Comparison and Design of Neyman-Pearson Classifiers

Clayton Scott*
cscott@rice.edu

October 2005

Abstract

In the Neyman-Pearson (NP) classification paradigm, the goal is to learn a classifier from labeled training data such that the probability of a false negative is minimized while the probability of a false positive is below a user-specified level $\alpha \in (0, 1)$. This work addresses the question of how to evaluate and compare classifiers in the NP setting. Simply reporting false positives and false negatives leaves some ambiguity about which classifier is best. Unlike conventional classification, however, there is no natural performance measure for NP classification. We cannot reject classifiers whose false positive rate exceeds α since, among other reasons, the false positive rate must be estimated from data and hence is not known with certainty.

We propose two families of performance measures for evaluating and comparing classifiers and suggest one criterion in particular for practical use. We then present general learning rules that satisfy performance guarantees with respect to these criteria. As in conventional classification, the notion of uniform convergence plays a central role, and leads to finite sample bounds, oracle inequalities, consistency, and rates of convergence. The proposed performance measures are also applicable to the problem of anomaly detection.

1 Introduction

In the Neyman-Pearson (NP) classification paradigm, the goal is to learn a classifier from labeled training data such that the probability (with respect to the true, unknown data distribution) of a false negative is minimized while the probability of a false negative is below a user-specified level $\alpha \in (0, 1)$. Unlike classical NP hypothesis testing, NP classification does not assume knowledge of the class-conditional densities, but only a finite sample of labeled observations.

This framework is an important alternative to other, more common approaches to classification that seek to minimize the probability of error or expected Bayes' cost. One advantage of the NP paradigm is that in many important applications, such as disease classification or network intrusion detection, it is more natural to specify a constraint on the false positive probability than to assign costs to the different kinds of errors. A second and no less important advantage is that, unlike decision-theoretic approaches, NP classification does not rely in any way on knowledge of a priori class probabilities. This is extremely important in applications where the class frequencies in the training data do not accurately reflect class frequencies in the larger population. For example, the frequencies of diseased and normal patients at a research hospital, where training data might be

*Department of Statistics, Rice University, 6100 Main St, MS-138, Houston, TX 77005. Supported by an NSF VIGRE postdoctoral training grant.

gathered, in all likelihood do not reflect the class frequencies in the population at large. In fact, it could probably be argued that most classification problems of interest fit this description.

In this paper we consider the problem of evaluating and comparing classifiers in the NP context. Suppose we train two classifiers (e.g., two support vector machines with different parameter settings) and compute empirical estimates of the false positive and false negative probabilities for each. Given α, which classifier should we choose? In some cases there may be an expert who can decide. But what if no expert is available? Or, what if we want to compare two learning algorithms across several datasets? Even if an expert is available, experts usually cost money or have limited time, and their decisions may be subjective or prone to human error. It seems desirable to have an objective criterion for evaluating and comparing the trained classifiers.

Before proceeding, let us briefly introduce some basic notations. Let f denote a classifier, and let $R_0(f)$ and $R_1(f)$ denote the true false positive and false negative probabilities, respectively. Also let $\tilde{R}_0(f)$ and $\tilde{R}_1(f)$ denote estimates of $R_0(f)$ and $R_1(f)$ (e.g., based on an independent test sample). Denote by f^*_α the classifier that minimizes $R_1(f)$ subject to $R_0(f) \leq \alpha$. This classifier is analogous to the Bayes classifier in conventional classification. Finally, set $\beta_\alpha = R_1(f^*_\alpha)$.

The approach adopted in this paper is to employ a scalar quantity that reflects the performance of a classifier and that can be estimated reliably. For example, in conventional classification, that scalar quantity is usually taken to be the probability of misclassification. By "reflects the performance" we mean, at the very least, that the global minimizer of the performance measure should be f^*_α.

One candidate for such a performance measure is to assign a classifier f a "score" of $R_1(f)$ if $R_0(f) \leq \alpha$, and ∞ otherwise. This is clearly minimized by f^*_α and it favors any classifier satisfying the constraint to any classifier that does not. Although such a measure may be appropriate for classical hypothesis testing, there are at least three reasons why it is impractical when learning from data.

First, learning rules produce classifiers from random training samples, and hence the false positive rate of a classifier is itself a random quantity. The performance of an algorithm on one dataset may not give a fair indication of its performance on others. This contrasts with conventional classification, where the probability of error for an unfavorable training sample will at least be somewhat close the typical probability of error for that distribution. In terms of expectation, the expected "score" of a learning rule will be infinite as long as there is some collection of training samples (that occurs with nonzero probability) for which $R_0(f) > \alpha$. It seems preferable for a performance measure to show more leniency to violations of the false positive constraint so that "rare" training samples do not mislead us about a learning rule's typical performance.

Second, it is not possible to estimate the performance from data precisely. Estimates of $R_0(f)$ and $R_1(f)$ are based on random data and thus will have some chance error. It is generally impossible to be certain whether a given classifier does or does not satisfy the false positive constraint. In fact, suppose $\tilde{R}_0(f)$ and $\tilde{R}_1(f)$ are estimates based on an independent test sample. If we estimate the "score" of f with $\tilde{R}_1(f)$ when $\tilde{R}_0(f) \leq \alpha$ and with ∞ otherwise, then the bias of this estimate is infinite whenever $0 < R_0(f) \leq \alpha$.

Third, many practitioners would be content to have the false positive rate slightly exceed the constraint if the decrease in the false negative rate is substantial. In other words, it might be nice to explore a small region along the receiver operating characteristic in the vicinity of α. Unlike cost-sensitive classification, however, which also allows trading off one error for another, we still require an operating point near α.

To alleviate these problems we could instead measure performance with $c\mathbb{I}_{R_0(f)>\alpha} + R_1(f)$, where $c > 1$ and $\mathbb{I}$ is an indicator. Yet the same three arguments used above for the case $c = \infty$ still apply. For example, suppose we estimate this quantity with a plug-in estimate based on an independent test sample. This estimate has a very large bias when the true R_0 is close to α, even for large test sample sizes. In fact, if $R_0(f) = \alpha$, the bias does not tend to zero as the test sample size increases.

In summary, a good performance measure for NP classification should not reject outright a classifier that appears to violate the false positive constraint. Such a violation may be (1) the result of a rare training sample on which an otherwise good learning rule performs poorly, (2) the result of chance error in estimating the false positive rate, and (3) acceptable to a practitioner.

Another possibility, which does show some leniency to classifiers violating the false positive constraint, is to measure performance with $|R_0(f) - \alpha| + |R_1(f) - \beta_\alpha|$ or some other "distance" in the (R_0, R_1) plane. Yet this kind of measure too has drawbacks. It requires knowledge of β_α which is unknown, and hence the distance cannot be estimated. In addition, it penalizes classifiers for having $R_0(f) < \alpha$ or $R_1(f) < \beta_\alpha$, which seems unreasonable.

In addition to the question of how to measure performance in NP classification, this paper addresses a second question that follows naturally from the first: Having settled on a performance criterion, is it possible to design a learning rule tailored to that criterion?

Previous work on NP classification has not dealt with either of these two questions. Both Cannon, Howse, Hush, and Scovel [1] and Scott and Nowak [2] prove generalization error bounds for learning rules $\widehat{f}_\alpha$ based on (penalized) empirical risk minimization. Yet these bounds are of the form "with high probability, $R_0(\widehat{f}_\alpha)$ is close to α and $R_1(\widehat{f}_\alpha)$ is close to its optimal value." The false positive and false negative probabilities are treated separately, and combining them into a single performance measure is not discussed. A more extensive comparison with these works is given in Section 2.4.

In Section 2 we propose two families of criteria for evaluating classifiers in the NP paradigm, and present learning rules that satisfy performance guarantees with respect to these criteria. The rules are given by (in one case constrained) penalized empirical risk minimization. The uniform convergence of certain empirical processes is seen to play a key role in bounding their performance, as in the theory of statistical learning for conventional classification [3, 4]. Uniform convergence in turn leads to finite sample bounds, oracle inequalities, consistency, and rates of convergence.

Section 3 gives examples of complexity penalties for NP classification, which are key ingredients in the proposed learning rules. Generalizations of one of the oracle inequalities are given in Section 4. Some practical aspects of estimating our performance measures from data are presented in Section 5. Section 6 discusses extensions and applications of the main results, including convex upper bounds and anomaly detection. The proofs are gathered in an appendix.

