Journal of Machine Learning Research v (2012) pp Submitted sub; Published pub

On the Necessity of Irrelevant Variables

David P. Helmbold DPH@SOE.UCSC.EDU
Department of Computer Science
University of California, Santa Cruz
Santa Cruz, CA 95064, USA

Philip M. Long PLONG@SV.NEC-LABS.COM
NEC Labs America
10080 N. Wolfe Rd, SW3-350
Cupertino, CA 95014, USA

Editor: Gábor Lugosi

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.

Keywords: Feature Selection, Generalization, Learning Theory

1. Introduction

When creating a classifier, a natural inclination is to only use variables that are obviously relevant since irrelevant variables typically decrease the accuracy of a classifier. On the other hand, this paper shows that the harm from irrelevant variables can be much less than the benefit from relevant variables and therefore it is possible to learn very accurate classifiers even when almost all of the variables are irrelevant. It can be advantageous to continue adding variables, even as their prospects for being relevant fade away. We show this with theoretical analysis and experiments using artificially generated data.

We provide an illustrative analysis that isolates the effects of relevant and irrelevant variables on a classifier's accuracy. We analyze the case in which variables complement one another, which we formalize using the common assumption of conditional independence given the class label. We focus on the situation where relatively few of the many variables are relevant, and the relevant variables are only weakly predictive.[1] Under these conditions, algorithms that cast a wide net can succeed while more selective algorithms fail.

1. Note that 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.

We prove upper bounds on the error rate of a very simple learning algorithm that may include many irrelevant variables in its hypothesis. We also prove a contrasting lower bound on the error of every learning algorithm that uses mostly relevant variables. The combination of these results show that the simple algorithm's error rate approaches zero in situations where every algorithm that predicts with mostly relevant variables has an error rate greater than a positive constant.

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). Consider for the moment a simple algorithm applied to a simple source. Each of two classes is equally likely, and there are 1000 relevant boolean 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 boolean 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. Figure 2 plots 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.[2] 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

2. In the first graph, only the results in which fewer than 1000 features were chosen are shown, since including larger feature sets obscures the shape of the graph in the most interesting region, where relatively few features are chosen.

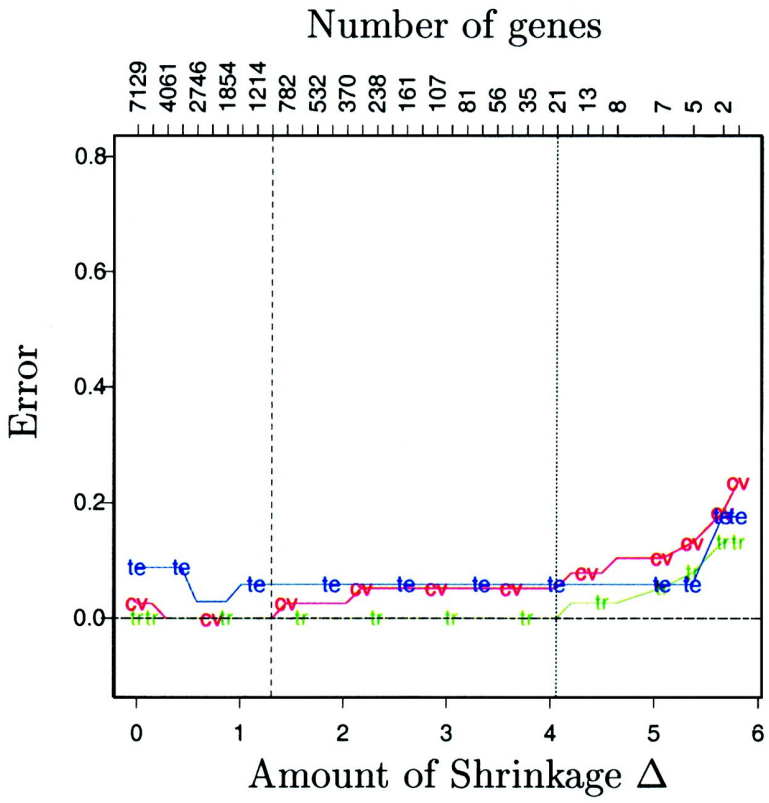

Figure 1: This graph is reproduced from (Tibshirani et al., 2002). For a microarray dataset, the training error, test error, and cross-validation error are plotted as a function both of the number of features (along the top) included in a linear model and a regularization parameter Δ (along the bottom).

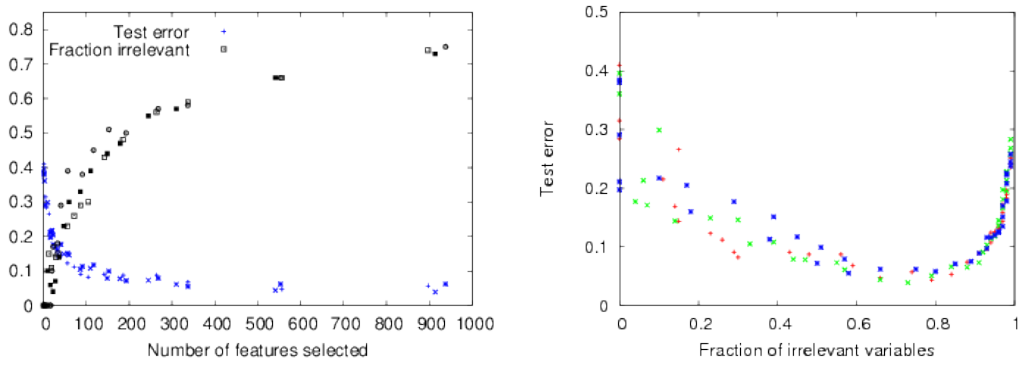

Figure 2: Left: Test error (blue) and fraction of irrelevant variables (black) as a function of the number of features. Right: Scatter plot of test error rates (vertical) against fraction of irrelevant variables (horizontal).

more variables than there are training examples. Also, accuracy over 90% is achieved even though there are few training examples and the correlation of the individual variables with the class label is weak. In fact, the number of examples is so small and the correlations are so weak that, for any individual feature, it is impossible to confidently tell whether or not the feature is relevant.

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 predominantly of irrelevant variables.

In Section 4 we consider learning from training data where the variables are conditionally independent given the class label. Whereas Equation (1) bounded the error as a function of the number of variables n and relevant variables k in the *model*, we now use capital N and capital K for the total number of variables and number of relevant variables in the *data*. The $N-K$ irrelevant variables are 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 $1/2-\gamma$. We analyze an algorithm that chooses a value $\beta \geq 0$ 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). Our Theorem 3 shows 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 $[z]_+ \stackrel{\text{def}}{=} \max\{z,0\}$. (Throughout the paper, the "big Oh" and other asymptotic notation will be for the case where γ is small, $K\gamma$ is large, and N/K is large. Thus the edge of the relevant features and the fraction of features that are relevant both approach zero while the total number of relevant features increases. If K is not large relative to $1/\gamma^2$, even the Bayes optimal classifier is not accurate. No other assumptions about the relationship between the parameters are needed.)

When $\beta \leq \gamma/2$ and the number m of training examples satisfies $m \geq c/\gamma^2$ for an absolute constant c, we also show in Theorem 8 that the error probability is at most

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

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.[3] Because bound (3) holds even when $\beta = 0$, it can be achieved by an algorithm that does not use knowledge of γ or K.

Our upper bounds illustrate the potential rewards for algorithms that are "inclusive", using many of the available variables in their classifiers – even when this means that most variables in the model are irrelevant. We also prove a complementary lower bound that illustrates the potential cost when algorithms are "exclusive". We say that an algorithm is λ-exclusive if the expectation of the fraction of the variables used in its model that are relevant is at least λ. We show that any λ-exclusive policy has an error probability bounded below by $\lambda/4$ 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 (where λ is a positive constant) can achieve a bound like (3).

