On the Necessity of Irrelevant Variables

David P. Helmbold DPH@SOE.UCSC.EDU
Computer Science Department, UC Santa Cruz

Philip M. Long PLONG@GOOGLE.COM
Google

Abstract

This work explores the effects of relevant and irrelevant boolean variables on the accuracy of classifiers. The analysis uses the assumption that the variables are conditionally independent given the class, and focuses on a natural family of learning algorithms for such sources when the relevant variables have a small advantage over random guessing. The main result is that algorithms relying predominately on irrelevant variables have error probabilities that quickly go to 0 in situations where algorithms that limit the use of irrelevant variables have errors bounded below by a positive constant. We also show that accurate learning is possible even when there are so few examples that one cannot determine with high confidence whether or not any individual variable is relevant.

1. Introduction

The purpose of this paper is to provide an illustrative analysis that isolates the effects of relevant and irrelevant variables on a classifier's accuracy. We show that, if variables complement one another (formalized with the usual assumption of conditional independence given the class label), then relevant variables can do much more good than irrelevant variables do harm. In many natural settings the individual variables are only weakly associated with the class label. This can happen when a lot of measurement error is present, as is seen in microarray data. In these settings it can be worthwhile for the classifier to use what at one time might have been thought an excessive number of variables, even if only an small fraction of them are relevant.

Over the past decade or so, a number of empirical and theoretical findings have challenged the traditional rule of thumb described by Bishop (2006) as follows.

> One rough heuristic that is sometimes advocated is that the number of data points should be no less than some multiple (say 5 or 10) of the number of adaptive parameters in the model.

The Support Vector Machine literature (see (Vapnik, 1998)) views algorithms that compute apparently complicated functions of a given set of variables as linear classifiers applied to an expanded, even infinite, set of features. These empirically perform well on test data, and theoretical accounts have been given for this. Boosting and Bagging algorithms also generalize well, despite combining large numbers of simple classifiers – even if the number of such "base classifiers" is much more than the number of training examples (Quinlan, 1996; Breiman, 1998; Schapire et al., 1998). This is despite the fact that Friedman et al. (2000) showed the behavior of such classifiers is closely related to performing logistic regression on a potentially vast set of features (one for each possible decision tree, for example).

Similar effects are sometimes found even when the features added are restricted to the original "raw" variables. Figure 1, which is reproduced from Tibshirani et al. (2002), is one example. The curve labelled "te" is the test-set error, and this error is plotted as a function of the number of features selected by the Shrunken Centroids algorithm. The best accuracy is obtained using a classifier that depends on the expression level of well over 1000 genes, despite the fact that there are only a few dozen training examples.

It is impossible to tell if most of the variables used by the most accurate classifier in Figure 1 are irrelevant. However, we do know which variables are relevant and irrelevant in synthetic data (and can generate as many test examples as desired). Figure 2 concerns a simple algorithm applied to a simple source. Each

Appearing in *Proceedings of the 28^{th} International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

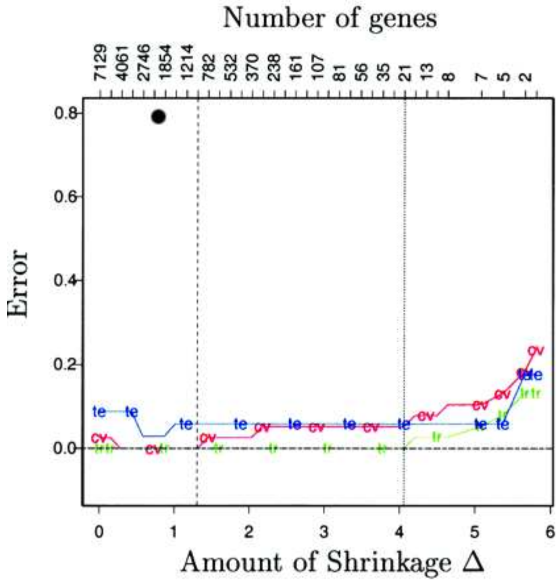

Figure 1. This graph is reproduced from Tibshirani et al. (2002). For a microarray dataset, the training error, test error, and cross-validation error are plotted as a function both of the number of features (along the top) included in a linear model and a regularization parameter Δ (along the bottom).

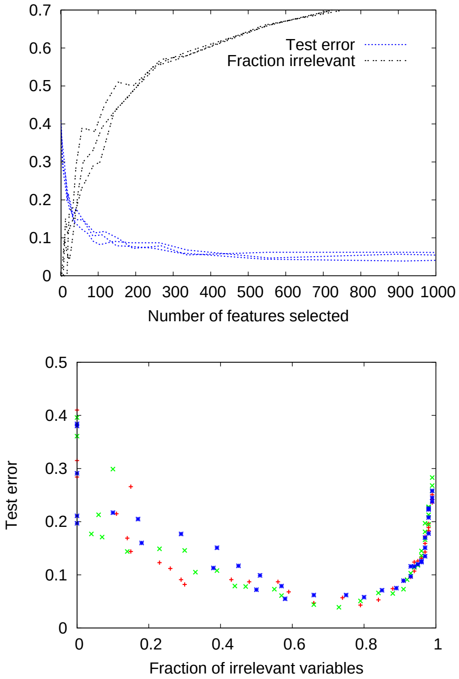

Figure 2. Top: Test error (blue) and fraction of irrelevant variables (black dashed) as a function of the number of features. Bottom: Scatter plot of test error rates (vertical) and fraction of irrelevant variables (horizontal).