2 Main Results

To formalize the problem, let $\mathcal{X} \subset \mathbb{R}^d$ be a space of possible patterns, and let X be a random variable on $\mathcal{X}$, corresponding to an observed pattern. Let $\mathbf{P}_0$ and $\mathbf{P}_1$ be two distributions of X. Let (x, y) denote a realization of X, where $y = 0, 1$ indicates the data generating distribution $\mathbf{P}_y$. A *classifier* is a Borel measurable function $f : \mathcal{X} \to \{0, 1\}$. If f is a classifier and (x, y) a realization, a *false positive* occurs when $f(x) = 1$ but $y = 0$. Similarly, a *false negative* occurs when $f(x) = 0$ but $y = 1$. Let $R_0(f) = \mathbf{P}_0(f(X) = 1)$ and $R_1(f) = \mathbf{P}_1(f(X) = 0)$ denote the false positive and

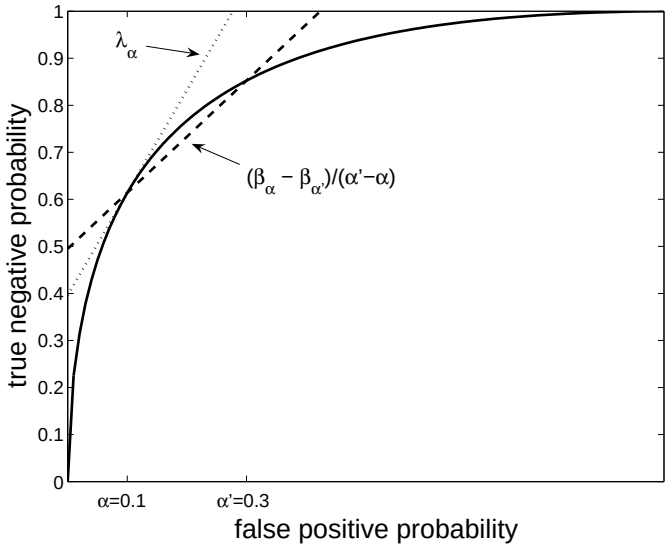


Figure 1: Lemma 1 can be understood in terms of properties of the ROC of f_α^*. The solid line is the ROC of a hypothetical f_α^*. The dashed line passes through the ROC at $\alpha = 0.1$ and $\alpha' = 0.3$ and has slope $(\beta_\alpha - \beta_{\alpha'})/(\alpha' - \alpha)$. The dotted line is tangent to the ROC at $\alpha = 0.1$ and has slope λ_α. Lemma 1 follows by concavity of the ROC.

false negative probabilities (sometimes called *rates*) of f.

Let $\alpha \in (0, 1)$. In Neyman-Pearson classification one seeks the classifier f_α^* for which $R_1(f)$ is minimized subject to $R_0(f) \leq \alpha$. Throughout this work we assume

A1 X has class-conditional Lebesgue densities h_0 and h_1.

A2 For each $\lambda > 0$, $\mathbf{P}_0(h_1(x) = \lambda h_0(x)) = \mathbf{P}_1(h_1(x) = \lambda h_0(x)) = 0$.

Under these assumptions, the Neyman-Pearson lemma [5] states that f_α^* is given by a likelihood ratio test (LRT): $f_\alpha^* = \mathbb{I}_{h_1(x) > \lambda_\alpha h_0(x)}$, where λ_α is the unique number such that $\int_{h_1(x) > \lambda_\alpha h_0(x)} h_0(x)\, dx = \alpha$. Assumption **A2** guarantees the existence of λ_α and also implies that we may also write $f_\alpha^* = \mathbb{I}_{h_1(x) \geq \lambda_\alpha h_0(x)}$. Recall $\beta_\alpha = R_1(f_\alpha^*) = \int_{h_1(x) < \lambda_\alpha h_0(x)} h_1(x)\, dx$ denotes the false negative rate of f_α^*.

We adopt **A1** and **A2** because they ensure that the receiver operating characteristic (ROC) of the LRT is concave. This allows us to bound $\beta_\alpha - \beta_{\alpha'}$ in terms of $\alpha - \alpha'$ using the slope of the ROC at α. In particular, we have the following.

Lemma 1. *If $\alpha' > \alpha$, then $\beta_\alpha - \beta_{\alpha'} \leq \lambda_\alpha(\alpha' - \alpha)$. If $\alpha' < \alpha$, then $\beta_{\alpha'} - \beta_\alpha \geq \lambda_\alpha(\alpha - \alpha')$.*

A geometric proof of the lemma comes from considering properties of the ROC of f_α^* [6]. In particular, the ROC of f_α^* is concave and has slope λ_α at α. The lemma follows from these properties and the realization that $(\beta_\alpha - \beta_{\alpha'})/(\alpha' - \alpha)$ is the slope of the line passing through the ROC at α and α'. See Fig. 1. An analytic proof of the lemma is given in the appendix.

Without assumptions **A1** and **A2** (for example, when X is discrete) the ROC of the LRT might have discontinuities and could be non-concave. One theoretical fix is the introduction of randomized classifiers that choose between two non-randomized classifiers based on a coin flip. This guarantees that the ROC of f_α^* is concave for general distributions. However, such classifiers

are not amenable to our analysis or require computationally implausible learning rules. That said, the results stated later in this section do extend to more general distributions, provided that α is such that $R_0(f^*_\alpha) = \alpha$, i.e., the optimal classifier satisfies the constraint with equality. In this special case, randomized classifiers are not needed.

Let $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ denote a training sample. In NP classification, two sampling plans may be considered [1]. Under the *joint* sampling plan, the numbers n_0 and n_1 of examples from each class are not known until the sample is gathered, and the data are iid realizations of some joint distribution $\mathbf{P}$ on (X, Y). Under the *conditional* sampling plan, n_0 and n_1 are known a priori, and the data are iid realizations of their respective conditional distributions. The difference is that in the former case, n_0 and n_1 are random, whereas in the latter case they are non-random. We will see later that, in the presented framework, the philosophical differences between the two sampling plans can be ignored.

The empirical errors with respect to S are defined by $\widehat{R}_0(f) = (1/n_0)\sum_{i:Y_i=0} \mathbb{I}_{f(X_i)=1}$ and $\widehat{R}_1(f) = (1/n_1)\sum_{i:Y_i=1} \mathbb{I}_{f(X_i)=0}$. A *learning rule* maps training samples to classifiers. We denote a learning rule by $\widehat{f}_\alpha$, and we also use the same notation to denote the classifier returned when the training sample is given.

2.1 Comparing classifiers

We propose two families of performance measures, each indexed by a parameter $0 < \kappa \leq \infty$. They capture what will be referred to as the *net NP error*, or simply net error. The first is

$$\mathcal{M}^\kappa(f) = \kappa \left(R_0(f) - \alpha\right)_+ + \left(R_1(f) - \beta_\alpha\right)_+ , \tag{1}$$

where $(x)_+ := \max(x, 0)$. The second is

$$\mathcal{N}^\kappa(f) = \kappa \left(R_0(f) - \alpha\right)_+ + R_1(f) - \beta_\alpha. \tag{2}$$

The parameter κ is a tradeoff parameter that controls the penalty for exceeding the constraint. We allow $\kappa = \infty$, which corresponds to strict enforcement of the constraint, although as argued in the introduction, this case has questionable value. Note that κ is different from a "cost" parameter used in cost-sensitive classification. If the $(\cdot)_+$ operators were not present, these criteria would be like cost-sensitive risks. Yet with the $(\cdot)_+$ operators, κ only penalizes false positive rates *in excess* of α.

Each of the two families of error measures has its pros and cons. The first family has the desirable property that, for any value of κ, it is minimized by the optimal classifier f^*_α. Yet it also has a serious drawback: it depends on β_α, which depends on the unknown distribution. Thus, it is useful for theoretical analysis and simulation studies, but it cannot be reliably estimated with real-world data.

For the second family the story is reversed. Even though β_α appears, it is simply a constant additive term that can be ignored for the purposes of minimization or comparison. Thus, reliable estimation (up to the unknown constant) based on an independent test sample is possible. On the downside, minimizing this net error does not result in f^*_α for all values of κ. If κ is small, and since (in contrast to the first family) it pays to take $R_1(f)$ below β_α, the minimizing classifier might have $R_0(f) > \alpha$. The following result tells us which values of κ are meaningful.

Proposition 1. *If $\kappa \geq \lambda_\alpha$ then $\mathcal{N}^\kappa(f^*_\alpha) \leq \mathcal{N}^\kappa(f)$ for all classifiers f.*

Proof. It suffices to consider only the classifiers $f^*_{\alpha'}, \alpha' \in [0,1]$, since each of these has optimal false negative rate for false positive rate α'. There are two cases. If $\alpha' < \alpha$, then $\mathcal{N}^\kappa(f^*_\alpha) - \mathcal{N}^\kappa(f^*_{\alpha'}) = \beta_\alpha - \beta_{\alpha'} \leq 0$. If $\alpha' > \alpha$, then by Lemma 1 we have

$$\begin{aligned} \mathcal{N}^\kappa(f^*_\alpha) - \mathcal{N}^\kappa(f^*_{\alpha'}) &= \kappa(\alpha - \alpha') + (\beta_\alpha - \beta_{\alpha'}) \\ &\leq \kappa(\alpha - \alpha') + \lambda_\alpha(\alpha' - \alpha) \\ &= (\lambda_\alpha - \kappa)(\alpha' - \alpha) \\ &\leq 0 \end{aligned}$$

since $\kappa \geq \lambda_\alpha$. □

Since λ_α is in general unknown, an upper bound on λ_α is needed. The following lemma is one (perhaps crude) possibility.