Relationship to Previous Work Donoho and Jin (see Donoho and Jin, 2008; Jin, 2009) and Fan and Fan (2008), building on a line of research on joint testing of multiple hypotheses (see Abramovich et al., 2006; Addario-Berry et al., 2010; Donoho and Jin, 2004, 2006; Meinshausen and Rice, 2006), performed analyses and simulations using sources with elements in common with the model studied here, including conditionally independent variables and a weak association between the variables and the class labels. Donoho and Jin also pointed out that their algorithm can produce accurate hypotheses while using many more irrelevant features than relevant ones. The main theoretical results proved in their papers describe conditions that imply that, if the relevant variables are too small a fraction of all the variables, and the number of examples is too small, then learning is impossible. The

3. 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.

emphasis of our theoretical analysis is the opposite: algorithms can tolerate a large number of irrelevant variables, while using a small number of examples, and algorithms that avoid irrelevant variables, even to a limited extent, cannot learn as effectively as algorithms that cast a wider net. In particular, ours is the first analysis that we are aware of to have a result qualitatively like Theorem 13, which demonstrates the limitations of exclusive algorithms.

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 are aware of (such as Schapire et al. (1998); Koltchinskii and Panchenko (2002); Dasgupta and 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 and Jordan (2001) showed that the Naive Bayes algorithm (which ignores class-conditional dependencies) converges relatively quickly, justifying its use when there are few examples. But their bound for Naive Bayes is also vacuous when $m = \Theta(1/\gamma^2)$. Bickel and 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 and 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). As partial illustration of the robustness of the effects analyzed here, we generalize upper bound (1) to this case in Section 6.1. There we prove an error 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 proving 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 easily give similar theorems when each relevant variable has a probability between $1/2+\gamma/2$ and $1/2+2\gamma$ of agreeing with the class label (as discussed in Section 6.2). Most of this paper uses the cleanest and simplest setting 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 gives bounds on the expected error of hypotheses learned from training data while Section 5 shows that, in certain situations, any exclusive algorithm must have high error while the error of some inclusive algorithms goes to 0. In Section 6.1 we bound the accuracy of voting classifiers under a weakened independence assumption and in Section 6.2 we consider relaxation of the assumption that all relevant variables have the same edge.

2. Tail bounds

This section gathers together the several tail bounds that will be used in various places in the analysis. These bounds all assume that $U_1, U_2, \ldots, U_\ell$ are ℓ independent $\{0,1\}$-valued random variables and $U = \sum_{i=1}^{\ell} U_i$. We start with some upper bounds.

- The Hoeffding bound, (see Pollard, 1984):

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

- The Chernoff bound, (Angluin and Valiant, 1979; Motwani and Raghavan, 1995, see) and Appendix A.1. For any $\eta > 0$:

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

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

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

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

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

We also use the following lower bounds on the tails of distributions.

- If $\mathbb{P}[U_i = 1] = 1/2$ for all i, $\eta > 0$, and $\ell \geq 1/\eta^2$ then (see Appendix A.3):

$$\mathbb{P}\left[\frac{1}{\ell}U - \frac{1}{\ell}\mathbb{E}\left(U\right) \geq \eta\right] \geq \frac{1}{7\eta\sqrt{\ell}}\exp\left(-2\eta^2\ell\right) - \frac{1}{\sqrt{\ell}}. \tag{8}$$

- If $\mathbb{P}[U_i = 1] = 1/2$ for all i, then for all $0 \leq \eta \leq 1/8$ such that $\eta\ell$ is an integer[4] (see Appendix A.4):

$$\mathbb{P}\left[\frac{1}{\ell}U - \frac{1}{\ell}\mathbb{E}(U) \geq \eta\right] \geq \frac{1}{5}e^{-16\eta^2\ell}. \tag{9}$$

4. For notational simplicity we omit the floors/ceilings implicit in the use of this bound.

- A consequence of Slud's Inequality (1977) gives the following (see Appendix A.5). If $0 \leq \eta \leq 1/5$ and $\mathbb{P}[U_i = 1] = 1/2 + \eta$ for all i then:

$$\mathbb{P}\left[\frac{1}{\ell}U < 1/2\right] \geq \frac{1}{4}e^{-5\eta^2\ell}. \tag{10}$$

Note that the constants in the above bounds were chosen to be simple and illustrative, rather than the best possible.

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 use the following generative model:

- a random class designation from $\{0, 1\}$ is chosen, with both classes equally likely, then
- each of K *relevant* variables are equal to the class designation with probability $1/2+\gamma$ (or with probability $1/2 - \gamma$), and
- the remaining $N - K$ *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 positively 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 now 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. Recall that $[z]_+ \stackrel{\text{def}}{=} \max\{z, 0\}$.

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). The probability that $\mathcal{M}$ predicts incorrectly is at most* $\exp\left(\frac{-2\gamma^2[k-\ell]_+^2}{n}\right)$.

Proof: If $\ell \geq k$ then the exponent is 0 and the bound trivially holds.

Suppose $k > \ell$. 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 2 *If γ is a 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 (as $n \to \infty$).*

For example, the conditions of Corollary 2 are satisfied 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[5] 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 in this section apply to algorithms that chose β as a function of the number of features N, the number of relevant features K, the edge of the relevant features γ, and training set size m, and then predict with $\mathcal{M}_\beta$. Note that this includes the algorithm that always choses $\beta = 0$ regardless of N, K, γ and m.

Recall that asymptotic notation will concern the case in which γ is small, $K\gamma$ is large, and N/K is large.

This section proves two theorems bounding the expected error rates of learned models. One can compare these bounds with a similar bound on the Bayes Optimal predictor that "knows" which features are relevant. This 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).

Theorem 3 *If $0 \leq \beta \leq \gamma$, then 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 3 starts with lemmas bounding the number of misleading, irrelevant, and relevant features in $\mathcal{M}_\beta$. These lemmas use a quantity $\delta > 0$ that will be determined later in the analysis.

Lemma 4 *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 to be included in $\mathcal{M}_\beta$, Algorithm A must overestimate the probability that misleading feature equals the class label by at least $\beta + \gamma$.

5. If $\mathcal{M}_\beta$ is empty or the vote is tied then any default prediction, such as 1, will do.

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 $\mathbb{E}(U) \leq Ke^{-2(\beta+\gamma)^2 m}$ to get the desired result. □

Lemma 5 *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: For a particular positive irrelevant feature to be included in $\mathcal{M}_\beta$, Algorithm A must overestimate the probability that the positive irrelevant feature equals the class label by β. Applying (4), this happens with probability at most $e^{-2\beta^2 m}$, so the expected number of irrelevant positive features in $\mathcal{M}_\beta$ is at most $(N-K)e^{-2\beta^2 m}$.

All of the events that variables agree with the label, for various variables, and various examples, are independent. So the events that various irrelevant variables are included in $\mathcal{M}_\beta$ are independent. Applying (7) with $\mathbb{E}(U) = (N-K)e^{-2\beta^2 m}$ gives that, with probability at least $1-\delta$, the number of irrelevant positive features in $\mathcal{M}_\beta$ is at most $4(N-K)e^{-2\beta^2 m}$.

A symmetric analysis establishes the same bound on the number of negative irrelevant features in $\mathcal{M}_\beta$. Adding these up completes the proof. □

Lemma 6 *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: For a particular relevant feature to be excluded from $\mathcal{M}_\beta$, Algorithm A must underestimate the probability that the relevant feature equals the class label by at least $\gamma - \beta$. Applying (4), this happens with probability at most $e^{-2(\gamma-\beta)^2 m}$, so the expected number of relevant variables excluded from $\mathcal{M}_\beta$ is at most $Ke^{-2(\gamma-\beta)^2 m}$. Applying (7) as in the preceding two lemmas completes the proof. □

Lemma 7 *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 $0 \leq \beta \leq \gamma$.