of two classes is equally likely, and there are 1000 relevant variables, 500 of which agree with the class label with probability $1/2 + 1/10$, and 500 which disagree with the class label with probability $1/2 + 1/10$. Another 99000 variables are irrelevant. The algorithm is equally simple: it has a parameter β, and outputs the majority vote over those features (variables or their negations) that agree with the class label on a $1/2+\beta$ fraction of the training examples. Plots are provided for three runs of this algorithm with 100 training examples, and 1000 test examples. Both the accuracy of the classifier and the fraction of relevant variables are plotted against the number of variables used in the model, for various values of β. Each time, the best accuracy is achieved when an overwhelming majority of the variables used in the model are irrelevant, and those models with few ($< 25\%$) irrelevant variables perform far worse. Furthermore, the best accuracy is obtained with a model that uses many more variables than there are training examples. Also, accuracy over 90% is achieved even though the correlation of the individual variables with the class label is so weak, and the number of training examples is so small, that it is impossible, for any individual feature, to tell confidently whether that feature is relevant or not.

Assume classifier f consists of a vote over n variables that are conditionally independent given the class label. Let k of the variables agree with the class label with probability $1/2+\gamma$, and the remaining $n-k$ variables agree with the label with probability $1/2$. Then the probability that f is incorrect is at most

$$\exp\left(\frac{-2\gamma^2 k^2}{n}\right) \tag{1}$$

(as shown in Section 3). The error bound decreases exponentially in the *square* of the number of relevant variables. The competing factor increases only *linearly* with the number of irrelevant variables. Thus, a very accurate classifier can be obtained with a feature set consisting predominately of irrelevant variables.

In Section 4 we consider learning from training data generated from a source in which N boolean variables are conditionally independent given the class label, and $N - K$ of them are also independent of the label, agreeing with it with probability $1/2$. The K relevant variables either agree with the label with probability $1/2 + \gamma$ or with probability $1/2 - \gamma$. Whereas Equation (1) bounded the error as a function of the number of relevant and irrelevant variables in the *model*, we are now discussing the number of relevant and irrelevant variables in the *data*. We analyze an algorithm that chooses a value of β and outputs a majority vote over all features that agree with the class label on at least $1/2+\beta$ of the training examples (as before, each feature is either a variable or its negation). We show that if $\beta \leq \gamma$ and the algorithm is given m training examples, then the probability that it makes an incorrect prediction on an independent test example is at most

$$(1+o(1)) \exp\left(-2\gamma^2 K \left(\frac{[1-8e^{-2(\gamma-\beta)^2 m}-\gamma)]_+^2}{1+8(N/K)e^{-2\beta^2 m}+\gamma}\right)\right), \tag{2}$$

where $[z]_+ \overset{\text{def}}{=} \max\{z, 0\}$. (Throughout the paper, the "big Oh" and other asymptotic notation will be for the case where γ is small, K/γ is large, and N/K is large. If K is not large relative to $1/\gamma^2$, even the Bayes optimal classifier is not accurate.) If $\beta = \gamma/2$, this implies a bound of

$$(1+o(1)) \exp\left(-2\gamma^2 K \left(\frac{1-O(e^{-\gamma^2 m/2})}{1+O\left((N/K)e^{-\gamma^2 m/2}\right)}\right)\right).$$

When $\beta \leq \gamma/2$ and $m \geq c/\gamma^2$, we also show that the error probability is at most

$$(1+o(1)) \exp\left(-\Omega\left(\gamma^2 K^2/N\right)\right). \tag{3}$$

If $N = o(\gamma^2 K^2)$, this error probability goes to zero. With only $\Theta(1/\gamma^2)$ examples, an algorithm cannot even tell with high confidence whether a relevant variable is positively or negatively associated with the class label, much less solve the more difficult problem of determining whether or not a variable is relevant. Indeed, this error bound is also achieved using $\beta = 0$, when, for each variable X_i, the algorithm includes either X_i or its negation in the vote.[1]

Our upper bounds illustrate the potential rewards for algorithms that are "inclusive", using many of the available variables in their classifiers. We also prove some lower bounds that illustrate the potential cost when algorithms are "exclusive". We say that a policy for setting β as a function of γ is λ-exclusive if the expected number of relevant variables in the resulting model divided by its expected total number of variables is at least λ. We show that any λ-exclusive policy has an error probability at least a constant as K and N/K go to infinity and γ goes to 0 in such a way that the error rate obtained by the more "inclusive" setting $\beta = \gamma/2$ goes to 0. In particular, no λ-exclusive algorithm can achieve a bound like (3).