Lemma 2. $\lambda_\alpha \leq 1/\alpha$.

Proof. Taking $\alpha' = 0$ in Lemma 1 we have $\lambda_\alpha \leq (1 - \beta_\alpha)/\alpha \leq 1/\alpha$. □

Thus, if $\kappa \geq 1/\alpha$, the criterion in (2) is meaningful (it is minimized by the target classifier f^*_α) and it can be estimated from data. For practical applications we suggest the *NP score*

$$\mathcal{E}(f) = \frac{1}{\alpha}\left(R_0(f) - \alpha\right)_+ + R_1(f) \tag{3}$$

as the criterion for evaluation and comparison of NP classifiers. In addition to being minimized by f^*_α, it can be accurately estimated from a test sample (see Section 5.1 for details). Furthermore, it has the appealing property that as α draws closer to 0, a stiffer penalty is exacted on classifiers that violate the constraint. This makes sense because exceeding α by 0.01, for example, is much more significant when $\alpha = 0.01$ than when $\alpha = 0.1$. Said another way, the NP score in (3) penalizes the *relative* error $(R_0(f) - \alpha)/\alpha$.

2.2 Designing classifiers

Having addressed the first of our motivating questions, we now turn to the second. How can one learn with respect to the error measures put forth above? More precisely, are there learning rules that obey performance guarantees when performance is measured in terms of these error measures?

The learning problem here is considerably different from decision-theoretic statistical estimation problems such as conventional classification. There, one seeks to optimize an expected loss, where the loss measures the cost of making an error. The $(\cdot)_+$ operator is also a kind of loss function, but in NP classification it operates after the expectation with respect to the random pattern.

We propose two general learning rules whose net NP errors are governed by oracle inequalities. Before introducing the rules, the following definition is needed. Let $\mathcal{F}$ be a set of classifiers, and let ϕ_0 and ϕ_1 be functions of $f \in \mathcal{F}$, the training sample S, and a confidence parameter δ.

Definition 1. *We say* (ϕ_0, ϕ_1) *are a (distribution free)* complexity penalty pair *for* $\mathcal{F}$ *if and only if for all distributions and all* $\delta_0, \delta_1 \in (0,1)$*,*

$$\mathbf{P}_S\left(\left\{S : \sup_{f \in \mathcal{F}}\left(\left|R_0(f) - \widehat{R}_0(f)\right| - \phi_0(f, S, \delta_0)\right) > 0\right\}\right) \leq \delta_0$$

and

$$\mathbf{P}_S\left(\left\{S : \sup_{f\in\mathcal{F}}\left(\left|R_1(f) - \widehat{R}_1(f)\right| - \phi_1(f,S,\delta_1)\right) > 0\right\}\right) \leq \delta_1,$$

where $\mathbf{P}_S$ *denotes the probability with respect to the appropriate (joint or conditional) sampling plan.*

Thus, ϕ_0 and ϕ_1 control the rate of uniform convergence of the class-conditional frequencies to the true class-conditional probabilities. In the results below we do not specify the sampling plan explicitly. Rather, we assume that (ϕ_0, ϕ_1) are a complexity penalty pair with respect to whatever the sampling plan happens to be. Examples of complexity penalties for NP classification under both sampling plans are discussed in Section 3.

Both learning rules that we propose assume a given collection $\mathcal{F}$ of candidate classifiers, as well as a complexity penalty pair for $\mathcal{F}$. Each rule has a separate formulation for the case $\kappa = \infty$, and these formulations happen to coincide, so there are three rules total. For now we consider only the first two; the case $\kappa = \infty$ is treated in Section 4.2. Each learning rule also has two parameters. The first, $\nu \in \mathbb{R}$, controls the tradeoff between false positives and false negatives. The second, $\gamma > 0$, is not a tuning parameter; for any given ν, κ, and α, the results below specify the appropriate choice of γ. It is treated as a variable to simplify the presentation.

The first learning rule, based on a constrained minimization and referred to as NP-CON, is

$$\widehat{f}_\alpha^c = \arg\min_{f\in\widehat{\mathcal{F}}_\alpha^\nu} \quad \widehat{R}_1(f) + \phi_1(f,S,\delta_1) + \gamma\phi_0(f,S,\delta_0). \tag{4}$$

Here we define $\widehat{\mathcal{F}}_\alpha^\nu = \{f \in \mathcal{F} : \widehat{R}_0(f) \leq \alpha + \nu\phi_0(f,S,\delta_0)\}$. The following result uses the notation $\mathcal{F}_\alpha = \{f \in \mathcal{F} : R_0(f) \leq \alpha\}$.

Theorem 1. *Assume* (ϕ_0, ϕ_1) *are a complexity penalty pair for* $\mathcal{F}$*. Fix* $0 < \kappa < \infty$*. Let* $\delta_0, \delta_1 > 0$ *and let* $\widehat{f}_\alpha^c$ *be the rule NP-CON* (4) *with* $\nu = 1$ *and* $\gamma \geq 2(\kappa + \lambda_\alpha)$*. Then*

$$\mathcal{M}^\kappa(\widehat{f}_\alpha^c) \leq \inf_{f\in\mathcal{F}_\alpha}\left\{\ \gamma\phi_0(f,S,\delta_0) + R_1(f) - \beta_\alpha + 2\phi_1(f,S,\delta_1)\ \right\}$$

with probability at least $1 - (\delta_0 + \delta_1)$ *with respect to the draw of the training sample* S*. Moreover, the same result holds if we replace* $\mathcal{M}^\kappa$ *by* $\mathcal{N}^\kappa$*, where now it suffices to have* $\gamma \geq 2\kappa$*.*

Brief remarks:

- The result only holds for $\nu = 1$. In Section 4 we present a more general oracle inequality for this rule that applies for all values of ν.
- Recall that γ is not a free parameter. We see that the appropriate choice of γ is $2(\kappa + \lambda_\alpha)$ if performance is to be measured by $\mathcal{M}^\kappa$, and 2κ if measured by $\mathcal{M}^\kappa$. Since λ_α is usually unknown, it may be replaced by the upper bound $1/\alpha$ of Lemma 2.

The upper bound of Theorem 1 is called an oracle bound. The second term is the (deterministic) approximation error for the false negative probability. It is the amount of "excess false negative probability" that results from a particular classifier satisfying the constraint. The first and third terms bound the (stochastic) estimation errors of the false positive and false negative probabilities, respectively. In other words, these terms bound the deviation between true and empirical error.

In general, the approximation and estimation errors will vary in magnitude depending on the unknown distribution. The oracle inequality says that $\widehat{f}^c_\alpha$ performs about as well as the classifier selected by an oracle to optimize the tradeoff between these error terms. This implies an adaptive rate of convergence, as discussed in Section 2.3 below.

The second learning rule, based on an unconstrained minimization and referred to as NP-UNC, is

$$\widehat{f}^u_\alpha = \arg\min_{f\in\mathcal{F}} \quad \kappa\left(\widehat{R}_0(f) - \alpha - \nu\phi_0(f,S,\delta_0)\right)_+ + \gamma\phi_0(f,S,\delta_0) + \widehat{R}_1(f) + \phi_1(f,S,\delta_1) \tag{5}$$

This rules obeys its own oracle inequality.

Theorem 2. *Assume (ϕ_0,ϕ_1) are a complexity penalty pair. Let $\delta_0,\delta_1>0$ and let $\widehat{f}^u_\alpha$ be as in (5) with $\nu\in\mathbb{R}$ and $\gamma=\kappa(1+\nu)$. Then*

$$\mathcal{N}^\kappa(\widehat{f}^u_\alpha) \le \inf_{f\in\mathcal{F}}\left\{\ \kappa\left(R_0(f)-\alpha\right)_+ + 2\kappa\phi_0(f,S,\delta_0) + R_1(f) - \beta_\alpha + 2\phi_1(f,S,\delta_1)\ \right\}$$

with probability at least $1-(\delta_0+\delta_1)$ with respect to the draw of S.

The two rules and corresponding bounds each have their pros and cons. Theorem 2 applies for all values of ν, whereas Theorem 1 assumes $\nu=1$. A generalization of Theorem 1 is given in Section 4, although this bound is somewhat less appealing than the form in Theorem 1. Even if we compare Theorem 1 and Theorem 2 for $\nu=1$, the inf in Theorem 2 is over all of $\mathcal{F}$. We can specialize to $\mathcal{F}_\alpha$ and recover the exact bound of Theorem 1. Thus the bound in Theorem 2 is at least as small as the bound in Theorem 1.

On the other hand, NP-CON has the advantage that Theorem 1 applies to both $\mathcal{M}^\kappa$ and $\mathcal{N}^\kappa$, whereas Theorem 2 only applies to $\mathcal{N}^\kappa$. Furthermore, as shall be made precise in Theorem 3 below, NP-CON enjoys a guaranteed bound on its false positive rate. Moreover, this bound applies to all ν, and when $\nu=-1$, the desired false positive constraint can be guaranteed with high probability. With NP-UNC, the ν parameter offers some control over the excess false positive rate, but no bound is known.