Proof: The bounds of Lemmas 4, 5, and 6 simultaneously hold with probability at least $1-4\delta$. Thus the error probability of $\mathcal{M}_\beta$ is at most 4δ plus the probability of error given that all three bounds hold. Plugging the three bounds into Theorem 1, (and over-estimating the number n of variables in the model with K plus the bound of Lemma 5 on the number of irrelevant variables) gives a bound on $\mathcal{M}_\beta$'s error probability of

$$\exp\left(\frac{-2\gamma^2\left[(K - 4Ke^{-2(\gamma-\beta)^2 m} - 3\ln(1/\delta)) - (4Ke^{-2(\gamma+\beta)^2 m} + 3\ln(1/\delta))\right]_+^2}{K + 8Ne^{-2\beta^2 m} + 6\ln(1/\delta)}\right) \qquad (11)$$

when all three bounds hold. Under-approximating $(\gamma+\beta)^2$ with $(\gamma-\beta)^2$ and simplifying yields:

$$(11) \leq \exp\left(\frac{-2\gamma^2\left[K - 8Ke^{-2(\gamma-\beta)^2 m} - 6\ln(1/\delta)\right]_+^2}{K + 8Ne^{-2\beta^2 m} + 6\ln(1/\delta)}\right).$$

Adding 4δ completes the proof. □

We are now ready to prove Theorem 3.

Proof (of Theorem 3): Using

$$\delta = \exp\left(-\frac{\gamma K}{6}\right)$$

in Lemma 7 bounds the probability that $\mathcal{M}_\beta$ makes a mistake by

$$\exp\left(\frac{-2\gamma^2\left[K - 8Ke^{-2(\gamma-\beta)^2 m} - \gamma K\right]_+^2}{K + 8Ne^{-2\beta^2 m} + \gamma K}\right) + 4\exp\left(-\frac{\gamma K}{6}\right)$$
$$< \exp\left(\frac{-2\gamma^2 K\left[1 - 8e^{-2(\gamma-\beta)^2 m} - \gamma\right]_+^2}{1 + \frac{8N}{K}e^{-2\beta^2 m} + \gamma}\right) + 4\exp\left(-\frac{\gamma K}{6}\right). \quad (12)$$

The first term is at least $e^{-2\gamma^2 K}$, and

$$4\exp\left(-\frac{\gamma K}{6}\right) = o\left(e^{-2\gamma^2 K}\right)$$

as $\gamma \to 0$ and $\gamma K \to \infty$, so (12) implies the bound

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

as desired. □

The following theorem bounds the error in terms of just K, N, and γ when m is sufficiently large.

Theorem 8 *Suppose algorithm A produces models $\mathcal{M}_\beta$ where $0 \leq \beta \leq c\gamma$ for a constant $c \in [0,1)$.*

- *Then there is a constant b (depending only on c) such that whenever $m \geq b/\gamma^2$ the error of A's model is at most $(1+o(1))\exp\left(-\frac{\gamma^2 K^2}{N}\right)$.*

- *If $m = \omega(1/\gamma^2)$ then the error of A's model is at most $(1 + o(1)) \exp\left(\frac{-(2-o(1))\gamma^2 K^2}{N}\right)$.*

Proof Combining Lemmas 4 and 6 with the upper bound of N on the number of features in $\mathcal{M}_\beta$ as in Lemma 7's proof gives the following error bound on $\mathcal{M}_\beta$

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

for any $\delta > 0$. Setting

$$\delta = \exp\left(-\frac{\gamma K}{6}\right)$$

and continuing as in the proof of Theorem 3 gives the bound

$$(1 + o(1)) \exp\left(\frac{-2\gamma^2 K^2\left[1 - 2\left(4e^{-2(\gamma-\beta)^2 m} + \gamma\right)\right]_+^2}{N}\right). \tag{13}$$

For the first part of the theorem, it suffices to show that the $[\cdots]_+^2$ term is at least $1/2$. Recalling that our analysis is for small γ, the term inside the $[\cdots]_+$ of (13) is at least

$$1 - 8e^{-2(1-c)^2\gamma^2 m} - o(1).$$

When

$$m \geq \frac{\ln(32)}{2(1-c)^2\gamma^2}, \tag{14}$$

this term is at least $3/4 - o(1)$, and thus its square is at least $1/2$ for small enough γ, completing the proof of the first part of the theorem.

To see the second part of the theorem, since $m \in \omega(1/\gamma^2)$, the term of (13) inside the $[\cdots]_+$ is $1 - o(1)$. ■

By examining inequality (14), we see that the constant b in Theorem 8 can be set to $\ln(32)/2(1-c)^2$.

Lemma 9 *The expected number of irrelevant variables in $\mathcal{M}_\beta$ is at least $(N - K)e^{-16\beta^2 m}$.*

Proof Follows from inequality (9). ■

Corollary 10 *If K, N, and m are functions of γ such that*

$$\begin{aligned}
&\gamma \to 0,\\
&K^2/N \in \omega(\ln(1/\gamma)/\gamma^2),\\
&K = o(N) \text{ and}\\
&m = 2\ln(32)/\gamma^2
\end{aligned}$$

then if an algorithm outputs $\mathcal{M}_\beta$ using a β in $[0, \gamma/2]$, it has an error that decreases super-polynomially (in γ), while the expected fraction of irrelevant variables in the model goes to 1.

Note that Theorem 8 and Corollary 10 include non-trivial error bounds on the model $\mathcal{M}_0$ that votes all N variables (for odd sample size m).

5. Lower bound

Here we show that any algorithm with an error guarantee like Theorem 8 must include many irrelevant features in its model. The preliminary version of this paper (Helmbold and Long, 2011) contains a related lower bound for algorithms that choose β as a function of N, K, m, and γ, and predict with $\mathcal{M}_\beta$. Here we present a more general lower bound that applies to algorithms outputting arbitrary hypotheses. This includes algorithms that use weighted voting (perhaps with L_1 regularization). In this section we

- set the number of features N, number of relevant features K, and sample size m as a functions of γ in such a way that Corollary 10 applies, and
- prove a constant lower bound for these combinations of values that holds for "exclusive" algorithms (defined below) when γ is small enough.

Thus, in this situation, "inclusive" algorithms relying on many irrelevant variables have error rates going to zero while every "exclusive" algorithm has an error rate bounded below by a constant.

The proofs in this section assume that all relevant variables are positively correlated with the class designation, so each relevant variable agrees with the class designation with probability $1/2 + \gamma$. Although not essential for the results, this assumption simplifies the definitions and notation[6]. We also set $m = 2\ln(32)/\gamma^2$. This satisfies the assumption of Theorem 8 when $\beta \leq \gamma/2$ (see Inequality (14)).

Definition 11 *We say a classifier f includes a variable x_i if there is an input $(x_1, ..., x_N)$ such that*

$$f(x_1, ..., x_{i-1}, x_i, x_{i+1}, ..., x_N) \neq f(x_1, ..., x_{i-1}, 1 - x_i, x_{i+1}, ..., x_N).$$

Let $V(f)$ be the set of variables included in f.

For a training set S, we will refer to the classifier output by algorithm A on S as $A(S)$. Let $\mathcal{R}$ be the set of relevant variables.

Definition 12 *We say that an algorithm A is λ-exclusive[7] if for every positive N, K, γ, and m, the expected fraction of the variables included in its hypothesis that are relevant is at least λ, i.e.* $\mathbb{E}\left(\frac{|V(A(S)) \cap \mathcal{R}|}{|V(A(S))|}\right) \geq \lambda$.

6. The assumption that each relevant variable agrees with the class label with probability $1/2 + \gamma$ gives a special case of the generative model described in Section 4, so the lower bounds proven here also apply to that more general setting.
7. The proceedings version of this paper (Helmbold and Long, 2011) used a different definition of λ-exclusive.

Our main lower bound theorem is the following.