Relationship to Previous Work For the sources studied in this paper, there is a linear classifier that classifies most random examples correctly with a large margin, i.e. most examples are not close to the decision boundary. The main motivation for our analysis was to understand the effects of relevant and irrelevant variables on generalization, but it is interesting to note that we get meaningful bounds in the extreme case that $m = \Theta(1/\gamma^2)$, whereas the margin-based bounds that we know (such as Schapire et al. (1998); Koltchinskii & Panchenko (2002); Dasgupta & Long (2003); Wang et al. (2008)) are vacuous in this case. (Since these other bounds hold more generally, their overall strength is incomparable to our results.) Ng & Jordan (2001) showed that the Naive Bayes algorithm (which ignores class-conditional dependencies) converges relatively quickly, justifying its use when there are few examples. (Their bound for Naive Bayes is also vacuous when $m = \Theta(1/\gamma^2)$.) Bickel & Levina (2004) studied the case in which the class conditional distributions are Gaussians, and showed how an algorithm which does not model class conditional dependencies can perform nearly optimally in this case, especially when the number of variables is large. Bühlmann & Yu (2002) analyzed the variance-reduction benefits of Bagging with primary focus on the benefits of the smoother classifier that is obtained when ragged classifiers are averaged. As such it takes a different form than our analysis.

Our analysis demonstrates that certain effects are possible, but how important this is depends on how closely natural learning settings resemble our theoretical setting and the extent to which our analysis can be generalized. The conditional independence assumption is one way to express the intuitive notion that variables are not too redundant. A limit on the redundancy is needed for results like ours since, for example, a collection of $\Theta(k)$ perfectly correlated irrelevant variables would swamp the votes of the k relevant variables. On the other hand, many boosting algorithms minimize the potential for this kind of effect by choosing features in later iterations that make errors on different examples then the previously chosen features. One relaxation of the conditional independence assumption is to allow each variable to conditionally depend on a limited number r of other variables, as is done in the formulation of the Lovasz Local Lemma (see Alon et al. (1992)). To illustrate the robustness of the effects analyzed here, we generalize (1) to this case in Section 5. There we prove a bound of $c(r+1)\exp\left(\frac{-2\gamma^2 k^2}{n(r+1)}\right)$ when each variable depends on most r others. There are a number of ways that one could imagine relaxing the conditional independence assumption while still prov-

[1] To be precise, the algorithm includes each variable or its negation when $\beta = 0$ and m is odd, and includes both the variable and its negation when m is even and the variable agrees with the class label exactly half the time. But, any time both a variable and its negation are included, their votes cancel. We will always use the smaller equivalent model obtained by removing such canceling votes.

ing theorems of a similar flavor.

Another obvious direction for generalization is to relax the strict categorization of variables into irrelevant and $(1/2 + \gamma)$-relevant classes. We believe that many extensions of this work with different coverage and interpretability tradeoffs are possible. For example, our proof techniques give similar theorems when each relevant variable has a probability between $1/2 + \gamma/2$ and $1/2 + \gamma$ of agreeing with the class label. Here we concentrate on some of the cleanest and simplest settings in order to focus attention on the main ideas.

We state some useful tail bounds in the next section, and Section 3 analyzes the error of simple voting classifiers. Section 4.1 gives bounds on the expected error of hypotheses learned from training data while 4.2 shows that, in certain situations, any exclusive algorithm must have high error while the error of some inclusive algorithms goes to 0. In section 5 we bound the accuracy of voting classifiers under a weakened independence assumption.

2. Tail bounds

These bounds all assume that $U_1, U_2, \ldots, U_\ell$ are ℓ independent $\{0,1\}$-valued random variables and $U = \sum_{i=1}^n U_i$. We start with some upper bounds.

The Hoeffding bound, see Pollard (1984):

$$\Pr\left[\frac{1}{\ell}U - \mathbb{E}\left(\frac{1}{\ell}U\right) \geq \gamma\right] \leq e^{-2\gamma^2\ell}. \tag{4}$$

The Chernoff bound, see Angluin & Valiant (1979); Motwani & Raghavan (1995), for any $\eta > 0$:

$$\Pr[U > (1+\eta)\mathbb{E}(U)] < \exp\left(-(1+\eta)\mathbb{E}(U)\ln\left(\frac{1+\eta}{e}\right)\right) \tag{5}$$

For any $0 \leq \eta \leq 4$ (see Appendix A.1):

$$\Pr[U > (1+\eta)\mathbb{E}(U)] < \exp\left(-\eta^2\mathbb{E}(U)/4\right). \tag{6}$$

For any $0 < \delta \leq 1$ (see Appendix A.2):

$$\Pr[U > 4\mathbb{E}(U) + 3\ln(1/\delta)] < \delta. \tag{7}$$

We will also need lower bounds on the tails of the distribution. Here $c_1, ..., c_7$ are absolute constants.