2.3 Consistency and rates of convergence

The finite sample bounds presented in this paper yield asymptotic results. Strong consistency follows easily via the Borel-Cantelli lemma, provided the class $\mathcal{F}$ grows with n in a suitable way. For example, consistency for VC classes or histograms holds under the conditions described in [2].

Rates of convergence also follow naturally.[1] For example, if the optimal classifier f^*_α belongs to a set with finite VC dimension, and if the learning rules (with the corresponding VC penalty) are applied to that VC class, the resulting net NP error decays to zero like $O(\sqrt{\log n/n})$, which is within a log factor of the standard $1/\sqrt{n}$ rate for VC classes (see [3], ch. 14).

More significantly, our oracle bounds give adaptive rates. Faster convergence is automatically achieved when the distribution is nice in some sense. For example, consider a learning rule based on $\mathcal{F}=\cup_{k=1}^K\mathcal{F}^k$, where each $\mathcal{F}^k$ is a VC class with VC dimension $\propto k$ (for example). Consider the class of distributions such that

$$R_1(f^k)-\beta_\alpha \approx k^{-s}$$

[1] "Rate" here refers to the decay of the performance measure to zero as the sample size increases, as opposed to false negative/positive "rates," which are just probabilities.

where

$$f^k = \arg\min_{f \in \mathcal{F}_\alpha^k}$$

and $s > 0$. Since the VC penalties are on the order of $\sqrt{k \log n/n}$, the oracle inequality (in Theorem 1, for example) is minimized by choosing $k \propto (n/\log n)^{1/(2s+1)}$, which yields a rate of $(\log n/n)^{s/(2s+1)}$. Clearly the optimal choice of k depends on s, which is unknown a priori. Yet because of the oracle inequality, our learning rules perform as if they knew s.

2.4 Relation to previous work

The theoretical foundations of NP classification were laid in [1] and [2]. Cannon et al. [1] consider learning a classifier from a fixed VC class $\mathcal{F}$ [3]. They study the rule

$$\widehat{f}_\alpha = \arg\min_{f \in \mathcal{F}} \{\widehat{R}_1(f) \mid \widehat{R}_0(f) \leq \alpha + \epsilon_0\}. \tag{6}$$

Under the conditional sampling plan, they are able to show that with high probability (going to one exponentially fast as $n_0, n_1 \to \infty$),

$$R_0(\widehat{f}_\alpha) \leq \alpha + 2\epsilon_0 \quad \text{and} \quad R_1(\widehat{f}_\alpha) \leq R_1(f_{\mathcal{F},\alpha}) + 2\epsilon_1,$$

for any $\epsilon_0, \epsilon_1 > 0$, and where $f_{\mathcal{F},\alpha} = \arg\min_{f \in \mathcal{F}} \{R_1(f) \mid R_0(f) \leq \alpha\}$. Under the joint sampling plan, they obtain much looser bounds that are impractical for not only their looseness but also their dependence on the unknown a priori class probabilities.

Scott and Nowak [2] show that if the tolerances ϵ_0 and ϵ_1 are chosen in a certain way as functions of n_0 and n_1, then $\widehat{f}_\alpha$ obeys the same performance guarantee under the joint sampling plan as under the conditional sampling plan. They also extend the analysis to learning from one of several candidate VC classes $\mathcal{F}^1, \ldots, \mathcal{F}^K$. For appropriate choices of $\epsilon_0(k), \epsilon_1(k), k = 1, \ldots, K$, they consider the rule that first computes $\widehat{f}_\alpha^k$ according to (6) for each k, and then selects

$$\widehat{f}_\alpha = \arg\min\{\widehat{R}_1(f) + \epsilon_1(k) \mid f = \widehat{f}_\alpha^k, k = 1, \ldots, K\}.$$

Here $\epsilon_0(k)$ and $\epsilon_1(k)$ are essentially a complexity penalty pair. Under both joint and conditional sampling plans, it is shown that with high probability,

$$R_0(\widehat{f}_\alpha) \leq \alpha + 2\epsilon_0(\widehat{k}) \quad \text{and} \quad R_1(\widehat{f}_\alpha) \leq \min_{k=1,\ldots,K} (R_1(f_{\mathcal{F},\alpha}) + 2\epsilon_1(k)) \tag{7}$$

where $\widehat{k}$ is the index of the class of the selected classifier. This might be called a semi-oracle bound, where the approximation and stochastic components of the excess false negative probability are balanced, but the excess false negative probability is not balanced with the excess false positive probability.

The present work improves on this earlier work in three respects. First, we allow for complexity penalties that are data-dependent or classifier-dependent, as opposed to restricting to VC or finite classes. Second, we balance the excess false positive rate with the excess false negative rate. The result in (7) balances the approximation and stochastic components of the excess false negative probability, but the excess false positive probability is not balanced with these two terms. Third, our rules have guaranteed performance with respect to the criteria proposed earlier.

3 Complexity Penalties for NP Classification

In this section we give several examples of complexity penalty pairs (ϕ_0, ϕ_1) under both conditional and joint sampling plans. The penalties have well known counterparts in conventional classification. Penalties under conditional sampling are essentially complexity penalties for classes of sets, and are closely related to concentration of measure inequalities. Furthermore, we show that a complexity penalty pair under the conditional sampling plan is also a complexity penalty pair under joint sampling. Thus, it is not necessary to be concerned with the philosophical differences between the two sampling plans.

3.1 Penalties under the conditional sampling plan

Partition the training sample S into S_0 and S_1 according to class. Suppose there are n_y samples of class $y = 0, 1$. Under the conditional sampling plan, (ϕ_0, ϕ_1) is a complexity penalty pair if for all $\delta_y \in (0, 1)$,

$$\mathbf{P}_y^{n_y}\left(\left\{S_y : \sup_{f\in\mathcal{F}}\left(\left|R_y(f) - \widehat{R}_y(f)\right| - \phi_y(f, S_y, \delta_y)\right) > 0\right\}\right) \le \delta_y,$$

$y = 0, 1$. Thus it suffices to choose ϕ_0 and ϕ_1 independently. Furthermore, the necessary condition above can be rephrased in terms of complexity penalties for classes of sets. The following definition was put forth in [7].

Definition 2. *Let $\mathcal{G}$ be a collection of measurable subsets of $\mathcal{X}$. Given a distribution $\mathbf{Q}$ on $\mathcal{X}$, let $T = \{X_1, \ldots, X_m\}$ denote an iid sample from $\mathbf{Q}$. We say ϕ is a (distribution free)* complexity penalty *for $\mathcal{G}$ if and only if for all distributions $\mathbf{Q}$ and all $\delta \in (0, 1)$,*

$$\mathbf{Q}_T\left(\left\{T : \sup_{G\in\mathcal{G}}\left(\left|\mathbf{Q}(G) - \widehat{\mathbf{Q}}(G)\right| - \phi(G, T, \delta)\right) > 0\right\}\right) \le \delta.$$

Here $\widehat{\mathbf{Q}}(G) = (1/m)\sum_{i=1}^m \mathbb{I}_{X_i\in G}$ is the empirical probability of G.

Let $\mathcal{G}_y$ be the collection of all sets of the form $G_f^y = \{x : f(x) = 1 - y\}$ ranging over $f \in \mathcal{F}$. If ϕ_y is a complexity penalty for the class of sets $\mathcal{G}_y$, $y = 0, 1$, then (ϕ_0, ϕ_1) is a complexity penalty pair according to Definition 1. When applying Definition 2, we take $\mathbf{Q} = \mathbf{P}_y$, $T = S_y$, $m = n_y$, $\delta = \delta_y$, and $\phi_y(f, S_y, \delta_y) = \phi(G_f^y, T, \delta)$. In summary, a complexity penalty pair for NP classification is nothing more than a pair of penalties for sets.

Several examples of complexity penalties for sets are given in [7]. We briefly recall some basic examples.

1. **VC penalty**: Suppose $\mathcal{F}$ has VC dimension $2 \le V < \infty$. Then $\mathcal{G}_y$ also has VC dimension V [3]. Let $\mathcal{G}$ be either $\mathcal{G}_0$ or $\mathcal{G}_1$. According to one version of the VC inequality [3], for any $\epsilon > 0$ and for any distribution $\mathbf{Q}$,

$$\mathbf{Q}_T\left(\sup_{G\in\mathcal{G}}\left|\mathbf{Q}(G) - \widehat{\mathbf{Q}}(G)\right| > \epsilon\right) \le 8m^V e^{-m\epsilon^2/32}.$$

Thus,

$$\phi(G, T, \delta) = \sqrt{32\frac{V\log m + \log(8/\delta)}{m}}$$

defines a complexity penalty for $\mathcal{G}$. The reader is reminded VC penalties were originally introduced as penalties for sets [8].