Theorem 13 *If*

$$K = \frac{1}{\gamma^2} \exp\left(\ln(1/\gamma)^{1/3}\right)$$
$$N = K \exp\left(\ln(1/\gamma)^{1/4}\right)$$
$$m = \frac{2\ln(32)}{\gamma^2}$$

then for any constant $\lambda > 0$ and any λ-exclusive algorithm A, the error rate of A is lower bounded by $\lambda/4 - o(1)$ as γ goes to 0.

Notice that this theorem provides a sharp contrast to Corollary 10. Corollary 10 shows that inclusive A using models $\mathcal{M}_\beta$ for any $0 \le \beta \le \gamma/2$ have error rates that goes to zero super-polynomially fast (in $1/\gamma$) under the assumptions of Theorem 13.

The values of K and N in Theorem 13 are chosen to make the proof convenient, but other values would work. For example, decreasing K and/or increasing N would make the lower bound part of Theorem 13 easier to prove. There is some slack to do so while continuing to ensure that the upper bound of Corollary 10 goes to 0.

As the correlation of variables with the label over the sample plays a central role in our analysis, we will use the following definition.

Definition 14 *If a variable agrees with the class label on $1/2 + \eta$ of the training set then it has* (empirical) edge η.

The proof of Theorem 13 uses a critical value of β, namely $\beta^* = \gamma \ln(N/K)/10\ln(32)$, with the property that both:

$$\frac{\mathbb{E}\left(|\mathcal{M}_{\beta^*} \cap \mathcal{R}|\right)}{\mathbb{E}\left(|\mathcal{M}_{\beta^*}|\right)} \to 0 \tag{15}$$
$$\mathbb{E}\left(|\mathcal{M}_{\beta^*} \cap \mathcal{R}|\right) \in o(1/\gamma^2) \tag{16}$$

as $\gamma \to 0$.

Intuitively, (15) means that any algorithm that uses most of the variables having empirical edge at least β^* cannot be λ-exclusive. On the other hand, (16) implies that if the algorithm restricts itself to variables with empirical edges greater than β^* then it does not include enough relevant variables to be accurate. The proof must show that *arbitrary* algorithms frequently include either too many irrelevant variables to be λ-exclusive or too few relevant ones to be accurate. See Figure 3 for some useful facts about γ, m, and β^*.

To prove the lower bound, borrowing a technique from Ehrenfeucht et al. (1989), we will assume that the K relevant variables are randomly selected from the N variables, and lower bound the error with respect to this random choice, along with the training and test data. This will then imply that, for each algorithm, there will be a choice of the K relevant variables giving the same lower bound with respect only to the random choice of the training and test data. We will always use relevant variables that are positively associated with the class label, agreeing with it with probability $1/2 + \gamma$.

$$b = 2\ln(32)$$

$$m = \frac{b}{\gamma^2} = \frac{2\ln(32)}{\gamma^2}$$

$$\beta^* = \frac{\gamma \ln(N/K)}{5b} = \frac{\gamma \ln(1/\gamma)^{1/4}}{10\ln(32)}$$

Figure 3: Some useful facts relating b, γ, m and β^* under the assumptions of Theorem 13.

Proof [of Theorem 13] Fix any learning algorithm A, and let $A(S)$ be the hypothesis produced by A from sample S. Let $n(S)$ be the number of variables included in $A(S)$ and let $\beta(S)$ be the $n(S)$'th largest empirical (w.r.t. S) edge of a variable.

Let q_γ be the probability that $\beta(S) \geq \beta^* = \gamma \ln(N/K)/10\ln(32)$. We will show in Section 5.2 that if A is λ-exclusive then $\lambda \leq q_\gamma + o(1)$ (as γ goes to 0). We will also show in Section 5.3 that the expected error of A is at least $q_\gamma/4 - o(1)$ as γ goes to 0. Therefore any λ-exclusive algorithm A has an expected error rate at least $\lambda/4 - o(1)$ as γ goes to 0. ■

Before attacking the two parts of the proof alluded to above, we need a subsection providing some basic results about relevant variables and optimal algorithms.

5.1 Relevant Variables and Good Hypotheses

This section proves some useful facts about relevant variables and good hypotheses. The first lemma is a lower bound on the accuracy of a model in terms of the number of relevant variables.

Lemma 15 *If $\gamma \in [0, 1/5]$ then any classifier using k relevant variables has an error probability at least $\frac{1}{4}e^{-5\gamma^2 k}$.*

Proof: The usual Naive Bayes calculation (see Duda et al., 2000) implies that the optimal classifier over a certain set V of variables is a majority vote over $V \cap \mathcal{R}$. Applying the lower tail bound (10) then completes the proof. □

Our next lemma shows that, given a sample, the probability that a variable is relevant (positively correlated with the class label) is monotonically increasing in its empirical edge.

Lemma 16 *For two variables x_i and x_j, and any training set S of m examples,*

- $\mathbb{P}[x_i \text{ relevant} \mid S] > \mathbb{P}[x_j \text{ relevant} \mid S]$ *if and only if the empirical edge of x_i in S is greater than the empirical edge of x_j in S, and*
- $\mathbb{P}[x_i \text{ relevant} \mid S] = \mathbb{P}[x_j \text{ relevant} \mid S]$ *if and only if the empirical edge of x_i in S is equal to the empirical edge of x_j in S.*

Proof Since the random choice of $\mathcal{R}$ does not effect that marginal distribution over the labels, we can generate S by picking the labels for all the examples first, then $\mathcal{R}$, and finally the values of the variables on all the examples. Thus if we can prove the lemma after conditioning on the values of the class labels, then scaling all of the probabilities by 2^{-m} would complete the proof. So, let us fix the values of the class labels, and evaluate probabilities only with respect to the random choice of the relevant variables $\mathcal{R}$, and the values of the variables.

Let

$$\Delta = \mathbb{P}[x_i \in \mathcal{R} | S] - \mathbb{P}[x_j \in \mathcal{R} | S] \,.$$

First, by subtracting off the probabilities that both variables are relevant, we have

$$\Delta = \mathbb{P}\big[x_i \in \mathcal{R}, x_j \notin \mathcal{R} \big| S\big] - \mathbb{P}\big[x_i \notin \mathcal{R}, x_j \in \mathcal{R} \big| S\big] \,.$$

Let ONE be the event that exactly one of x_i or x_j is relevant. Then

$$\Delta = (\mathbb{P}\big[x_i \in \mathcal{R}, x_j \notin \mathcal{R} \big| S, \mathrm{ONE}\big] - \mathbb{P}\big[x_i \notin \mathcal{R}, x_j \in \mathcal{R} \big| S, \mathrm{ONE}\big])\,\mathbb{P}[\mathrm{ONE}] \,.$$

So $\Delta > 0$ if and only if

$$\Delta' \stackrel{\mathrm{def}}{=} \mathbb{P}\big[x_i \in \mathcal{R}, x_j \notin \mathcal{R} \big| S, \mathrm{ONE}\big] - \mathbb{P}\big[x_i \notin \mathcal{R}, x_j \in \mathcal{R} \big| S, \mathrm{ONE}\big] > 0$$

(and similarly for $\Delta = 0$ if and only if $\Delta' = 0$). If $\mathbb{Q}$ is the distribution obtained by conditioning on ONE, then

$$\Delta' = \mathbb{Q}\big[x_i \in \mathcal{R}, x_j \notin \mathcal{R} \big| S\big] - \mathbb{Q}\big[x_i \notin \mathcal{R}, x_j \in \mathcal{R} \big| S\big] \,.$$

Let S_i be the values of variable i in S, and define S_j similarly for variable j. Let S' be the values of the other variables. Since we have already conditioned on the labels, after also conditioning on ONE (i.e., under the distribution $\mathbb{Q}$), the pair (S_i, S_j) is independent of S'. For each S_i we have $\mathbb{P}[S_i \mid x_i \notin \mathcal{R}] = \mathbb{Q}[S_i \mid x_i \notin \mathcal{R}]$. Furthermore, by symmetry,

$$\mathbb{Q}\big[x_i \in \mathcal{R}, x_j \notin \mathcal{R} \big| S'\big] = \mathbb{Q}\big[x_i \notin \mathcal{R}, x_j \in \mathcal{R} \big| S'\big] = \frac{1}{2}.$$