If $\Pr(U_i = 1) = 1/2$ for all i, $\eta > 0$, and $\ell \geq 1/\eta^2$ then (see Appendix A.3):

$$\Pr\left[\frac{1}{\ell}U - \frac{1}{\ell}\mathbb{E}(U) \geq \eta\right] \geq \frac{c_1}{\eta\sqrt{\ell}}\exp\left(-2\eta^2\ell\right) - \frac{c_2}{\sqrt{\ell}}. \tag{8}$$

Slud's Theorem (Slud, 1977), if $0 \leq \eta \leq c_3$ and $\Pr[U_i = 1] = 1/2 + \eta$ for all i then:

$$\Pr\left[\frac{1}{\ell}U < 1/2\right] \geq c_4 e^{-c_5\eta^2\ell}. \tag{9}$$

If $\Pr[U_i = 1] = 1/2$ for all i, then for all $0 \leq \eta \leq 1/2$ (see Appendix A.4):

$$\Pr\left[\frac{1}{\ell}U - \frac{1}{\ell}\mathbb{E}(U) \geq \eta\right] \geq c_6 e^{-c_7\eta^2\ell}. \tag{10}$$

3. The accuracy of models containing relevant and irrelevant variables

In this section we analyze the accuracy of the models (hypotheses) produced by the algorithms in Section 4. Each example is represented by a vector of N binary *variables* and a class designation. We assume a simple generative model with parameter $\gamma > 0$ and:

- random $\{0,1\}$ class designations: both classes are equally likely;
- the K *relevant* variables are equal to the class designation either with probability $1/2+\gamma$ or with probability $1/2 - \gamma$;
- the $N - K$ *irrelevant* variables are equal to the class label with probability $1/2$;
- all variables are conditionally independent given the class designation.

Which variables are relevant and whether each one is positive or negatively correlated with the class designations are chosen arbitrarily ahead of time.

A *feature* is either a variable or its complement. The $2(N - K)$ *irrelevant* features come from the irrelevant variables, the K *relevant* features agree with the class labels with probability $1/2 + \gamma$, and the K *misleading* features agree with the class labels with probability $1/2 - \gamma$.

We consider models $\mathcal{M}$ predicting with a majority vote over a subset of the features. We use n for the total number of features in model $\mathcal{M}$, k for the number of relevant features, and ℓ for the number of misleading features (leaving $n-k-\ell$ irrelevant features). Since the votes of a variable and its negation "cancel out," we assume without loss of generality that models include at most one feature for each variable.

Theorem 1. *Let $\mathcal{M}$ be a majority vote of n features, k of which are relevant and ℓ of which are misleading (and $n - k - \ell$ are irrelevant). If $\ell \leq k$,* *the probability that $\mathcal{M}$ predicts incorrectly is at most* $\exp\left(\frac{-2\gamma^2(k-\ell)^2}{n}\right)$.

Proof: Model $\mathcal{M}$ predicts incorrectly only when at most half of its features are correct. The expected fraction of correct voters is $1/2 + \frac{\gamma(k-\ell)}{n}$, so, for $\mathcal{M}$'s prediction to be incorrect, the fraction of correct voters must be at least $\gamma(k-\ell)/n$ less than its expectation. Applying (4), this probability is at most $\exp\left(\frac{-2\gamma^2(k-\ell)^2}{n}\right)$. □

The next corollary shows that even models where most of the features are irrelevant can be highly accurate.

Corollary 1. *If γ is constant, $k-\ell = \omega(\sqrt{n})$, and $k = o(n)$, then the accuracy of the model approaches* 100% *while its fraction of irrelevant variables approaches* 1.

The hypothesis of Corollary 1 is satisfied, for example, when $\gamma = 1/4$, $k = 2n^{2/3}$ and $\ell = n^{2/3}$.

4. Learning

We now consider the problem of learning a model $\mathcal{M}$ from data. We assume that the algorithm receives m i.i.d. examples generated as described in Section 3. One test example is independently generated from the same distribution, and we evaluate the algorithm's *expected error*, the probability over training set and test example that its model makes an incorrect prediction on the test example (the "prediction model" of Haussler et al. (1994)).

We define $\mathcal{M}_\beta$ to be the majority vote[2] of all features that equal the class label on at least $1/2+\beta$ of the training examples. To keep the analysis as clean as possible, our results apply to algorithms that chose β as a function of N, K, γ, and training set size m, and then predict with $\mathcal{M}_\beta$.

4.1. The accuracy of $\mathcal{M}_\beta$

This section proves two theorems bounding the expected error rates of learned models. We note that the Bayes Optimal predictor for our generative model is a majority vote of the K relevant features, and has an error rate bounded by $e^{-2\gamma^2 K}$ (a bound as tight as the Hoeffding bound). We also use Hoeffding bounds in our results and will state them in a similar form.

Theorem 2. *If $0 \le \beta \le \gamma$, the expected error rate of $\mathcal{M}_\beta$ is at most*

$$(1+o(1)) \exp\left(-2\gamma^2 K \left(\frac{[1 - 8e^{-2(\gamma-\beta)^2 m} - \gamma]_+^2}{1 + 8(N/K)e^{-2\beta^2 m} + \gamma}\right)\right).$$

Our proof of Theorem 2 starts with lemmas bounding the number of misleading, irrelevant, and relevant features in $\mathcal{M}_\beta$.