2. **Occam's Razor penalty**: Suppose $\mathcal{F}$ is countable and let $\{\pi_k\}_{k\geq 1}$ be a prior for $\mathcal{F}$: $\pi_k \geq 0$ and $\sum_k \pi_k = 1$. Let $\mathcal{G}$ be either $\mathcal{G}_0$ or $\mathcal{G}_1$. By Chernoff's bound [9], for any fixed $G_k \in \mathcal{G}$ and any $\epsilon > 0$ we have

$$\mathbf{Q}_T\left(\left|\mathbf{Q}(G_k) - \widehat{\mathbf{Q}}(G_k)\right| > \sqrt{\frac{\log(1/\pi_k) + \log(2/\delta)}{2m}}\right) \leq \pi_k \delta.$$

By the union bound

$$\phi(G, T, \delta) = \sqrt{\frac{\log(1/\pi_k) + \log(2/\delta)}{2m}}$$

defines a complexity penalty for $\mathcal{G}$.

3. **Rademacher penalty**: Let $\mathcal{G}$ be either $\mathcal{G}_0$ or $\mathcal{G}_1$, and assume that $\mathcal{G}$ satisfies the property that $G \in \mathcal{G} \Rightarrow \overline{G} \in \mathcal{G}$, where $\overline{G}$ denotes the compliment of G. Let $\sigma_1, \ldots, \sigma_n$ be Rademacher random variables, i.e., independent random variables taking on the values 1 and -1 with equal probability. Denote $\widehat{\mathbf{Q}}_{(\sigma_i)}(G) = \frac{1}{n}\sum_{i=1}^n \sigma_i \mathbb{I}_{X_i \in G}$. Define the conditional Rademacher average

$$\widehat{\rho}(\mathcal{G}, T) = \mathbf{E}_{(\sigma_i)}\left[\sup_{G \in \mathcal{G}} \widehat{\mathbf{Q}}_{(\sigma_i)}(G)\right],$$

where the expectation is with respect the Rademacher random variables. As discussed in [7],

$$\phi(G, T, \delta) = 2\widehat{\rho}(\mathcal{G}, T) + \sqrt{\frac{2\log(2/\delta)}{m}}$$

defines a complexity penalty for $\mathcal{G}$.

4. **Union penalty**: Suppose $\mathcal{F} = \cup_{k=1}^K \mathcal{F}^k$, where perhaps $K = \infty$, and (ϕ_0^k, ϕ_1^k) is a complexity penalty pair for each k. Let $\{\pi_k\}_{k\geq 1}$ be a prior. Then

$$\phi_y(f, S_y, \delta_y) = \phi_y^k(f, S_y, \delta_y \pi^k)$$

defines a complexity penalty pair for $\mathcal{F}$, where on the right hand side k is such that $f \in \mathcal{F}^k$.

3.2 Penalties under the joint sampling plan

The definition of complexity penalty is such that a simple conditioning argument gives us penalties under the joint sampling plan for free.

Proposition 2. *If (ϕ_0, ϕ_1) are a complexity penalty pair for $\mathcal{F}$ under the* conditional *sampling plan, then they are also a complexity penalty pair for $\mathcal{F}$ under the* joint *sampling plan.*

Proof. Define the event

$$\Omega_0 = \left\{S : \sup_{f \in \mathcal{F}} \left(\left|R_0(f) - \widehat{R}_0(f)\right| - \phi_0(f, S, \delta_0)\right) > 0\right\}.$$

By the law of total probability we have

$$\begin{aligned}\mathbf{P}_S(\Omega_0) &= \sum_{m=0}^{n} \mathbf{P}_S(S \in \Omega_0 \,|\, n_0 = m)\mathbf{P}_S(n_0 = m) \\ &\leq \sum_{m=0}^{n} \delta_0 \mathbf{P}_S(n_0 = m) \\ &= \delta_0.\end{aligned}$$

A similar argument applies to ϕ_1. □

4 Generalizations of Theorem 1

We present two generalizations of Theorem 1. The first extends NP-CON to the case $\nu \geq -1$, while the second extends both NP-CON and NP-UNC to the case $\kappa = \infty$.

4.1 NP-CON with $\nu \geq -1$

Theorem 1 is a special case of the following result. This more general result was deferred until now because its interpretation when $\nu \neq 1$ is somewhat less transparent. Introduce the notation $\mathcal{F}_\alpha^\nu = \{f \in \mathcal{F} : R_0(f) \leq \alpha - (1-\nu)\phi_0(f, S, \delta_0)\}$. Note that $\mathcal{F}_\alpha^1$ coincides with $\mathcal{F}_\alpha$ which was introduced with the statement of Theorem 1.

Theorem 3. *Let (ϕ_0, ϕ_1) be a complexity penalty pair, and let $\delta_0, \delta_1 > 0$. Define the* oracle bound

$$\mathcal{B}(\mathcal{F}, S, \nu, \gamma, \delta_0, \delta_1) = \inf_{f \in \mathcal{F}_\alpha^\nu} \left\{ \gamma\phi_0(f, S, \delta_0) + R_1(f) - \beta_\alpha + 2\phi_1(f, S, \delta_1) \right\}.$$

Let $\widehat{f}_\alpha^c$ be as in (4) with $\nu \geq -1$. If $\gamma \geq (\kappa + \lambda_\alpha)(1+\nu)$, then

$$R_0(\widehat{f}_\alpha^c) \leq \alpha + (1+\nu)\phi_0(\widehat{f}_\alpha^c, S, \delta_0)$$

and

$$\mathcal{M}^\kappa(\widehat{f}_\alpha^c) \leq \mathcal{B}(\mathcal{F}, S, \nu, \gamma, \delta_0, \delta_1)$$

with probability at least $1 - (\delta_0 + \delta_1)$ with respect to the draw of S. The same statement holds if we replace $\mathcal{M}^\kappa$ by $\mathcal{N}^\kappa$, where now it suffices to take $\gamma \geq \kappa(1+\nu)$.

The interpretation of the oracle when $\nu \neq 1$ is essentially the same as the case $\nu = 1$. To gain some insight, assume that $\mathcal{F} = \cup_{k=1}^K \mathcal{F}^k$, where ϕ_0 and ϕ_1 are constant on each $\mathcal{F}^k$. That is, for any S, $\phi_0(f, S, \delta_0) = \epsilon_0^k$ and $\phi_1(f, S, \delta_1) = \epsilon_1^k$ for $f \in \mathcal{F}^k$. This happens, for example, when $\mathcal{F}$ is a union of VC classes.

Further assume that the approximation error of the false negative probability decays (as a function of k) at a rate that is independent of α. Formally, define

$$\beta_\alpha^k = \inf_{\substack{f \in \mathcal{F}^k \\ R_0(f) \leq \alpha}} R_1(f).$$

Assume there exist constants $\underline{c}$ and $\overline{c}$ and a function τ_k tending to zero such that, for all $\alpha \in (0,1)$ and all k,

$$\underline{c}\tau^k \leq \beta_\alpha^k - \beta_\alpha \leq \overline{c}\tau^k.$$

Under these assumptions, and denoting $\alpha(k,\nu) = \alpha - (1-\nu)\epsilon_0^k$, we have

$$\mathcal{B}(\mathcal{F}, S, \nu, \gamma, \delta_0, \delta_1) = \min_{1\leq k\leq K} \left[\beta_{\alpha(k,\nu)}^k - \beta_\alpha + 2\epsilon_1^k + \gamma\epsilon_0^k\right].$$

By Lemma 1,

$$\begin{aligned}\beta_{\alpha(k,\nu)}^k - \beta_\alpha &= \beta_{\alpha(k,\nu)}^k - \beta_{\alpha(k,\nu)} + \beta_{\alpha(k,\nu)} - \beta_\alpha \\ &\leq \overline{c}\tau_k + \lambda_{\alpha(k,\nu)}(1-\nu)\epsilon_0^k.\end{aligned}$$

Using $\tau_k \leq \frac{1}{\underline{c}}(\beta_\alpha^k - \beta_\alpha)$, and introducing $C = (\overline{c}/\underline{c})$ and $C_k = \gamma + \lambda_\alpha(k,\nu)(1-\nu)$, we conclude

$$\begin{aligned}\mathcal{B}(\mathcal{F}, S, \nu, \gamma, \delta_0, \delta_1) &\leq \min_{1\leq k\leq K}\left[C(\beta_\alpha^k - \beta_\alpha) + 2\epsilon_1^k + C_k\epsilon_0^k\right] \\ &= \min_{1\leq k\leq K}\left[C\inf_{f\in\mathcal{F}_\alpha^k}(R_0(f) - \beta_\alpha) + \epsilon_1^k + C_k\epsilon_0^k\right] \\ &= \inf_{f\in\mathcal{F}_\alpha}\left[C_k\phi_0(f,S,\delta_0) + C\left(R_0(f) - \beta_\alpha\right) + 2\phi_1(f,S,\delta_1)\right].\end{aligned}$$

Thus, up to constants, the oracle bound when $\nu \neq 1$ has the same interpretation as when $\nu = 1$.

4.2 Absolute certainty: $\kappa = \infty$

Both NP-CON and NP-UNC have extensions, which happen to coincide, to the case $\kappa = \infty$. The basic idea is to take $\nu = -1$ and set $\gamma = 0$ in Equations (4) or (5).