Thus, by using Bayes' Rule on each term, we have

$$\begin{aligned}
\Delta' &= \mathbb{Q}\big[x_i \in \mathcal{R}, x_j \notin \mathcal{R} \big| S_i, S_j, S'\big] - \mathbb{Q}\big[x_i \notin \mathcal{R}, x_j \in \mathcal{R} \big| S_i, S_j, S'\big] \\
&= \frac{\mathbb{Q}\big[S_i, S_j \big| x_i \in \mathcal{R}, x_j \notin \mathcal{R}, S'\big] - \mathbb{Q}\big[S_i, S_j \big| x_i \notin \mathcal{R}, x_j \in \mathcal{R}, S'\big]}{2\,\mathbb{Q}[S_i, S_j | S']} \\
&= \frac{(1/2+\gamma)^{m_i}(1/2-\gamma)^{m-m_i} - (1/2+\gamma)^{m_j}(1/2-\gamma)^{m-m_j}}{2^{m+1}\,\mathbb{Q}[S_i, S_j]},
\end{aligned}$$

where m_i and m_j are the numbers of times that variables x_i and x_j agree with the label in sample S. The proof concludes by observing that Δ' is positive exactly when $m_i > m_j$ and zero exactly when $m_i = m_j$. ■

Because, in this lower bound proof, relevant variables are always positively associated with the class label, we will use a variant of $\mathcal{M}_\beta$ which only considers positive features.

Definition 17 *Let $\mathcal{V}_\beta$ be a vote over the variables with empirical edge at least β.*

When there is no chance of confusion, we will refer to the set of variables in $\mathcal{V}_\beta$ also as $\mathcal{V}_\beta$ (rather than $V(\mathcal{V}_\beta)$).

We now establish lower bounds on the probability of variables being included in $\mathcal{V}_\beta$ (here β can be a function of γ, but does not depend on the particular sample S).

Lemma 18 *If $\gamma \leq 1/8$ and $\beta \geq 0$ then the probability that a given variable has empirical edge at least β is at least*

$$\frac{1}{5}\exp\left(-16\beta^2 m\right).$$

If in addition $m \geq 1/\beta^2$, then the probability that a given variable has empirical edge at least β is at least

$$\frac{1}{7\beta\sqrt{m}}\exp\left(-2\beta^2 m\right) - \frac{1}{\sqrt{m}}.$$

Proof: Since relevant variables agree with the class label with probability $1/2+\gamma$, the probability that a relevant variable has empirical edge at least β is lower bounded by the probability that an irrelevant variable has empirical edge at least β. An irrelevant variable has empirical edge at least β only when it agrees with the class on $1/2+\beta$ of the sample. Applying Bound (9), this happens with probability at least $\frac{1}{5}\exp\left(-16\beta^2 m\right)$. The second part uses Bound (8) instead of (9). □

We now upper bound the probability of a relevant variable being included in $\mathcal{V}_\beta$, again for β that does not depend on S.

Lemma 19 *If $\beta \geq \gamma$, the probability that a given relevant variable has empirical edge at least β is at most $e^{-2(\beta-\gamma)^2 m}$.*

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. □

5.2 Bounding λ-Exclusiveness

Recall that $n(S)$ is the number of variables used by $A(S)$, and $\beta(S)$ is the edge of the variable whose rank, when the variables are ordered by their empirical edges, is $n(S)$. We will show that: if $A(S)$ is λ-exclusive, then there is reasonable probability that $\beta(S)$ is at least the critical value $\beta^* = \gamma\ln(N/K)/5b$. Specifically, if A is λ-exclusive, then, for any small enough γ, we have $\mathbb{P}[\beta(S) \geq \beta^*] > \lambda/2$.

Suppose, given the training set S, the variables are sorted in decreasing order of empirical edge (breaking ties arbitrarily, say using the variable index). Let $\mathcal{V}_{S,k}$ consist of the first k variables in this sorted order, the "top k" variables.

Since for each sample S and each variable x_i, the probability $\mathbb{P}\left[x_i \text{ relevant} \middle| S\right]$ decreases as the empirical edge of x_i decreases (Lemma 16), the expectation $\mathbb{E}\left(\frac{|\mathcal{V}_{S,k}\cap\mathcal{R}|}{|\mathcal{V}_{S,k}|}\,\middle|\, S\right)$ is non-increasing with k.

Furthermore, Lemma 16 also implies that for each sample S, we have

$$\mathbb{E}\left(\frac{|V(A(S))\cap\mathcal{R}|}{|V(A(S))|}\;\middle|\;S\right)\le\mathbb{E}\left(\frac{|\mathcal{V}_{S,n(S)}\cap\mathcal{R}|}{|\mathcal{V}_{S,n(S)}|}\;\middle|\;S\right).$$

Therefore, by averaging over samples, for each γ we have

$$\mathbb{E}\left(\frac{|V(A(S))\cap\mathcal{R}|}{|V(A(S))|}\right)\le\mathbb{E}\left(\frac{|\mathcal{V}_{S,n(S)}\cap\mathcal{R}|}{|\mathcal{V}_{S,n(S)}|}\right). \tag{17}$$

Note that the numerators in the expectations are never greater than the denominators. We will next give upper bounds on $|\mathcal{V}_{\beta^*}\cap\mathcal{R}|$ and lower bounds on $|\mathcal{V}_{\beta^*}|$ that each hold with probability $1-\gamma$.

The next step is a high-confidence upper bound on $|\mathcal{V}_{\beta^*}\cap\mathcal{R}|$. From Lemma 19, the probability that a particular relevant variable is in $\mathcal{V}_{\beta^*}$ is at most (recall that $m=b/\gamma^2$)

$$\begin{aligned}
e^{-2(\beta^*-\gamma)^2m} &= e^{-2b(\beta^*/\gamma-1)^2}\\
&= \exp\left(-2b\left(\frac{\ln(1/\gamma)^{1/4}}{5b}-1\right)^2\right)\\
&= \exp\left(\frac{-2\ln(1/\gamma)^{1/2}}{25b}+\frac{4\ln(1/\gamma)^{1/4}}{5}-2b\right)\\
&< \exp\left(\frac{-2\ln(1/\gamma)^{1/2}}{25b}+\frac{4\ln(1/\gamma)^{1/4}}{5}\right).
\end{aligned}$$

Let $p_{\rm rel}=\exp\left(\frac{-2\ln(1/\gamma)^{1/2}}{25b}+\frac{4\ln(1/\gamma)^{1/4}}{5}\right)$ be this upper bound, and note that $p_{\rm rel}$ drops to 0 as γ goes to 0, but a rate slower than γ^ϵ for any ϵ. The number of relevant variables in $\mathcal{V}_{\beta^*}$ has a binomial distribution with parameters K and p where $p<p_{\rm rel}$. The standard deviation of this distribution is

$$\sigma=\sqrt{Kp(1-p)}<\sqrt{Kp}<\frac{\exp\left(\frac{\ln(1/\gamma)^{1/3}}{2}\right)\sqrt{p_{\rm rel}}}{\gamma}. \tag{18}$$

Using the Chebyshev bound,

$$\mathbb{P}[|X-\mathbb{E}(X)|>a\sigma]\le\frac{1}{a^2} \tag{19}$$

with $a=1/\sqrt{\gamma}$ gives that

$$\begin{aligned}
\mathbb{P}\left[|\mathcal{V}_{\beta^*}\cap\mathcal{R}|-Kp>\frac{\sigma}{\sqrt{\gamma}}\right] &\le \gamma\\
\mathbb{P}\left[|\mathcal{V}_{\beta^*}\cap\mathcal{R}|>Kp_{\rm rel}+\frac{\sigma}{\sqrt{\gamma}}\right] &\le \gamma.
\end{aligned} \tag{20}$$

Since $\sigma < \sqrt{Kp} < \sqrt{Kp_{\text{rel}}}$ by (18), we have $\sigma\sqrt{Kp_{\text{rel}}} < Kp_{\text{rel}}$. Substituting the values of K and p_{rel} into the square-root yields

$$\begin{aligned} Kp_{\text{rel}} &> \sigma \times \frac{\exp\left(\frac{\ln(1/\gamma)^{1/3}}{2}\right)\exp\left(\frac{-\ln(1/\gamma)^{1/2}}{25b} + \frac{2\ln(1/\gamma)^{1/4}}{5}\right)}{\gamma} \\ &> \sigma/\sqrt{\gamma}, \end{aligned}$$

for small enough γ. Combining with (20), we get that

$$\mathbb{P}[|\mathcal{V}_{\beta^*} \cap \mathcal{R}| > 2Kp_{\text{rel}}] \leq \gamma \tag{21}$$

holds for small enough γ.