Lemma 1. *With probability at least $1-\delta$, the number of misleading features in $\mathcal{M}_\beta$ is at most $4Ke^{-2(\gamma+\beta)^2 m} + 3\ln(1/\delta)$.*

Proof: For a particular misleading feature L in $\mathcal{M}_\beta$, Algorithm A must overestimate the probability that $L = Y$ by $\beta+\gamma$. Applying (4), this happens with probability at most $e^{-2(\beta+\gamma)^2 m}$, so the expected number of misleading features in $\mathcal{M}_\beta$ is at most $Ke^{-2(\beta+\gamma)^2 m}$. Since each misleading feature is associated with a different independent variable, we can apply (7) with $\mathbb{E}(U) = Ke^{-2(\beta+\gamma)^2 m}$ to get the desired result. □

Lemma 2. *With probability at least $1-2\delta$, the number of irrelevant features in $\mathcal{M}_\beta$ is at most $8Ne^{-2\beta^2 m} + 6\ln(1/\delta)$.*

Proof: Separately bound the number of positive and negative irrelevant features in the model as in Lemma 1. With probability at least $1-2\delta$ the total number of irrelevant features is at most the sum of the bounds. □

Lemma 3. *With probability at least $1-\delta$, the number of relevant features in $\mathcal{M}_\beta$ is at least $K - 4Ke^{-2(\gamma-\beta)^2 m} - 3\ln(1/\delta)$.*

Proof: Bound the number of relevant features *not* in the model as in Lemma 1. The number of relevant features remaining in the model is at least K minus this bound. □

Lemma 4. *The probability that $\mathcal{M}_\beta$ makes an error is at most*

$$\exp\left(\frac{-2\gamma^2\left[K - 8Ke^{-2(\gamma-\beta)^2 m} - 6\ln(1/\delta)\right]_+^2}{K + 8Ne^{-2\beta^2 m} + 6\ln(1/\delta)}\right) + 4\delta.$$

for any $\delta > 0$ and $0 \le \beta \le \gamma$.

Proof: Applying Theorem 1 with the lower bound on $k-\ell$ from Lemmas 1 (under-approximating $(\gamma+\beta)^2$ with $(\gamma-\beta)^2$) and 3 and upper bounding n with K plus the bound of Lemma 2 shows that, with probability at least $1-4\delta$, the error is at most the first term. (Note that the lemma is vacuous unless the bound on $k-\ell$ is positive.) □

Proof (of Theorem 2): Using $\delta = \exp(-\gamma K/6)$ in Lemma 4 and simplifying, the expected error rate of

[2] If $\mathcal{M}_\beta$ is empty, then any default prediction, such as 1, will do.

$\mathcal{M}_\beta$ is at most

$$\exp\left(\frac{-2\gamma^2 K\left[1-8e^{-2(\gamma-\beta)^2 m}-\gamma\right]_+^2}{1+\frac{8N}{K}e^{-2\beta^2 m}+\gamma}\right)+4e^{-\gamma K/6}.$$

The first term is at least $e^{-2\gamma^2 K}$ and $e^{-\gamma K/6} = o(e^{-2\gamma^2 K})$ as γK gets large, implying the bound

$$(1+o(1))\exp\left(\frac{-2\gamma^2 K\left[1-8e^{-2(\gamma-\beta)^2 m}-\gamma\right]_+^2}{1+\frac{8N}{K}e^{-2\beta^2 m}+\gamma}\right).$$

□

Theorem 3. *Suppose A uses a β where $0 \le \beta \le c\gamma$ for a constant $c \in [0,1)$. There is a constant c' (depending only on c) such that, if $m = c'/\gamma^2$, the error of A is at most $(1+o(1))\exp\left(-\frac{\gamma^2 K^2}{N}\right)$. Furthermore, if $m = \omega(1/\gamma^2)$, the error of A is at most $(1+o(1))\exp\left(\frac{-(2-o(1))\gamma^2 K^2}{N}\right)$.*

Proof. Since $\mathcal{M}_\beta$ contains at most N features, the bound of Lemma 4 with the denominator replaced by N also holds. Continuing as in the proof of Theorem 2 and using $(\gamma-\beta)^2 \ge (1-c)^2\gamma^2$ yields a bound of

$$(1+o(1))\exp\left(-2\gamma^2 K^2\left(\frac{[1-O(e^{-2(1-c)^2\gamma^2 m})]_+^2}{N}\right)\right).$$

Setting $m = c'/\gamma^2$ for a large enough value of c' suffices to make the $[\cdots]_+^2$ term at least $1/2$, and when $m = \omega(1/\gamma^2)$ it is $1-o(1)$. □

Note that Theorem 3 includes non-trivial bounds for $\mathcal{M}_0$ that votes all N variables (for odd sample size m).

4.2. Lower bound

In this subsection, we show that any algorithm with an error guarantee like Theorem 3 must include many irrelevant features in its model.

Definition 1. *Let $\mathcal{R}$ be the set of relevant features, and recall that $\mathcal{M}_\beta$ is the set of features in the model (which depends on the random training data). We say that an algorithm A is $\underline{\lambda\text{-exclusive}}$ if for every positive N, K, γ, and m, A uses a $\beta \in [0,1/2]$ such that $\frac{\mathbb{E}(|\mathcal{M}_\beta \cap \mathcal{R}|)}{\mathbb{E}(|\mathcal{M}_\beta|)} \ge \lambda$.*

Our main lower bound theorem is the following.

Theorem 4. *There are absolute positive constants c_1 and $c_2 \in [0,1)$, and functions $K(\gamma)$ and $N(\gamma)$ tying K and N to γ such that if $m = c_1/\gamma^2$ the following hold.*

If $\lambda > 0$, then the error rate of any λ-exclusive A is at least a positive constant for all small enough γ.

