How many faces can be recognized? Performance extrapolation for multi-class classification

Charles Y. Zheng
Department of Statistics
Stanford University
Stanford, CA 94305
snarles@stanford.edu

Rakesh Achanta
Department of Statistics
Stanford University
Stanford, CA 94305
rakesha@stanford.edu

Yuval Benjamini
Department of Statistics
Hebrew University
Jerusalem, Israel
yuval.benjamini@mail.huji.ac.il

Abstract

The difficulty of multi-class classification generally increases with the number of classes. Using data from a subset of the classes, can we predict how well a classifier will scale with an increased number of classes? Under the assumption that the classes are sampled exchangeably, and under the assumption that the classifier is generative (e.g. QDA or Naive Bayes), we show that the expected accuracy when the classifier is trained on k classes is the $k-1$st moment of a *conditional accuracy distribution*, which can be estimated from data. This provides the theoretical foundation for developing estimation approaches based on pseudolikelihood, unbiased estimation, and high-dimensional asymptotics. We find empirically that these methods work well for generative classifiers and non-generative classifiers alike, suggesting a possible extension of our theory to a larger class of *asymptotically generative* classifiers.

1 Introduction

In multi-class classification, one observes pairs (z, y) where $y \in \mathcal{Y} \subset \mathbb{R}^p$ are feature vectors, and z are unknown labels, which lie in a countable label set $\mathcal{Z}$. The goal is to construct a classification rule for predicting the label of a new data point; generally, the classification rule $h : \mathcal{Y} \to \mathcal{Z}$ is learned from previously observed data points. In many applications of multi-class classification, such as face recognition or image recognition, the space of potential labels is practically infinite. However, one considers the classification problem on a finite subset of the labels $\mathcal{Z}_1 \subset \mathcal{Z}$: for instance, classifying the faces of 100 selected individuals from the population. At a later time, one might consider a larger (but still finite) classification problem on $\mathcal{Z}_2 \subset \mathcal{Z}$ with $\mathcal{Z}_2 \supset \mathcal{Z}_1$. In general, consider an infinite sequence of classification problems on subsets $\mathcal{Z}_1 \subset \cdots \subset \mathcal{Z}_t \subset \cdots$. Let S_i represent the training data available for the ith classification problem, and let $h^{(i)} : \mathcal{Y} \to \mathcal{Z}_i$ be the learned classification rule. Define the accuracy for the ith problem as

$$\text{acc}^{(i)} = \Pr[h^{(i)}(Y) = Z | Z \in \mathcal{Z}_i].$$

where the probability is taken over the joint distribution of (Z, Y). Using data from only $\mathcal{Z}_k$, can one predict the accuracy achieved on the larger label set $\mathcal{Z}_K$, with $K > k$? This is the problem of *prediction extrapolation*.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

A practical instance of prediction extrapolation occurs in neuroimaging studies, Kay et al. (2008) obtain fMRI brain scans which record how a single subject's visual cortex responds to natural images. The label set $\mathcal{Z}$ corresponds to the space of all grayscale photographs of natural images, and the set $\mathcal{Z}_1$ is a subset of 1750 photographs used in the experiment. Kay et al. construct a classifier based on a combination of regularized multiple-response regression and Naive Bayes: they achieve over 0.75 accuracy on the subset of 1750 photographs, which by itself is already a convincing demonstration of the richness of the information contained in the fMRI scan. However, it would also be of interest to know what accuracy could be achieved on a larger set of photographs. Kay et al. calculated (based on exponential extrapolation) that it would take on the order of $10^{9.5}$ photographs before the accuracy of the model drops below 0.10! Directly validating this estimate would take immense resources, so it would be useful to develop the theory needed to understand how to compute such extrapolations in a principled way.