Using similar reasoning, we now obtain a lower bound on the expected number of variables in $\mathcal{V}_{\beta^*}$. Lemma 18 shows that, for each variable, the probability of the variable having empirical edge β^* is at least

$$\begin{aligned} \frac{1}{7\beta^*\sqrt{m}}\exp\left(-2\beta^{*2}m\right) - \frac{1}{\sqrt{m}} &= \frac{5\sqrt{b}}{7\ln(1/\gamma)^{1/4}}\exp\left(-2\frac{\ln(1/\gamma)^{1/2}}{25b}\right) - \frac{\gamma}{\sqrt{b}} \\ &> \frac{\sqrt{b}}{2\ln(1/\gamma)^{1/4}}\exp\left(-2\frac{\ln(1/\gamma)^{1/2}}{25b}\right) \end{aligned}$$

for sufficiently small γ. Since the empirical edges of different variables are independent, the probability that at least n variables have empirical edge at least β^* is lower bounded by the probability of at least n successes from the binomial distribution with parameters N and p_{irrel} where

$$p_{\text{irrel}} = \frac{\sqrt{b}}{2\ln(1/\gamma)^{1/4}}\exp\left(-2\frac{\ln(1/\gamma)^{1/2}}{25b}\right).$$

If, now, we define σ to be the standard deviation of this binomial distribution, then, like before, $\sigma = \sqrt{Np_{\text{irrel}}(1-p_{\text{irrel}})} < \sqrt{Np_{\text{irrel}}}$, and

$$\begin{aligned} Np_{\text{irrel}}/2 &> \sigma\sqrt{Np_{\text{irrel}}}/2 \\ &= \frac{\sigma}{2} \times \frac{\exp((1/2)(\ln(1/\gamma)^{1/4} + \ln(1/\gamma)^{1/3}))}{\gamma} \times \frac{b^{1/4}}{\sqrt{2}\ln(1/\gamma)^{1/8}}\exp\left(-\frac{\ln(1/\gamma)^{1/2}}{25b}\right), \end{aligned}$$

so that, for small enough γ, $Np_{\text{irrel}}/2 > \sigma/\sqrt{\gamma}$. Therefore applying the Chebyshev bound (19) with $a = 1/\sqrt{\gamma}$ gives (for sufficiently small γ)

$$\mathbb{P}\left[|\mathcal{V}_{\beta^*}| < \frac{Np_{\text{irrel}}}{2}\right] \leq \mathbb{P}[|\mathcal{V}_{\beta^*}| < Np_{\text{irrel}} - \sigma/\sqrt{\gamma}] < \gamma. \tag{22}$$

Recall that

$$q_\gamma = \mathbb{P}[\beta(S) \geq \beta^*] = \mathbb{P}[n(S) \leq |\mathcal{V}_{\beta^*}|].$$

If A is λ-exclusive then, using (17), we have

$$\begin{aligned}
\lambda \le \mathbb{E}\left(\frac{|V(A(S)) \cap \mathcal{R}|)}{|V(A(S))|}\right) &\le \mathbb{E}\left(\frac{|\mathcal{V}_{S,n(S)} \cap \mathcal{R}|}{|\mathcal{V}_{S,n(S)}|}\right) \\
&\le (1-q_\gamma)\mathbb{E}\left(\frac{|\mathcal{V}_{S,n(S)} \cap \mathcal{R}|}{|\mathcal{V}_{S,n(S)}|} \;\middle|\; |\mathcal{V}_{\beta^*}| < n(S)\right) + q_\gamma \\
&\le (1-q_\gamma)\mathbb{E}\left(\frac{|\mathcal{V}_{\beta^*} \cap \mathcal{R}|}{|\mathcal{V}_{\beta^*}|} \;\middle|\; |\mathcal{V}_{\beta^*}| < n(S)\right) + q_\gamma \\
&\le (1-q_\gamma)\left(\frac{2Kp_{\text{rel}}}{Np_{\text{irrel}}/2} + 2\gamma\right) + q_\gamma
\end{aligned}$$

where we use the upper and lower bounds from Equations (21) and (22) that each hold with probability $1-\gamma$. Note that the ratio

$$\begin{aligned}
\frac{2Kp_{\text{rel}}}{Np_{\text{irrel}}/2} &\le \frac{2\dfrac{e^{\ln(1/\gamma)^{1/3}}}{\gamma^2}\exp\left(\dfrac{-2\ln(1/\gamma)^{1/2}}{25b} + \dfrac{4\ln(1/\gamma)^{1/4}}{5}\right)}{\dfrac{e^{\ln(1/\gamma)^{1/3}}e^{\ln(1/\gamma)^{1/4}}}{4\gamma^2}\dfrac{\sqrt{b}}{\ln(1/\gamma)^{1/4}}\exp\left(-2\dfrac{\ln(1/\gamma)^{1/2}}{25b}\right)} \\
&= \frac{8\ln(1/\gamma)^{1/4}\exp\left(\dfrac{-2\ln(1/\gamma)^{1/2}}{25b} + \dfrac{4\ln(1/\gamma)^{1/4}}{5}\right)}{\sqrt{b}\,e^{\ln(1/\gamma)^{1/4}}\exp\left(-2\dfrac{\ln(1/\gamma)^{1/2}}{25b}\right)} \\
&= \frac{8\ln(1/\gamma)^{1/4}\exp\left(\dfrac{-\ln(1/\gamma)^{1/4}}{5}\right)}{\sqrt{b}}
\end{aligned}$$

which goes to 0 as γ goes to 0. Therefore,

$$\lambda \le \mathbb{E}\left(\frac{|V(A(S)) \cap \mathcal{R}|)}{|V(A(S))|}\right) \le q_\gamma + o(1)$$

which implies that,

$$q_\gamma = \mathbb{P}[\beta(S) \ge \beta^*] \ge \lambda - o(1)$$

as γ goes to 0.

5.3 Large Error

Call a variable *good* if it is relevant and its empirical edge is at least β^* in the sample. Let p be the probability that a relevant variable is good. Thus the number of good variables is binomially distributed with parameters K and p. We have that the expected number of good variables is pK and the variance is $Kp(1-p) < Kp$. By Chebyshev's inequality, we have

$$\mathbb{P}\Big[\#\text{ good vars} \ge Kp + a\sqrt{Kp}\Big] \le \mathbb{P}\Big[\#\text{ good vars} \ge Kp + a\sqrt{Kp(1-p)}\Big] \le \frac{1}{a^2}, \tag{23}$$

and setting $a = \sqrt{Kp}$, this gives

$$\mathbb{P}[\#\text{ good vars} \geq 2Kp] \leq \frac{1}{Kp}. \tag{24}$$

By Lemma 19, $Kp \leq Ke^{-2(\beta^*-\gamma)^2 m} = Ke^{-2b\left(\ln(1/\gamma)^{1/4}/5b-1\right)^2}$, so

$$\ln(Kp) \leq \ln K - 2b\left(\frac{\ln(1/\gamma)^{1/2}}{25b^2} - \frac{2\ln(1/\gamma)^{1/4}}{5b} + 1\right)$$

$$\ln(Kp) \leq 2\ln(1/\gamma) + \ln(1/\gamma)^{1/3} - 2b\left(\frac{\ln(1/\gamma)^{1/2}}{25b^2} - \frac{2\ln(1/\gamma)^{1/4}}{5b} + 1\right).$$

So for small enough γ,

$$\ln(Kp) \leq 2\ln(1/\gamma) - \frac{\ln(1/\gamma)^{1/2}}{25b}$$

and thus $Kp \in o(1/\gamma^2)$.