Consider the learning rule

$$\widehat{f}_\alpha^i = \arg\min_{f\in\widehat{\mathcal{F}}_\alpha^{-1}} \quad \widehat{R}_1(f) + \phi_1(f,S,\delta_1). \tag{8}$$

Taking $\nu = -1$ and set $\gamma = 0$ in the proof of Theorem 3 gives the following.

Corollary 1. *Let (ϕ_0, ϕ_1) be a complexity penalty pair, and let $\delta_0, \delta_1 > 0$. Let $\widehat{f}_\alpha^i$ be as in (8). Then*

$$R_0(\widehat{f}_\alpha^i) \leq \alpha$$

and

$$R_1(\widehat{f}_\alpha^i) - \beta_\alpha \leq \inf_{f\in\mathcal{F}_\alpha^{-1}}\left\{\ R_1(f) - \beta_\alpha + 2\phi_1(f,S,\delta_1)\ \right\}$$

with probability at least $1-(\delta_0+\delta_1)$ with respect to the draw of S.

The interpretation of the oracle bound in this case is the same as for Theorem 3.

5 Practical Matters

The NP score

$$\mathcal{E}^{\kappa}(f) = \kappa\,(R_0(f) - \alpha)_+ + R_1(f),$$

$\kappa \geq 1/\alpha$, is envisioned not only as a theoretical tool but a performance criterion for the evaluation of NP learning algorithms with real data. This section examines some practical aspects of working with this error quantity.

5.1 Error estimation

Suppose we are given a classifier f and a test sample $\tilde{S}$ (independent of the training sample). How should we estimate the error $\mathcal{E}^{\kappa}(f) = \kappa\,(R_0(f) - \alpha)_+ + R_1(f)$? Our proposal is to do the obvious: compute the sample-based frequencies $\tilde{R}_0(f)$ and $\tilde{R}_1(f)$ and plug in to obtain the estimate

$$\tilde{\mathcal{E}}^{\kappa}(f) = \kappa\left(\tilde{R}_0(f) - \alpha\right)_+ + \tilde{R}_1(f). \tag{9}$$

As seen below, this estimator implements the maximum likelihood principle and, although biased, it has a smaller mean squared error (MSE) than if the $(\cdot)_+$ operator was not present.

Let us recast the problem in more general terms. Consider the problem of estimating $\theta = \theta_0 + \theta_1$, where $\theta_0 = \kappa\,(p_0 - \alpha)_+$ and $\theta_1 = p_1$, and p_0 and p_1 are the success probabilities of two Bernoulli trials. We observe n_y outcomes of the respective Bernoulli trials, of which k_y are assumed to be successes, $y = 0, 1$. Moreover, k_0 and k_1 are independent.

The maximum likelihood estimate of θ is $\kappa\,(k_0/n_0 - \alpha)_+ + k_1/n_1$. This is seen as follows. Since k_0 and k_1 are independent, the MLE of θ is the sum of the MLE's of θ_0 and θ_1. Clearly k_1/n_1 is the MLE of θ_1. The MLE of p_0 is also clearly k_0/n_0. Since θ_1 is simply a transformation of the parameter p_0, the claim follows from the invariance property of the MLE.

Observe that $\widehat{\theta_0} = \kappa\,(k_0/n_0 - \alpha)_+$ is biased. Fig. 2 (a) shows a plot of the bias of $\widehat{\theta_0}$ as a function of p_0 for $\alpha = 0.1$ and $n_0 = 100$. The bias is greatest when p_0 is near α, although it is still quite small, and this maximum bias will decrease for larger n_0. Furthermore, when bias and variance are considered together, it becomes apparent that the $(\cdot)_+$ operator actually makes θ easier to estimate. In particular, the following is true.

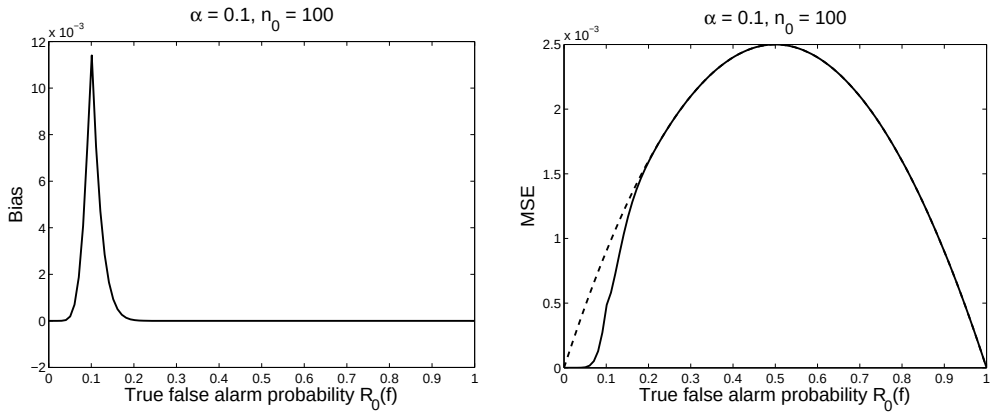
Proposition 3.

$$\mathbf{E}[(\widehat{\theta} - \theta)^2] \leq \kappa^2 \frac{p_0(1-p_0)}{n_0} + \frac{p_1(1-p_1)}{n_1}.$$

Note that the right-hand side is just the MSE of the standard unbiased estimator of $\kappa(p_0 - \alpha) + p_1$. Fig. 2 (b) is a plot of the MSE of $\widehat{\theta_0}$ (with $\kappa = 1$ for simplicity) as a function of p_0 together with a plot of the MSE of the MLE of $p_0 - \alpha$. The two graphs coincide for p_0 much larger than α, but the former MSE is clearly less when p_0 is near or below α.

5.2 Model selection

If $\mathcal{E}^{\kappa}$ is used as a criterion for model/parameter selection (for any learning algorithm, not necessarily the rules previously studied in this paper), the actual value of $R_0(\widehat{f_\alpha})$ will be less than the estimated value on average. Since having $R_0(\widehat{f_\alpha}) \leq \alpha$ is so important in NP classification, this conservative bias may actually be helpful in designing the classifier. This remark also applies to other error estimates, including those based on training data such as cross-validation or bootstrap.



(a) (b)

Figure 2: (a) The bias of the MLE of $(R_0(f) - \alpha)_+$ as a function of $R_0(f)$. Here $\alpha = 0.1$ and $n_0 = 100$. (b) The mean squared error (solid line) of the MLE of $(R_0(f) - \alpha)_+$ as a function of $R_0(f)$. Here $\alpha = 0.1$ and $n_0 = 100$. The dashed line represents the MSE the MLE of $R_0(f)$, for comparison.

5.3 Confidence intervals

Hoeffding's inequality and related concentration inequalities for bounded random variables may be applied to $\tilde{\mathcal{E}}^{\kappa}(f)$ to obtain distribution free confidence intervals [3, 10]. If one is interested in a confidence interval for $\tilde{R}_0(f) = \widehat{p_0}$, one may also appeal to the central limit theorem (CLT) and use the standard error estimate $\sqrt{\hat{p}_0(1-\hat{p}_0)/n_0}$. A common rule of thumb states that for the CLT approximation to apply to the binomial $\text{bin}(n_0, p_0)$ one needs $n_0 p_0 \geq 5$. If it can be assumed that $p_0 \approx \alpha$, then this gives a lower bound on the class zero test sample size n_0. If $\alpha = 0.1$, for example, than at least $5/\alpha = 50$ class zero test examples are need to apply the standard error-based confidence interval. Recently Langford has argued that the CLT approach is suspect for standard classification because the (unknown) probability of error could be very close to zero [10]. In NP classification, since α is known, this argument can be circumvented.

6 Extensions

In this concluding section we briefly discuss some extensions of our main results.

6.1 Criteria for Anomaly Detection

The proposed error measures can be easily applied to the problem of evaluating and comparing set estimates for anomaly detection. In anomaly detection, one is given training data from one distribution, and the objective is to construct a set where data following this distribution tend to be located. In one formulation, the goal is to construct a small region containing at least $100\alpha\%$ of the total probability mass of the normal data, where α is set by the user.

Performance can be measured in at least two ways. In the minimum volume set paradigm [11, 12], the objective is to find the set with minimum volume, where volume is calculated according to some (user-specified) reference measure, such as Lebesgue. The performance is summarized in

terms of the volume of the estimate and the closeness of the mass of the estimate to α. A second approach measures performance by counting false positives and false negatives on a new dataset that contains both normal and abnormal data.

In both cases, there is a second measure that comes in to play, in addition to the data-generating probability measure. Let us call these the inlier and outlier measures, $\mathbf{Q}_0$ and $\mathbf{Q}_1$. Then the problem is equivalent to Neyman-Pearson classification at the false positive constraint $1-\alpha$, where $1-\mathbf{Q}_0$ corresponds to the false positive probability and $\mathbf{Q}_1$ corresponds to the false negative probability. Thus, the quantity

$$\frac{1}{1-\alpha}\left(\alpha - \mathbf{Q}_0(G)\right)_+ + \mathbf{Q}_1(G)$$

is an objective criterion for comparing two set estimates. It is minimized by the set that is optimal with respect to $\mathbf{Q}_0$ and $\mathbf{Q}_1$, can be estimated from test data, and places a stronger penalty on violations of the mass constraint when α is larger.