However, in the fully general setting, it is impossible on construct non-trivial bounds on the accuracy achieved on the new classes $\mathcal{Z}_K \setminus \mathcal{Z}_k$ based only on knowledge of $\mathcal{Z}_k$: after all, $\mathcal{Z}_k$ could consist entirely of well-separated classes while the new classes $\mathcal{Z}_K \setminus \mathcal{Z}_k$ consist entirely of highly inseparable classes, or vice-versa. Thus, the most important assumption for our theory is that of *exchangeable sampling*. The labels in $\mathcal{Z}_i$ are assumed to be an exchangeable sample from $\mathcal{Z}$. The exchangeability further implies that the marginal distributions of $z \in \mathcal{Z}$ are equiprobable within every subset $\mathcal{Z}_i$. The condition of exchangeability ensures that the separability of random subsets of $\mathcal{Z}$ can be inferred by looking at the empirical distributions in $\mathcal{Z}_k$, and therefore that some estimate of the achievable accuracy on $\mathcal{Z}_K$ can be obtained.

Unfortunately, the assumption of exchangeability is clearly violated in a majority of instances of multi-class classification. Many multi-class classification problems have a hierarchical structure, where the initial label set $\mathcal{Z}_1$ corresponds to a coarse-grained partition of the instances, and an expanded label set $\mathcal{Z}_2$ corresponds to a refinement of the partition induced by $\mathcal{Z}_1$: for instance, $\mathcal{Z}_1$ consists of the categories $\{\text{animal}, \text{vegetable}, \text{mineral}\}$, while $\mathcal{Z}_2$ consists of subcategories $\{\text{mammal}, \text{bird}, \text{insect}, \text{reptile}, \text{fungus}, \text{tree}, \text{flower}, \text{rock}, \text{metal}\}$. Not only is $\mathcal{Z}_2$ not a superset of $\mathcal{Z}_1$, but the marginal distributions within $\mathcal{Z}_2$ are necessarily more concentrated than the marginals of $\mathcal{Z}_1$. Many non-hierarchical classification problems are also excluded by the requirement of exchangeability. Consider the problem of annotating spoken words: the set $\mathcal{Z}_1$ might consist of data from the 100 most common words, while the set $\mathcal{Z}_2$ consists of data from the 1000 most common words. Exchangeability is violated because the words $z \in \mathcal{Z}$ are not equiprobable, but rather follow a long-tail law. It would be interesting to extend our theory to the hierarchical setting, or to handle non-hierarchical settings with non-uniform prior class probabilities, but we leave the subject for future work.

In addition to the assumption of exchangeability, we restrict the set of classifiers considered. We focus on *generative classifiers*, which are classifiers which work by training a model separately on each class. This convenient property allows us to characterize the accuracy of the classifier by selectively conditioning on one class at a time: in section 3, we use this technique to reveal an equivalence between the expected accuracies of $\mathcal{Z}_k$ to moments of a common distribution. This moment equivalence result allows standard approaches in statistics, such as U-statistics and nonparametric pseudolikelilood, to be directly applied to the extrapolation problem, as we discuss in section 4. In non-generative classifiers, the classification rule has a joint dependence on the entire set of classes, and cannot be analyzed by conditioning on individual classes. Nevertheless, in Section 5, we see that our methods achieve similarly accurate extrapolation for both generative and non-generative classifiers in real data examples. Section 6 concludes.

2 Setting

2.1 Prediction extrapolation

Having motivated the problem of prediction extrapolation, we now reformulate the problem for notational and theoretical convenience. Instead of requiring $\mathcal{Z}_k$ to be a random subset of $\mathcal{Z}$ as we did in section 1, take $\mathcal{Z} = \mathbb{N}$ and $\mathcal{Z}_k = \{1, \ldots, k\}$. We fix the size of $\mathcal{Z}_k$ without losing generality, since any monotonic sequence of finite subsets can be embedded in a sequence with $|\mathcal{Z}_k| = k$. In addition, rather than randomizing the labels, we will randomize the marginal distribution of each label; Towards that end, let $\mathcal{Y} \subset \mathbb{R}^p$ be a space of feature vectors, and let $\mathcal{P}(\mathcal{Y})$ be a measurable

space of probability distributions on $\mathcal{Y}$. Let $\mathcal{F}$ be a probability measure on $\mathcal{P}$, and let $F_1, F_2, \ldots$ be an infinite sequence of i.i.d. draws from $\mathbb{F}$. We refer to $\mathbb{F}$, a probability measure on probability measures, as a *meta-distribution*. The distributions $F_1, \ldots, F_k$ are the marginal distributions of the first k classes. We therefore rewrite the accuracy as

$$\text{acc}^{(i)} = \frac{1}{t} \sum_{i=1}^{t} \Pr_{F_i}[h^{(t)}(Y) = i].$$

where the probabilities are taken over $Y \sim F_i$.