So if $Kp > 1/\gamma$, then with probability at least $1-\gamma$, there are less than $2Kp \in o(1/\gamma^2)$ good variables. On the other hand, if $Kp < 1/\gamma$, then, setting $a = \sqrt{1/\gamma}$ in bound (23) gives that the probability that there are more than $2/\gamma$ good variables is at most γ. So in either case the probability that there are more than $\frac{2}{\gamma^2}\exp\left(-\ln(1/\gamma)^{1/2}/25b\right)$ good variables is at most γ (for small enough γ).

So if $\mathbb{P}[\beta(S) \geq \beta^*] \geq q_\gamma$, then with probability at least $q_\gamma - \gamma$ algorithm A is using a hypothesis with at most $\frac{2}{\gamma^2}\exp\left(-\ln(1/\gamma)^{1/2}/25b\right)$ relevant variables. Applying Lemma 15 yields the following lower bound on the probability of error:

$$(q_\gamma - \gamma)\frac{1}{4}\exp\left(-10\exp\left(-\ln(1/\gamma)^{1/2}/25b\right)\right). \tag{25}$$

Since the limit of (25) for small γ is $q_\gamma/4$, this completes the proof of Theorem 13.

6. Relaxations of some assumptions

To keep the analysis clean, and facilitate the interpretation of the results, we have analyzed an idealized model. In this section, we briefly consider the consequences of some relaxations of our assumptions.

6.1 Conditionally dependent variables

Theorem 1 can be generalized to the case in which there is limited dependence among the variables, after conditioning on the class designation, in a variety of ways. For example, suppose 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*. Then applying a Chernoff-Hoeffding bound for such sets of random variables due to Pemmaraju (2001), if $r \leq n/2$, one gets a bound of $c(r+1)\exp\left(\frac{-2\gamma^2k^2}{n(r+1)}\right)$ the probability of error.

6.2 Variables with different strengths

We have previously assumed that all relevant variables are equally strongly associated with the class label. Our analysis is easily generalized to the situation when the strengths of associations fall in an interval $[\gamma_{\min}, \gamma_{\max}]$. Thus relevant variables agree with the class label with probability at least $1/2 + \gamma_{\min}$ and misleading variables agree with the class label with probability at least $1/2 - \gamma_{\max}$. Although a sophisticated analysis would take each variable's degree of association into account, it is possible to leverage our previous analysis with a simpler approach. Using the $1/2 + \gamma_{\min}$ and $1/2 - \gamma_{\max}$ underestimates on the probability that relevant variables and misleading variables agree with the class label leads to an analog of Theorem 1. This analog says that models voting n variables, k of which are relevant and ℓ of which are misleading, have error probabilities bounded by

$$\exp\left(\frac{-2[\gamma_{\min}k - \gamma_{\max}\ell]_+^2}{n}\right).$$

We can also use the upper and lower bounds on association to get high-confidence bounds (like those of Lemmas 4 and 6) on the numbers of relevant and misleading features in models $\mathcal{M}_\beta$. This leads to an analog of Theorem 3 bounding the expected error rate of $\mathcal{M}_\beta$ by

$$(1+o(1))\exp\left(\frac{-2\gamma_{\min}^2 K\left[1 - 4(1+\gamma_{\max}/\gamma_{\min})e^{-2(\gamma_{\min}-\beta)^2 m} - \gamma_{\min}\right]_+^2}{1 + \frac{8N}{K}e^{-2\beta^2 m} + \gamma_{\min}}\right)$$

when $0 \le \beta \le \gamma$ and $\gamma_{\max} \in o(1)$. Note that γ in Theorem 3 is replaced by $\gamma_{\min}$ here, and $\gamma_{\max}$ only appears in the $4(1+\gamma_{\max}/\gamma_{\min})$ factor (which replaces an "8" in the original theorem).

Continuing to mimic our previous analysis gives analogs to Theorem 8 and Corollary 10. These analogs imply that if $\gamma_{\max}/\gamma_{\min}$ is bounded then algorithms using small β perform well in the same limiting situations used in Section 5 to bound the effectiveness of exclusive algorithms.

A more sophisticated analysis keeping better track of the degree of association between relevant variables and the class label may produce better bounds. In addition, if the variables have varying strengths then it makes sense to consider classifiers that assign different voting weights to the variables based on their estimated strength of association with the class label. An analysis that takes account of these issues is a potentially interesting subject for further research.

7. Conclusions

We analyzed learning when there are few examples, a small fraction of the variables are relevant, and the relevant variables are only weakly correlated with the class label. In this situation, algorithms that produce hypotheses consisting predominately of irrelevant variables can be highly accurate (with error rates going to 0). Furthermore, this inclusion of many irrelevant variables is essential. Any algorithm limiting the expected fraction of irrelevant variables in its hypotheses has an error rate bounded below by a constant. This

is in stark contrast with many feature selection heuristics that limit the number of features to a small multiple of the number of examples, or that limit the classifier to use variables that pass stringent statistical tests of association with the class label.

These results have two implications on the practice of machine learning. First, they show that the engineering practice of producing models that include enormous numbers of variables is sometimes justified. Second, they run counter to the intuitively appealing view that accurate class prediction "validates" the variables used by the predictor.

Acknowledgements

We thank Aravind Srinivasan for his help.

Appendix A. Appendices

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

Equation 4.1 from (Motwani and Raghavan, 1995) is

$$\mathbb{P}[U > (1+\eta)\mathbb{E}(U)] < \left(\frac{e^{\eta}}{(1+\eta)^{1+\eta}}\right)^{\mathbb{E}(U)} \tag{26}$$

and holds for independent 0-1 valued U_i's each with (possibly different) probabilities $p_i = P(U_i = 1)$ where $0 < p_i < 1$ and $\eta > 0$. Taking the logarithm of the RHS, we get

$$\begin{aligned} \ln(\text{RHS}) &= \mathbb{E}(U)\left(\eta - (1+\eta)\ln(1+\eta)\right) \\ &< \mathbb{E}(U)\left(\eta + 1 - (1+\eta)\ln(1+\eta)\right) \\ &= -\mathbb{E}(U)(\eta+1)(\ln(1+\eta) - 1), \end{aligned} \tag{27}$$

which implies (5). From (26), when $0 \leq \eta \leq 4$ (since $\eta - (1+\eta)\ln(1+\eta) < -\eta^2/4$ there), $\mathbb{P}[U > (1+\eta)\mathbb{E}(U)] < \exp\left(-\eta^2\mathbb{E}(U)/4\right)$ showing (6).

A.2 Proof of (7)

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

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

using the fact that $\ln(4/e) \approx 0.38 > 1/3$.

A.3 Proof of (8)

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

Lemma 20 ((see DasGupta, 2008, Theorem 11.1)) *Under the assumptions of Section 2 with each $\mathbb{P}[U_i = 1] = 1/2$, let:*

$$T_i = 2(U_i - 1/2),$$
$$T = \sqrt{\frac{1}{\ell}} \sum_{i=1}^{\ell} T_i, \text{ and}$$
$$Z \text{ be a standard normal random variable.}$$

Then for all η, we have $\left|\mathbb{P}[T > \eta] - \mathbb{P}[Z > \eta]\right| \leq \frac{1}{\sqrt{\ell}}$.

