(S) - \mathbb{E}[\ell_v(S')|\kappa(S,S')]| \;\le\; p_v\, m^{-1/4}$. 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. □

Lastly, we bound A_4 in the next lemma. See [13] for the proof.

Lemma 13. *For $m \ge 8$,*

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

5 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} was proposed and its convergence was analyzed. We showed that the expected error of h_S^{Bayes} is optimal and that its concentration is weaker than that of h_S^{Gini}. 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. This is especially important when ranking groups of categorical features. 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.

Acknowledgments

We would like to thank Yishay Mansour for helpful discussions.

References

1. A. Antos, L. Devroye, and L. Gyorfi. Lower bounds for bayes error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(7), 1999.
2. A. Antos and I. Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Struct. Algorithms*, 19(3-4), 2001.
3. R. Lopez de Mantaras. A distance-based attribute selection measure for decision tree induction. *Machine Learning Journal*, 1991.
4. E. Drukh and Y. Mansour. Concentration bounds for unigrams language model. *JMLR*, 2005.
5. I.J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 1953.
6. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.
7. M. Kearns and Y. Mansour. On the boosting ability of top-down decision tree learning algorithms. In *STOC*, 1996.
8. D.A. McAllester and R.E. Schapire. On the convergence rate of good-turing estimators. In *COLT*, 2000.
9. C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, pages 148–188, 1989.
10. J. Mingers. An empirical comparison of selection measures for decision-tree induction. *Machine Learning*, 1989.
11. Tom M. Mitchell. *Machine Learning*. McGraw Hill, 1997.
12. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
13. S. Sabato and S. Shalev-Shwartz. Prediction by categorical features. Technical report, 2007.
14. K. Torkkola. *Feature Extraction, Foundations and Applications*, chapter Information-Theoretic Methods. Springer, 2006.