In order to construct the classification rule $h^{(t)}$, we need data from the classes $F_1, \ldots, F_t$. In most instances of multi-class classification, one observes independent observations from each F_i which are used to construct the classifier. Since the order of the observations does not generally matter, a sufficient statistic for the training data for the tth classification problem is the collection of empirical distributions $\hat{F}_1^{(t)}, \ldots, \hat{F}_t^{(t)}$ for each class. Henceforth, we make the simplifying assumption that the training data for the ith class remains fixed from $t = i, i+1, \ldots$, so we drop the superscript on $\hat{F}_i^{(t)}$. Write $\hat{\mathbb{F}}(F)$ for the conditional distribution of $\hat{F}_i$ given $F_i = F$; also write $\hat{\mathbb{F}}$ for the marginal distribution of $\hat{F}$ when $F \sim \mathbb{F}$. As an example, suppose every class has the number of training examples $r \in \mathbb{N}$; then $\hat{F}$ is the empirical distribution of r i.i.d. observations from F, and $\hat{\mathbb{F}}(F)$ is the *empirical meta-distribution* of $\hat{F}$. Meanwhile, $\hat{\mathbb{F}}$ is the meta-distribution of the empirical distribution of r i.i.d. draws from a random $F \sim \mathbb{F}$.

2.2 Multi-class classification

Combining the formalism of Tewari and Bartlett (2007), we define a classifier as a collection of mappings $\mathcal{M}_i : \mathcal{P}(\mathcal{Y})^k \times \mathcal{Y} \to \mathbb{R}$ called *margin functions.* Intuitively speaking, each margin function *learns a model* from the first k arguments, which are the empirical marginals of the k classes, which it uses to assign a *margin* or *score* to the *query point* $y \in \mathcal{Y}$. A higher score $\mathcal{M}_i(\hat{F}_1, \ldots, \hat{F}_k, y)$ indicates a higher estimated probability that y belongs to the kth class. Therefore, the classification rule corresponding to a classifier $\mathcal{M}_i$ assigns a class with maximum margin to y:

$$h(y) = \text{argmax}_{i \in \{1,\ldots,k\}} \mathcal{M}_i(y).$$

For our purposes, it is not important how ties are resolved. We will also neglect discussion of randomized classifiers, except to mention that they can be treated in our framework as probability distributions over deterministic classifiers; we also neglect the incorporation of prior class probabilities into the classifier, since in our setting the prior class probabilities are uniform.

One reason why we formalize a classifier in terms of *empirical distributions* rather than data points is to formalize the notion of *continuity*. The level sets $\{y : \text{argmax}_i \mathcal{M}_i(\hat{F}_1, \ldots, \hat{F}_k, y) = k\}$ are called *decision regions.* We say the classifier is *continuous* if and only if the *decision regions* of $\mathcal{M}_{i=1}^k$ are continuous in the first k arguments with respect to the topology of weak convergence.

To acquaint the reader with our formalism, we redefine a number of familiar classifiers according to our notation.

Example 1. (OVO) Let $\mathcal{B}$ (the *base classifier*) be a binary-valued mapping with three arguments: distributions $\hat{F}_0$, $\hat{F}_1$, and query y. A one-vs-one (OVO) classifier is defined by

$$\mathcal{M}_i(\hat{F}_1, \ldots, \hat{F}_k, y) = \sum_{j \neq i} \mathcal{B}(\hat{F}_j, \hat{F}_i, y).$$

Example 2. (OVA) Let $\mathcal{D}$ (*the base classifier*) be a real-valued mapping with three arguments: distributions F_0, F_1, and query y. A one-vs-all (OVA) classifier is defined by

$$\mathcal{M}_i(\hat{F}_1, \ldots, \hat{F}_k, y) = \mathcal{D}(\hat{F}_i, \frac{1}{k} \sum_{j=1}^{k} \hat{F}_j, y).$$

It is immediate that an OVO classifier is continuous if and only if the level sets of $\mathcal{B}$ are continuous with respect to the topology of weak convergence, and similarly that an OVA classifier is continuous if and only if $\mathcal{D}$ is continuous.