6.2 Convex upper bounds

Much attention has been paid lately to the minimization of differentiable convex upper bounds on the probability of misclassification. We briefly remark on a property that may be of interest when attempting to apply this principle to NP classification. In particular, supposed that the non-differentiable convex function $(\cdot)_+$ is replaced by the differentiable convex function

$$\psi_\theta(x) = \frac{1}{\theta}\ln(1+\exp(\theta x)), \quad \theta > 0,$$

which majorizes $(\cdot)_+$ and tends to it in the limit as $\theta \to \infty$. Consider the modified net NP error criterion

$$\mathcal{N}_\theta^\kappa(f) = \kappa\psi_\theta(R_0(f)-\alpha) + R_1(f) - \beta_\alpha.$$

Is this error measure useful for NP classification? Is it minimized by f_α^* for any κ? The answer is a qualified yes.

Proposition 4. *If $\kappa = 2\lambda_\alpha$, then $\mathcal{N}_\theta^\kappa(f_\alpha^*) \le \mathcal{N}_\theta^\kappa(f)$ for all classifiers f.*

The problem of course is that λ_α is unknown in practice. Whereas before we only needed an upper bound on λ_α, minimizing the modified error appears to require exact knowledge of λ_α. As a side note, the factor of 2 occurs because it is equal to $1/\psi_\theta'(0)$. The derivative could easily be changed and the factor would change accordingly, but knowledge of λ_α would still be required.

6.3 Scaling the penalties

In practice it may be desirable to scale the penalties ϕ_0 and ϕ_1 by arbitrary constants $\gamma_0, \gamma_1 > 0$. Consider NP-CON for example and the net error $\mathcal{M}^\kappa$. Theorem 1 holds when $\gamma_1 \ge 1$ and $\gamma_0 = \gamma \ge 2(\kappa + \lambda_\alpha)$ If instead we allow γ_0 and γ_1 to be arbitrary, then Theorem 1 still holds but with an additional constant

$$C = \left(\max\left\{\frac{2(\kappa+\lambda_\alpha)}{\gamma_0}, \frac{1}{\gamma_1}, 1\right\}\right)^2$$

out in front. This fact follows easily by applying the elementary inequality

$$a_0A_0 + a_1A_1 + A_2 \le \max\left\{\frac{1}{a_0}, \frac{1}{a_1}, 1\right\}(A_0 + A_1 + A_2)$$

twice at appropriate places in the proof.

Similar remarks hold for the other oracle bounds. In conclusion, scaling the penalties (which can lead to improved performance on real world data) only affects the oracle inequalities by constants and therefore does not change consistency properties or rates of convergence derived from them.

Acknowledgment

The author is grateful to Robert Nowak for his helpful feedback.

A Proofs

Recall that conditions **A1** and **A2** are assumed throughout.

A.1 Proof of Lemma 1

Set $G_{0,\alpha} = \{x : f^*_\alpha(x) = 0\}$ and $G_{1,\alpha} = \overline{G_{0,\alpha}} = \{x : f^*_\alpha(x) = 1\}$. Define $G_{0,\alpha'}$ and $G_{1,\alpha'}$ similarly. Recall that $h_1(x) \leq \lambda_\alpha h_0(x)$ on $G_{0,\alpha}$. We have

$$\begin{aligned}
\beta_\alpha - \beta_{\alpha'} &= \int_{G_{0,\alpha}} h_1(x)dx - \int_{G_{0,\alpha'}} h_1(x)dx \\
&= \int_{G_{0,\alpha}\setminus G_{0,\alpha'}} h_1(x)dx \\
&\leq \lambda_\alpha \int_{G_{0,\alpha}\setminus G_{0,\alpha'}} h_0(x)dx \\
&= \lambda_\alpha \int_{G_{1,\alpha'}\setminus G_{1,\alpha}} h_0(x)dx \\
&= \lambda_\alpha \left(\int_{G_{1,\alpha'}} h_0(x)dx - \int_{G_{1,\alpha}} h_0(x)dx \right) \\
&= \lambda_\alpha(\alpha' - \alpha).
\end{aligned}$$

This proves the first part of the lemma. The second part follows a similar argument.

A.2 Proofs of Oracle Inequalities

Given $\delta_0, \delta_1 > 0$, define

$$\Omega_0 = \left\{ S : \sup_{f\in\mathcal{F}} \left(\left| R_0(f) - \widehat{R}_0(f) \right| - \phi_0(f, S, \delta_0) \right) > 0 \right\}$$

and

$$\Omega_1 = \left\{ S : \sup_{f\in\mathcal{F}} \left(\left| R_1(f) - \widehat{R}_1(f) \right| - \phi_1(f, S, \delta_1) \right) > 0 \right\}.$$

The basic strategy for proving the oracle inequalities is to show that they hold when $S \in \overline{\Omega_0} \cap \overline{\Omega_1}$, which, by Definition 1, occurs with probability at least $1 - (\delta_0 + \delta_1)$. But first, a couple of lemmas.

Lemma 3. *If $S \in \overline{\Omega_0}$ and $f \in \widehat{\mathcal{F}}_\alpha^\nu$, then $R_0(f) \le \alpha + (1+\nu)\phi_0(f, S, \delta_0)$.*

Proof. This follows from $\widehat{R}_0(f) \le \alpha + \nu\phi_0(f, S, \delta_0)$ and the definition of Ω_0. □

Lemma 4. *Let $\nu \ge -1$. If $S \in \overline{\Omega_0}$ and $f \in \widehat{\mathcal{F}}_\alpha^\nu$, then $\beta_\alpha - R_1(f) \le \lambda_\alpha(1+\nu)\phi_0(f, S, \delta_0)$.*

Proof. Let $S \in \overline{\Omega_0}$ and $f \in \widehat{\mathcal{F}}_\alpha^\nu$. Observe

$$R_0(f) \le \widehat{R}_0(f) + \phi_0(f, S, \delta_0) \le \alpha + (1+\nu)\phi_0(f, S, \delta_0) =: \alpha',$$

which implies $R_1(f) \ge \beta_{\alpha'}$. By Lemma 1, we have

$$\beta_\alpha - R_1(f) \le \beta_\alpha - \beta_{\alpha'} \le \lambda_\alpha(\alpha' - \alpha) = \lambda_\alpha(1+\nu)\phi_0(f, S, \delta_0).$$

□

A.3 Proof of Theorem 3

Assume $S \in \overline{\Omega_0} \cap \overline{\Omega_1}$. We consider three separate cases: (1) $R_1(\widehat{f}_\alpha^c) < \beta_\alpha$ and $R_0(\widehat{f}_\alpha^c) > \alpha$, (2) $R_1(\widehat{f}_\alpha^c) \ge \beta_\alpha$ and $R_0(\widehat{f}_\alpha^c) > \alpha$, and (3) $R_1(\widehat{f}_\alpha^c) \ge \beta_\alpha$ and $R_0(\widehat{f}_\alpha^c) \le \alpha$. Note that the case in which both $R_0(\widehat{f}_\alpha^c) \le \alpha$ and $R_1(\widehat{f}_\alpha^c) < \beta_\alpha$ is impossible.

In the first case, we have

$$\begin{aligned}
\mathcal{M}^\kappa(\widehat{f}_\alpha^c) &= \kappa(R_0(\widehat{f}_\alpha^c) - \alpha) \\
&\le \kappa(1+\nu)\phi_0(\widehat{f}_\alpha^c, S, \delta_0) + \beta_\alpha - R_1(\widehat{f}_\alpha^c) + R_1(\widehat{f}_\alpha^c) - \beta_\alpha \\
&\le (\kappa + \lambda_\alpha)(1+\nu)\phi_0(\widehat{f}_\alpha^c, S, \delta_0) + R_1(\widehat{f}_\alpha^c) - \beta_\alpha \\
&\le \gamma\phi_0(\widehat{f}_\alpha^c, S, \delta_0) + \widehat{R}_1(\widehat{f}_\alpha^c) - \beta_\alpha + \phi_1(\widehat{f}_\alpha^c, S, \delta_1) \\
&= \inf_{f \in \widehat{\mathcal{F}}_\alpha^\nu} \left\{ \gamma\phi_0(f, S, \delta_0) + \widehat{R}_1(f) - \beta_\alpha + \phi_1(f, S, \delta_1) \right\} \\
&\le \inf_{f \in \widehat{\mathcal{F}}_\alpha^\nu} \left\{ \gamma\phi_0(f, S, \delta_0) + R_1(f) - \beta_\alpha + 2\phi_1(f, S, \delta_1) \right\} \\
&\le \inf_{f \in \mathcal{F}_\alpha^\nu} \left\{ \gamma\phi_0(f, S, \delta_0) + R_1(f) - \beta_\alpha + 2\phi_1(f, S, \delta_1) \right\}.
\end{aligned}$$

The first two inequalities follow from Lemmas 3 and 4. The next two inequalities follow from $\gamma \ge (\kappa + \lambda_\alpha)(1+\nu)$ and from $S \in \overline{\Omega_1}$ (used twice). The final inequality follows from $S \in \overline{\Omega_0}$, which implies $\mathcal{F}_\alpha^\nu \subset \widehat{\mathcal{F}}_\alpha^\nu$.