Inclusive A using models $\mathcal{M}_\beta$ with $\beta \le c_2\gamma$ have error rates that goes to zero super-polynomially fast (in $1/\gamma$).

We will prove Theorem 4 using a series of lemmas. The first is a lower bound in terms of the number of relevant variables.

Lemma 5. *There are absolute positive constants c_1, c_2, c_3 such that if $\gamma \in [0, c_1]$ then any model with k relevant features has an error probability at least $c_2 e^{-c_3\gamma^2 k}$.*

Proof: If there are no irrelevant or misleading features, applying (9) yields the Lemma. Adding irrelevant or misleading features only increases the error probability. □

The next step is a lower bound on the number of irrelevant variables.

Lemma 6. *Suppose $\gamma \le 1/4$, $N \ge 2K$, and $\beta \ge 0$. The expected number of irrelevant features in $\mathcal{M}_\beta$'s model is at least $N\left(c_1\exp\left(-c_2(\beta/\gamma)^2\right)\right)$ and also at least*

$$N\left(\frac{c_3}{\beta\sqrt{m}}\exp\left(-2\beta^2 m\right)-\frac{c_4}{\sqrt{m}}\right)$$

where c_1, c_2, c_3 and c_4 are absolute positive constants.

Proof: A positive irrelevant feature is selected if it agrees with the class label at least $1/2+\beta$ of the time. Applying Bound (10) and linearity of expectation, gives a lower bound of $(N-K)\left(c_6\exp\left(-c_7\beta^2 m\right)\right)$ and the assumptions ensure $N-K \ge N/2$ and $m = c/\gamma^2$. The second part uses Bound (8) instead of (10). □

We now upper bound the number of relevant variables.

Lemma 7. *If $\beta \ge \gamma$, the expected number of relevant variables in A_β's model is at most $Ke^{-2(\beta-\gamma)^2 m}$.*

Proof: Use (4) to bound the probability that a relevant feature agrees with the class label $\beta-\gamma$ more often than its expected fraction of times and the linearity of expectation. □

Lemma 8. *The ratio $\frac{\beta}{\gamma} = \Omega\left(\min\left\{\ln\left(\frac{N}{K}\right), \sqrt{\ln\frac{1}{\gamma}}\right\}\right)$ (as $1/\gamma$ and N/K go to infinity) if $\frac{\mathbb{E}(|\mathcal{M}_\beta \cap \mathcal{R}|)}{\mathbb{E}(|\mathcal{M}_\beta|)} \ge \lambda$ for constant $\lambda \in (0,1)$.*

Proof: Let $\mathcal{I}$ be the set of irrelevant features. If $\frac{\mathbb{E}(|\mathcal{M}_\beta\cap\mathcal{R}|)}{\mathbb{E}(|\mathcal{M}_\beta|)} \ge \lambda$ then $\frac{\mathbb{E}(|\mathcal{M}_\beta\cap\mathcal{R}|)}{\mathbb{E}(|\mathcal{M}_\beta\cap\mathcal{R}|)+\mathbb{E}(|\mathcal{M}_\beta\cap\mathcal{I}|)} \ge \lambda$ which implies $\frac{\mathbb{E}(|\mathcal{M}_\beta\cap\mathcal{R}|)}{\mathbb{E}(|\mathcal{M}_\beta\cap\mathcal{I}|)} \ge \frac{\lambda}{1-\lambda}$.

We first show $\beta > \gamma$. Assume to the contrary that $\beta \le \gamma$. Use Lemma 6 and note there are at most

K relevant features, so $\dfrac{K}{cN} \geq \dfrac{\lambda}{1-\lambda}$ for an absolute constant c. This is contradicted for a large enough value of N/K, so $\beta > \gamma$ when N/K is large.

We apply Lemma 6 and 7 getting, for absolute positive constants c and c', that

$$\frac{Ke^{-2(\beta-\gamma)^2 m}}{N\left(\frac{c}{\beta\sqrt{m}}\exp(-2\beta^2 m) - \frac{c'}{\sqrt{m}}\right)} \geq \frac{\mathbb{E}(|\mathcal{M}_\beta \cap \mathcal{R})|}{\mathbb{E}(|\mathcal{M}_\beta \cap \mathcal{I}|)} \geq \frac{\lambda}{1-\lambda}.$$

Solving for the $\exp(-2\beta^2 m)$ term and setting $\kappa = \frac{(1-\lambda)K}{N\lambda}$ and $m = c''/\gamma^2$ gives

$$\exp(-2c''(\beta/\gamma)^2) \leq \frac{\sqrt{c''}\beta\kappa}{c\gamma} e^{-2c''(\beta/\gamma-1)^2} + \frac{c'\beta}{c}.$$

This implies at least one of:

$$\begin{aligned} \exp(-2c''(\beta/\gamma)^2) &\leq \frac{\sqrt{c''}\beta\kappa}{2c\gamma} e^{-2c''(\beta/\gamma-1)^2} \\ \text{or } \exp(-2c''(\beta/\gamma)^2) &\leq \frac{c'\beta}{2c}. \end{aligned} \tag{11}$$

The first of these implies, after taking logs and some algebra (e.g. canceling the (β/γ^2) terms), that $\dfrac{\beta}{\gamma} \geq \left(\dfrac{\ln\left(\frac{2c\gamma}{\sqrt{c''}\beta\kappa}\right)}{4c''} - \dfrac{1}{2}\right)$. Note the RHS is increasing in β/γ while the LHS is decreasing in β/γ. Furthermore, setting $\beta/\gamma = \ln(1/\kappa)$ leads to a contradiction as $\kappa \to 0$. Therefore $r = \Omega\left(\ln(1/\kappa)\right)$ and $\beta/\gamma = \Omega\left(\ln(N/K)\right)$.