Example 3. (ϵ-NN) ϵ-nearest neighbors can be thought of as k-nearest neighbors with $k = \epsilon n$ for fixed ϵ. Let d be a distance metric on $\mathcal{Y}$. Let $D(y)$ denote the induced distribution of $d(Y, y)$ when $Y \sim \frac{1}{k}\sum_{i=1}^{k} \hat{F}_i$, and let $d_\epsilon(y)$ denote the ϵ-quantile of $D(y)$. An ϵ-nearest neighbor classifier is defined by

$$\mathcal{M}_i(\hat{F}_1, \ldots, \hat{F}_k, y) = \Pr_{Y \sim \hat{F}_i}[d(Y, y) < d_\epsilon(y)].$$

Note that the ϵ-NN classifier is also an example of a OVA classifier, with $\mathcal{D}(F_0, F_1, y) = F_1(B_{d_\epsilon(y)})$, where B_r is the d-ball of radius r, since one can define $d_{\epsilon,t}(y) = \sup_r$ s.t. $\frac{k-1}{k}F_0 + \frac{1}{k}F_1)(B_r(y)) \leq \epsilon$.

Example 4. (Multinomial logistic regression.) Assume WLOG that $y_1 = 1$ for all $y \in \mathcal{Y}$, and let B be a $p \times k$ matrix which minimizes the objective function

$$-\mathbf{E}_{\hat{F}_j}\left[\langle Y, B_j\rangle - \log\left[\sum_{\ell=1}^{k} \exp[\langle Y, B_\ell\rangle]\right]\right].$$

A multinomial logistic regression classifier is defined by

$$\mathcal{M}_i(\hat{F}_1, \ldots, \hat{F}_k, y) = \langle y, B_i\rangle.$$

By the convexity of the objective function, multinomial logistic regression is continuous wherever the minimizer B is unique.

Throughout these examples we have neglected to discuss model selection or parameter tuning. It is admittedly non-trivial to formalize procedures such as cross-validation using our formalism, and since model selection lies beyond the scope of our paper, we omit the definitions.

2.3 Generative classifiers

For some classifiers, the margin function $\mathcal{M}_i$ is especially simple in that $\mathcal{M}_i$ is only a function of $\hat{F}_i$ and y. Furthermore, due to symmetry, in such cases one can write

$$\mathcal{M}_i(\hat{F}_1, \ldots, \hat{F}_k, y) = \mathcal{Q}(\hat{F}_i, y)$$

where $\mathcal{Q}$ is called a *single-class margin* (or simply *margin*.) For notational convenience, we assume that ties occur with probability zero: that is, $\hat{\mathbb{F}}$ and $\mathcal{Q}$ jointly satisfy the *tie-breaking* property:

$$\Pr[\mathcal{Q}(\hat{F}, y) = \mathcal{Q}(\hat{F}', y)] = 0. \tag{1}$$

for all $y \in \mathcal{Y}$, where $\mathbb{F}, \mathbb{F}' \overset{iid}{\sim} \hat{\mathbb{F}}$. Quadratic discriminant analysis and Naive Bayes are two examples of generative classifiers. For QDA, the margin is given by

$$\mathcal{Q}_{QDA}(\hat{F}, y) = -(y - \mu(\hat{F}))^T \Sigma(\hat{F})^{-1}(y - \mu(\hat{F})) - \log\det(\Sigma(\hat{F}))$$

where $\mu(F) = \int y dF(y)$ and $\Sigma(F) = \int (y - \mu(F))(y - \mu(F))^T dF(y)$. In Naive Bayes, the margin is

$$\mathcal{Q}_{NB}(\hat{F}, y) = \sum_{i=1}^{n} \log \hat{f}_i(y_i)$$

where $\hat{f}_i$ is a density estimate for the ith component of $\hat{F}$.

The *generative* property allows us to prove strong results about the accuracy of the classifier under the exchangeable sampling assumption. For starters, we can show that the expected accuracy follows a *mixed exponential decay*, as stated by the following theorem.