For the second case we have

$$\begin{aligned}
\mathcal{M}^\kappa(\widehat{f}_\alpha^c) &= \kappa(R_0(\widehat{f}_\alpha^c) - \alpha) + R_1(\widehat{f}_\alpha^c) - \beta_\alpha \\
&\le \kappa(1+\nu)\phi_0(\widehat{f}_\alpha^c, S, \delta_0) + R_1(\widehat{f}_\alpha^c) - \beta_\alpha \\
&\le \gamma\phi_0(\widehat{f}_\alpha^c, S, \delta_0) + R_1(\widehat{f}_\alpha^c) - \beta_\alpha.
\end{aligned}$$

Now proceed as in the first case.

For the third case, note

$$\begin{aligned}
\mathcal{M}^\kappa(\widehat{f}_\alpha^c) &= R_1(\widehat{f}_\alpha^c) - \beta_\alpha \\
&\le \kappa(1+\nu)\phi_0(\widehat{f}_\alpha^c, S, \delta_0) + R_1(\widehat{f}_\alpha^c) - \beta_\alpha
\end{aligned}$$

and proceed as in the second case.

To prove the theorem for the net error $\mathcal{N}^\kappa$, the strategy above can be used but with a simplification. The first case above need not be considered, which allows γ to be smaller.

A.4 Proof of Theorem 2

Below we use the elementary fact $(x+y)_+ \le (x)_+ + (y)_+$ for any x, y, which follows by convexity of $(\cdot)_+$. Assume $S \in \overline{\Omega_0} \cap \overline{\Omega_1}$. Then

$$\begin{aligned}
\mathcal{N}^\kappa(\widehat{f}^u_\alpha) &= \kappa\left(R_0(\widehat{f}^u_\alpha) - \alpha\right)_+ + R_1(\widehat{f}^u_\alpha) - \beta_\alpha \\
&\le \kappa\left(\widehat{R}_0(\widehat{f}^u_\alpha) - \alpha + \phi_0(\widehat{f}^u_\alpha, S, \delta_0)\right)_+ + \widehat{R}_1(\widehat{f}^u_\alpha) - \beta_\alpha + \phi_1(\widehat{f}^u_\alpha, S, \delta_1) \\
&\le \kappa\left(\widehat{R}_0(\widehat{f}^u_\alpha) - \alpha - \nu\phi_0(\widehat{f}^u_\alpha, S, \delta_0)\right)_+ + \kappa(1+\nu)\phi_0(\widehat{f}^u_\alpha, S, \delta_0) + \widehat{R}_1(\widehat{f}^u_\alpha) - \beta_\alpha + \phi_1(\widehat{f}^u_\alpha, S, \delta_1) \\
&= \inf_{f\in\mathcal{F}}\left\{\kappa\left(\widehat{R}_0(f) - \alpha - \nu\phi_0(f, S, \delta_0)\right)_+ + \kappa(1+\nu)\phi_0(f, S, \delta_0) + \widehat{R}_1(f) - \beta_\alpha + \phi_1(f, S, \delta_1)\right\} \\
&\le \inf_{f\in\mathcal{F}}\left\{\kappa\left(R_0(f) - \alpha\right)_+ + 2\kappa\phi_0(f, S, \delta_0) + R_1(f) - \beta_\alpha + 2\phi_1(f, S, \delta_1)\right\}.
\end{aligned}$$

A.5 Proof of Proposition 3

The MSE of the combined estimator $\widehat{\theta} = \widehat{\theta_0} + \widehat{\theta_1}$ is

$$\begin{aligned}
\mathbf{E}[(\widehat{\theta} - \theta)^2] &= \mathbf{E}[(\widehat{\theta_0} - \theta_0)^2] + \mathbf{E}[(\widehat{\theta_1} - \theta_1)^2] + \mathbf{E}[(\widehat{\theta_0} - \theta_0)(\widehat{\theta_1} - \theta_1)] \\
&= \mathbf{E}[(\widehat{\theta_0} - \theta_0)^2] + \mathbf{E}[(\widehat{\theta_1} - \theta_1)^2]
\end{aligned}$$

where in the last step we use independence and unbiasedness of $\widehat{\theta_1}$. Thus, it suffices to bound the MSE's of $\widehat{\theta_0}$ and $\widehat{\theta_1}$ separately. It is well known that $MSE(\widehat{\theta_1}) = p_1(1-p_1)/n_1$.

Writing out the definition of MSE we have

$$\begin{aligned}
\mathbf{E}[(\widehat{\theta_0} - \theta_0)^2] &= \kappa^2 \sum_{k=0}^{n_0}\left(\left(\frac{k}{n_0} - \alpha\right)_+ - (p_0 - \alpha)_+\right)^2 \mathrm{bin}(n_0, p_0, k) \\
&\le \kappa^2 \sum_{k=0}^{n_0}\left(\left(\frac{k}{n_0} - p_0\right)_+\right)^2 \mathrm{bin}(n_0, p_0, k) \\
&\le \kappa^2 \sum_{k=0}^{n_0}\left(\frac{k}{n_0} - p_0\right)^2 \mathrm{bin}(n_0, p_0, k) \\
&= \kappa^2 \frac{p_0(1-p_0)}{n_0},
\end{aligned}$$

where in the second step we use the fact $(a)_+ - (b)_+ \le (a-b)_+$. This concludes the proof.

A.6 Proof of Proposition 4

It suffices to consider only the classifiers $f^*_{\alpha'}, \alpha' \in [0,1]$, since each of these has optimal false negative rate for false positive rate α'. There are two cases. If $\alpha' < \alpha$, then

$$\begin{aligned}
\mathcal{N}^\kappa_\theta(f^*_\alpha) - \mathcal{N}^\kappa_\theta(f^*_{\alpha'}) &= \kappa(\psi_\theta(0) - \psi_\theta(\alpha' - \alpha)) + (\beta_\alpha - \beta_{\alpha'}) \\
&\le \kappa\psi'_\theta(0)(\alpha - \alpha') + \lambda_\alpha(\alpha' - \alpha) \\
&= (\lambda_\alpha - \tfrac{1}{2}\kappa)(\alpha' - \alpha) \\
&= 0,
\end{aligned}$$

where the inequality follows from the fact that ψ is convex and increasing, and from Lemma 1.

Similarly, if $\alpha' > \alpha$, then

$$\begin{aligned}
\mathcal{N}^{\kappa}(f^*_{\alpha}) - \mathcal{N}^{\kappa}(f^*_{\alpha'}) &= \kappa(\psi_\theta(0) - \psi_\theta(\alpha' - \alpha)) + (\beta_\alpha - \beta_{\alpha'}) \\
&\leq \kappa\psi'_\theta(0)(\alpha - \alpha') + \lambda_\alpha(\alpha' - \alpha) \\
&= (\lambda_\alpha - \tfrac{1}{2}\kappa)(\alpha' - \alpha) \\
&= 0.
\end{aligned}$$

References

[1] A. Cannon, J. Howse, D. Hush, and C. Scovel, "Learning with the Neyman-Pearson and min-max criteria," Los Alamos National Laboratory, Tech. Rep. LA-UR 02-2951, 2002. [Online]. Available: http://www.c3.lanl.gov/~kelly/ml/pubs/2002_minmax/paper.pdf

[2] C. Scott and R. Nowak, "A Neyman-Pearson approach to statistical learning," *IEEE Trans. Inform. Theory*, vol. 51, no. 8, pp. 3806–3819, 2005.

[3] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition.* New York: Springer, 1996.

[4] V. Vapnik, *Statistical Learning Theory.* New York: Wiley, 1998.

[5] E. Lehmann, *Testing statistical hypotheses.* New York: Wiley, 1986.

[6] H. V. Trees, *Detection, Estimation, and Modulation Theory: Part I.* New York: John Wiley & Sons, 2001.

[7] C. Scott and R. Nowak, "Learning minimum volume sets," UW-Madison, Tech. Rep. ECE-05-2, 2005. [Online]. Available: http://www.stat.rice.edu/~cscott

[8] V. Vapnik and C. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and Its Applications*, vol. 16, no. 2, pp. 264–280, 1971.

[9] H. Chernoff, "A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations," *Annals of Mathematical Statistics*, vol. 23, pp. 493–507, 1952.

[10] J. Langford, "Tutorial on practical prediction theory for classification," *J. Machine Learning Research*, vol. 6, pp. 273–306, 2005.

[11] C. Scott and R. Nowak, "Learning minimum volume sets," *to appear at* Neural Information Processing Systems 19 – *NIPS '05*, 2005.

[12] W. Polonik, "Minimum volume sets and generalized quantile processes," *Stochastic Processes and their Applications*, vol. 69, pp. 1–24, 1997.