We now turn to the $\exp(-2c''(\beta/\gamma)^2) \leq \frac{c'\beta}{2c}$ case. Note that the RHS is decreasing in β and the LHS is increasing in β. Therefore any value of β where this fails gives a lower bound on β. If $\beta = \gamma\sqrt{\ln(1/\gamma)/4c''}$ then the inequality becomes $\gamma^{1/2} \leq \dfrac{c'\gamma\sqrt{\ln(1/\gamma)}}{4c\sqrt{c''}}$ which fails for small enough γ. So $\beta/\gamma = \Omega\left(\sqrt{\ln(1/\gamma)}\right)$, completing the proof. □

Proof (of Theorem 4): Set $K = \frac{1}{\gamma^2}\exp((\ln(1/\gamma))^{1/3})$ and $N = K\exp((\ln(1/\gamma)^{1/4}))$. Theorem 3 now implies that the probability of error for $\mathcal{M}_\beta$ with $\beta \leq c_2\gamma$ is $\exp(-\Theta(\exp(\ln(1/\gamma)^{1/3})/\exp(\ln(1/\gamma)^{1/4})))$.

Now let us consider any λ-exclusive algorithm. The Chernoff bound together with Lemma 7 implies that there is a constant c such that with probability $3/4$, the number of relevant variables in $A_{\beta(\gamma)}$ is at most $Ke^{-c(\beta/\gamma-1)^2}$. Lemma 8 implies that this is at most $K\exp(-\Omega(\min\{\ln(N/K)^2, \ln(1/\gamma)\})) = \frac{1}{\gamma^2}\exp(-\Omega(\sqrt{\ln(1/\gamma)}\}))$. Applying Lemma 5 completes the proof. □

5. Conditionally dependent variables

Assume that there is a degree-r graph G whose nodes are variables, and such that, conditioned on the label, each variable is independent of all variables not connected to it by an edge in G. Assume that k variables agree with the label with probability $1/2+\gamma$, and the $n-k$ agree with the label with probability $1/2$. Let us say that a source like this *has r-local dependence* (we will also overload "r-local dependence" to refer to the constraint on each of the conditional distributions).

Theorem 5. *For a source that has r-local dependence for $r \leq n/2$, the probability that f predicts incorrectly is at most $c(r+1)\exp\left(\frac{-2\gamma^2 k^2}{n(r+1)}\right)$ for a positive constant c.*

Proof Sketch: Replace the Hoeffding bound in the proof of Theorem 1 with a similar bound for r-local dependence due to Pemmaraju (2001) . □

Acknowledgements

We thank Aravind Srinivasan for his help.

References

Alon, N., Spencer, J. H., and Erdös, P. *The Probabilistic Method*. Wiley, 1992.

Angluin, D. and Valiant, L. Fast probabilistic algorithms for Hamiltonion circuits and matchings. *J. Comp. Sys. Sci.*, 18(2):155–193, 1979.

Bickel, P. and Levina, E. Some theory of Fisher's linear discriminant function, 'Naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010, 2004.

Bishop, Christopher M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Breiman, Leo. Arcing classifiers. *The Annals of Statistics*, 1998.

Bühlmann, P. and Yu, B. Analyzing bagging. *Annals of Statistics*, 30:927–961, 2002.

DasGupta, A. *Asymptotic theory of statistics and probability*. Springer, 2008.

Dasgupta, S. and Long, P. M. Boosting with diverse base classifiers. *COLT*, 2003.

Feller, W. Generalization of a probability limit theorem of cramér. *Trans. Am. Math. Soc.*, 54:361–372, 1943.

Feller, W. *An introduction to probability theory and its applications*. John Wiley & Sons, 1968.

Friedman, J., Hastie, T., and Tibshirani, R. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 38(2):337–407, 2000.

Haussler, D., Littlestone, N., and Warmuth, M. K. Predicting $\{0,1\}$-functions on randomly drawn points. *In-*

formation and Computation, 115(2):129–161, 1994.

Koltchinskii, V. and Panchenko, D. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30(1), 2002.

Matoušek, J. and Vondrak, J. The probabilistic method, 2011. Lecture notes.

Motwani, R. and Raghavan, P. *Randomized Algorithms*. Cambridge University Press, 1995.

Ng, A. Y. and Jordan, M. I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *NIPS*, 2001.

Pemmaraju, S. Equitable coloring extends Chernoff-Hoeffding bounds. *RANDOM*, 2001.

Pollard, D. *Convergence of Stochastic Processes*. Springer Verlag, 1984.