Theorem 2.1 *Let $\mathcal{Q}$ be the scoring function of a generative classifier, and assume that $F_1, \ldots, F_k \overset{iid}{\sim} \mathbb{F}$ and $\hat{F}_i \sim \hat{\mathbb{F}}(F_i)$ independently, following the notation of section 2. Further assume that $\mathcal{Q}$ and $\mathbb{F}$ satisfy the tie-breaking property* (1)*. Then, recalling the definition of accuracy,*

$$acc^{(t)} = \frac{1}{t}\sum_{i=1}^{t} \Pr_{Y \sim F_i}[\mathcal{Q}(\hat{F}_i, Y) > argmax_{i>j}\mathcal{Q}(\hat{F}_i, Y)],$$

there exists a measure α on $[0, \infty)$ such that

$$\boldsymbol{E}[acc^{(t)}] = \int_{\mathbb{R}^+} e^{-\kappa t} d\alpha(\kappa).$$

(We give the proof in section 3.)

The theorem immediately suggests a method for predicting $\text{acc}^{(K)}$: fit a mixed exponential decay to $a(t) = \text{acc}^{(t)}$ for $t = 2, \ldots, k$ and extrapolate the curve to $t = K$. This is just one of many methods which can be developed for generative classifiers, which we discuss more fully in Section 3 and 4.

3 Prediction extrapolation for generative classifiers

Let us specialize to the case of a generative classifier, with scoring rule $\mathcal{Q}$. Consider estimating the expected accuracy at time t,

$$p_t \stackrel{def}{=} \mathbf{E}[\text{acc}^{(t)}].$$

Define the *conditional accuracy* function $u(\hat{F}, y)$ which maps a distribution $\hat{F}$ on $\mathcal{Y}$ and a *test* observation y to a real number in $[0, 1]$. The conditional accuracy gives the probability that for independently drawn $\hat{F}'$ from $\hat{\mathbb{F}}$, that $\mathcal{Q}(\hat{F}, y)$ will be greater than $\mathcal{Q}(\hat{F}', y)$:

$$u(\hat{F}, y) = \Pr_{\hat{F} \sim \hat{\mathbb{F}}}[\mathcal{Q}(\hat{F}, y) > \mathcal{Q}(\hat{F}', y)].$$

Define the *conditional accuracy* distribution ν as the law of $u(\hat{F}, Y)$ where $\hat{F}$ and Y are generated as follows: (i) a true distribution F is drawn from $\mathbb{F}$; (ii) the query Y is drawn from F, and (iii) the empirical distribution $\hat{F}$ is drawn from $\hat{\mathbb{F}}(F)$ (e.g., the distribution of the empirical distribution of r i.i.d. observations drawn from F), with Y independent of $\hat{F}$. The significance of the conditional accuracy distribution is that the expected generalization error p_t can be written in terms of its moments.

Theorem 3.1. *Let $\mathcal{Q}$ be a single-distribution margin, and let $\mathbb{F}$, $\hat{F}(F)$ be a distribution on $\mathcal{P}(\mathcal{Y})$. Let U be defined as the random variable*

$$U = u(\hat{F}, Y)$$

for $F \sim \mathbb{F}$, $Y \sim F$, and $\hat{F} \sim \hat{\mathbb{F}}(F)$ with $Y \perp \hat{F}$. Recall the definition

$$p_k = \boldsymbol{E}[acc^{(k)}] = \boldsymbol{E}\left[\frac{1}{t}\sum_{i=1}^{k} \Pr_{Y \sim F_i}[\mathcal{Q}(\hat{F}_i, Y) > \max_{j \neq i} \mathcal{Q}(\hat{F}_j, Y)]\right].$$