Lemma 21 ((Feller, 1968, Chapter VII, section 1)) *If Z is a standard normal random variable and $x > 0$, then*

$$\frac{1}{\sqrt{2\pi}} \left(\frac{1}{x} - \frac{1}{x^3}\right) e^{-x^2/2} < \mathbb{P}[Z > x] < \frac{1}{\sqrt{2\pi}} \left(\frac{1}{x}\right) e^{-x^2/2}.$$

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 20 implies that, for all κ

$$\left|\mathbb{P}\left[2\sqrt{\ell}M > \kappa\right] - \mathbb{P}[Z > \kappa]\right| \leq \frac{1}{\sqrt{\ell}}.$$

Using $\kappa = 2\eta\sqrt{\ell}$,

$$\mathbb{P}[M > \eta] \geq \mathbb{P}\left[Z > 2\eta\sqrt{\ell}\right] - \frac{1}{\sqrt{\ell}}. \tag{28}$$

Applying Lemma 21, we get

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

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

$$\begin{aligned}\mathbb{P}\left[Z > 2\eta\sqrt{\ell}\right] &\geq \frac{1}{\sqrt{2\pi}} \left(\frac{1}{2} - \frac{1}{8}\right) \frac{1}{\eta\sqrt{\ell}} e^{-2\eta^2\ell} \\ &\geq \frac{1}{7\eta\sqrt{\ell}} e^{-2\eta^2\ell}.\end{aligned}$$

Combining with (28) completes the proof of (8). □

A.4 Proof of (9)

We follow the proof of Proposition 7.3.2 in (Matoušek and Vondrak, 2011, Page 46).

Lemma 22 *For n even, let $U_1, \ldots, U_n$ be i.i.d. RVs with $\mathbb{P}[U_1 = 0] = \mathbb{P}[U_1 = 1] = 1/2$ and $U = \sum_{i=1}^{n}$. Then for integer $t \in [0, \frac{n}{8}]$,*

$$\mathbb{P}\left[U \geq \frac{n}{2} + t\right] \geq \frac{1}{5} e^{-16t^2/n}.$$

Proof Let integer $m = n/2$.

$$\mathbb{P}[U \geq m + t] = 2^{-2m} \sum_{j=t}^{m} \binom{2m}{m+j} \tag{29}$$

$$\geq 2^{-2m} \sum_{j=t}^{2t-1} \binom{2m}{m+j} \tag{30}$$

$$= 2^{-2m} \sum_{j=t}^{2t-1} \binom{2m}{m} \cdot \frac{m}{m+j} \cdot \frac{m-1}{m+j-1} \cdots \frac{m-j+1}{m+1} \tag{31}$$

$$\geq \frac{1}{2\sqrt{m}} \sum_{j=t}^{2t-1} \prod_{i=1}^{j} \left(1 - \frac{j}{m+1}\right) \qquad \text{using } \tbinom{2m}{m} \geq 2^{2m}/2\sqrt{m} \tag{32}$$

$$\geq \frac{t}{2\sqrt{m}} \left(1 - \frac{2t}{m}\right)^{2t} \tag{33}$$

$$\geq \frac{t}{2\sqrt{m}} e^{-8t^2/m} \qquad \text{since } 1 - x \geq e^{-2x} \text{ for } 0 \leq x \leq 1/2. \tag{34}$$

For $t \geq \frac{1}{2}\sqrt{m}$, the last expression is at least $\frac{1}{4}e^{-16t^2/n}$.

Note that $\mathbb{P}[U = m] = 2^{-2m}\binom{2m}{m} \leq 1/\sqrt{\pi m}$. Thus for $0 \leq t < \frac{1}{2}\sqrt{m}$, we have

$$\mathbb{P}[U \geq m + t] \geq \frac{1}{2} - t\,\mathbb{P}[U = m] \tag{35}$$

$$\geq \frac{1}{2} - \frac{1}{2}\sqrt{m}\frac{1}{\sqrt{\pi m}} \tag{36}$$

$$\geq \frac{1}{2} - \frac{1}{2\sqrt{\pi}} \approx 0.218 \geq \frac{1}{5} \geq \frac{1}{5}e^{-16^2/n} \tag{37}$$

Thus the bound $\frac{1}{5}e^{-16t^2/n}$ holds for all $0 \leq t \leq m/4$. ■

A.5 Proof of (10)

The proof of (10) uses the next two lemmas and and follows the proof of Lemma 5.1 in (Anthony and Bartlett, 1999).

Lemma 23 (Slud's Inequality, (Slud, 1977)) *Let B be a binomial (ℓ, p) random variable with $p \leq 1/2$. Then for $\ell(1-p) \geq j \geq \ell p$,*

$$\mathbb{P}[B \geq j] \geq \mathbb{P}\left[Z \geq \frac{j - \ell p}{\sqrt{\ell p(1-p)}}\right]$$

where Z is a standard normal random variable.

Lemma 24 ((see Anthony and Bartlett, 1999, Appendix 1)) *If Z is a standard normal and $x > 0$ then*

$$\mathbb{P}[Z \geq x] \geq \frac{1}{2}\left(1 - \sqrt{1 - e^{-x^2}}\right).$$

Recall that in (10) U the sum of the ℓ i.i.d. boolean random variables, each of which is 1 with probability $\frac{1}{2}+\eta$. Let B be a random variable with the binomial $(\ell, \frac{1}{2}-\eta)$ distribution.

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{\ell}U < 1/2\right] &= \mathbb{P}[B \geq \ell/2] \\
&\geq \mathbb{P}\left[N \geq \frac{\ell/2 - \ell(1/2-\eta)}{\sqrt{\ell(1/2+\eta)(1/2-\eta)}}\right] && \text{Slud's Inequality} \\
&= \mathbb{P}\left[N \geq \frac{2\eta\sqrt{\ell}}{\sqrt{(1-4\eta^2)}}\right] \\
&\geq \frac{1}{2}\left(1 - \sqrt{1 - \exp\left(-\frac{4\eta^2\ell}{1-4\eta^2}\right)}\right) \\
&\geq \frac{1}{4}\exp\left(-\frac{4\eta^2\ell}{1-4\eta^2}\right) && \text{since } 1-\sqrt{1-x} > x/2 \\
&\geq \frac{1}{4}\exp\left(-5\eta^2\ell\right) && \text{when } \eta \leq 1/5
\end{aligned}$$

completing the proof of (10).

References

F. Abramovich, Y. Benjamini, D. Donoho, and I. M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Annals of Statistics*, 34:584–653, 2006.

L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi. On combinatorial testing problems. *The Annals of Statistics*, 38(5):3063–3092, 2010.

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

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

M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

P. Bickel and E. Levina. 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.

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

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

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

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

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

D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32:952–994, 2004.

D. Donoho and J. Jin. Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *Annals of Statistics*, 34, 2006.

D. Donoho and J. Jin. Higher criticism thresholding: optimal feature selection when useful features and rare and weak. *PNAS*, 105(39):14790–14795, 2008.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd ed.)*. Wiley, 2000.

A. Ehrenfeucht, D. Haussler, M. Kearns, and L. G. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–251, 1989.

J. Fan and Y. Fan. High dimensional classification using features annealed independence rules. *The Annals of Statistics*, 36(6):2605Äì–2637, 2008.

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

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

D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):129–161, 1994.

D. P. Helmbold and P. M. Long. On the necessity of irrelevant variables. *ICML*, 2011.

J. Jin. Impossibility of successful classication when useful features are rare and weak. *PNAS*, 106(22):8859–8864, 2009.

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

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

N. Meinshausen and J. Rice. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Annals of Statistics*, 34(1):373–393, 2006.

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

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

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

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

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

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

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

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

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

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