Quinlan, J. Bagging, boosting and C4.5. *AAAI*, 1996.

Schapire, R. E., Freund, Y., Bartlett, P. L., and Lee, W. S. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26 (5):1651–1686, 1998.

Slud, E. Distribution inequalities for the binomial law. *Annals of Probability*, 5:404–412, 1977.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS*, 99(10):6567–72, 2002.

Vapnik, V. N. *Statistical Learning Theory*. New York, 1998.

Wang, L., Sugiyama, M., Yang, C., Zhou, Z., and Feng, J. On the margin explanation of boosting algorithms. *COLT*, 2008.

A. Appendices

A.1. Proof of (5) and (6)

Equation 4.1 from (Motwani & Raghavan, 1995) is

$$\Pr[U > (1+\eta)\mathbb{E}(U)] < \left(\frac{e^{\eta}}{(1+\eta)^{1+\eta}}\right)^{\mathbb{E}(U)}. \qquad (12)$$

which implies (5). From (12), when $0 \leq \eta \leq 4$ (since $\eta - (1+\eta)\ln(1+\eta) < -\eta^2/4$ there), $\Pr[U > (1+\eta)\mathbb{E}(U)] < \exp\left(-\eta^2\mathbb{E}(U)/4\right)$ showing (6).

A.2. Proof of (7)

Using (5) with $\eta = 3 + 3\ln(1/\delta)/\mathbb{E}(U)$,

$$\begin{aligned}&\Pr[U > 4\mathbb{E}(U) + 3\delta] \\ &< \exp\left(-(4\mathbb{E}(U) + 3\ln\delta)\ln\left(\frac{4 + 3\ln(1/\delta)/\mathbb{E}(U)}{e}\right)\right) \\ &< \exp\left(-(3\ln(1/\delta)\ln\left(\frac{4}{e}\right)\right) < \delta.\end{aligned}$$

A.3. Proof of (8)

The following is a straightforward consequence of the Berry-Esseen inequality.

Lemma 9 (see DasGupta (2008))**.** *Under the assumptions of Section 2 with each* $\Pr[U_i = 1] = 1/2$*, let:* $T_i = 2(U_i - 1/2)$ *and* $T = \sqrt{\frac{1}{\ell}}\sum_{i=1}^{\ell} T_i$*, and* Z *be a standard normal random variable.*
There is an absolute positive constant $c < 1$ *such that, for all* η*, we have* $|\Pr[T > \eta] - \Pr[Z > \eta]| \leq \frac{c}{\sqrt{\ell}}$.

Lemma 10. (Feller, 1968) *If* Z *is a standard normal random variable and* $\eta > 0$*, then* $\Pr[Z > \eta] \geq \frac{1}{\sqrt{2\pi}}\left(\frac{1}{\eta} - \frac{1}{\eta^3}\right)e^{-\eta^2/2}$.

Now, to prove (8), let $M = \frac{1}{\ell}\sum_{i=1}^{\ell}(U_i - \frac{1}{2})$ and let Z be a standard normal random variable. Then Lemma 9 implies that, for all κ

$$\left|\Pr\left[2\sqrt{\ell}M > \kappa\right] - \Pr[Z > \kappa]\right| \leq \frac{c}{\sqrt{\ell}}$$

for an absolute constant $c > 0$. Using $\kappa = 2\eta\sqrt{\ell}$,

$$\Pr[M > \eta] \geq \Pr\left[Z > 2\eta\sqrt{\ell}\right] - \frac{c}{\sqrt{\ell}}. \qquad (13)$$

Applying Lemma 10, we get

$$\Pr\left[Z > 2\eta\sqrt{\ell}\right] \quad \geq \quad \frac{1}{\sqrt{2\pi}}\left(\frac{1}{2\eta\sqrt{\ell}} - \left(\frac{1}{2\eta\sqrt{\ell}}\right)^3\right)e^{-2\eta^2\ell}.$$

Since $\ell \geq 1/\eta^2$, we get

$$\Pr\left[Z > 2\eta\sqrt{\ell}\right] \quad \geq \quad \frac{1}{\sqrt{2\pi}}\left(\frac{1}{2} - \frac{1}{8}\right)\frac{1}{\eta\sqrt{\ell}}e^{-2\eta^2\ell}.$$

Combining with (13) completes the proof of (8). □

A.4. Proof of (10)

When $\eta \leq 1/8$, Inequality (10) follows from a result of Feller (1943) (see Matoušek & Vondrak (2011)).

When $\eta > 1/8$, we have $\Pr\left[\frac{1}{\ell}U - \frac{1}{\ell}\mathbb{E}(U) \geq \eta\right] = \frac{1}{2^\ell}\sum_{i=0}^{(1/2-\eta)\ell}\binom{\ell}{i} \geq \frac{1}{2^\ell}\binom{\ell}{(1/2-\eta)\ell} \geq \frac{1}{2^\ell}\left(\frac{1}{1/2-\eta}\right)^{(1/2-\eta)\ell} = \exp(-\ell(\ln(2)+(1/2-\eta)\ln(1/2-\eta))) \geq \exp(-16\eta^2)$, where the last step can be verifying using Calculus (since $\eta > 1/8$).