Then

$$p_k = \boldsymbol{E}[U^{k-1}].$$

Proof. Write $q^{(i)}(y) = \mathcal{Q}(\hat{F}_i, y)$. By using conditioning and conditional independence, p_k can be written

$$\begin{aligned}
p_k &= \mathbf{E}\left[\frac{1}{k}\sum_{i=1}^{k} \Pr_{F_i}[q^{(i)}(Y) > \max_{j \neq i} q^{(j)}(Y)]\right] \\
&= \mathbf{E}\left[\Pr_{F_1}[q^{(1)}(Y) > \max_{j \neq 1} q^{(j)}(Y)]\right] \\
&= \mathbf{E}_{F_1}[\Pr[q^{(1)}(Y) > \max_{j \neq 1} q^{(j)}(Y) | \hat{F}_1, Y]] \\
&= \mathbf{E}_{F_1}[\Pr[\cap_{j>1} q^{(1)}(Y) > q^{(j)}(Y) | \hat{F}_1, Y]] \\
&= \mathbf{E}_{F_1}[\prod_{j>1} \Pr[q^{(1)}(Y) > q^{(j)}(Y) | \hat{F}_1, Y]] \\
&= \mathbf{E}_{F_1}[\Pr[q^{(1)}(Y) > q^{(2)}(Y) | \hat{F}_1, Y]^{k-1}] \\
&= \mathbf{E}_{F_1}[u(\hat{F}_1, Y)^{k-1}] = \mathbf{E}[U^{k-1}].
\end{aligned}$$

□

Theorem 3.1 tells us that the problem of extrapolation can be approached by attempting to estimate the conditional accuracy distribution. The $(t-1)$th moment of U gives us p_t, which will in turn be a good estimate of $\text{acc}^{(t)}$.

While $U = u(\hat{F}, Y)$ is not directly observed, we can obtain unbiased estimates of $u(\hat{F}_i, y)$ by using test data. For any $\hat{F}_1, \ldots, \hat{F}_k$, and independent test point $Y \sim F_i$, define

$$\hat{u}(\hat{F}_i, Y) = \frac{1}{k-1} \sum_{j \neq i} I(\mathcal{Q}(\hat{F}_i, Y) > \mathcal{Q}(\hat{F}_j, Y)). \tag{2}$$

Then $\hat{u}(\hat{F}_i, Y)$ is an unbiased estimate of $u(\hat{F}_i, Y)$, as stated in the following theorem.

Theorem 3.2. *Assume the conditions of theorem 3.1. Then defining $V = (k-1)\hat{u}(\hat{F}_i, y)$, we have*

$$V \sim \text{Binomial}(k, u(\hat{F}_i, y)).$$

Hence,

$$\boldsymbol{E}[\hat{u}(\hat{F}_i, y)] = u(\hat{F}_i, y).$$

In section 4, we will use this result to estimate the moments of U. Meanwhile, since U is a random variable on $[0, 1]$, we can immediately prove theorem 2.1.

Proof of theorem 2.1. Let α be the law of $-\log(U)$. Then from change-of-variables $\kappa = -\log(u)$, we get

$$\mathbf{E}[\text{acc}^{(t)}] = \mathbf{E}[U^{t-1}] = \int_0^1 u^{t-1} d\nu(u) = \int_0^1 e^{t \log(u)} \frac{1}{u} d\nu(u) = \int_{\mathbb{R}^+} e^{-\kappa t} d\alpha(\kappa).$$

□

3.1 Properties of the conditional accuracy distribution

The conditional error distribution ν is determined by $\mathbb{F}$ and $\mathcal{Q}$. What can we say about the the conditional accuracy distribution without making any assumptions on either $\mathbb{F}$ or $\mathcal{Q}$? The answer is: not much–for an arbitrary probability measure ν' on $[0, 1]$, one can construct $\mathbb{F}$ and $\mathcal{Q}$ such that $\nu = \nu'$, even if one makes the *perfect sampling assumption* that $\hat{F} = F$.

Theorem 3.3. *Let U be defined as in Theorem 3.1, and let ν denote the law of U. Then, for any probability distribution ν' on $[0, 1]$, one can construct a meta-distribution $\mathbb{F}$ and a scoring rule $\mathcal{Q}$ such that $\nu = \nu'$ under perfect sampling (that is, $\hat{F} = F$.)*

Proof. Let G be the cdf of ν, $G(x) = \int_0^x d\nu(x)$, and let $H(u) = \sup_x\{G(x) \leq u\}$. Define $\mathcal{Q}$ by

$$\mathcal{Q}(\hat{F}, y) = \begin{cases} 0 & \text{if } \mu(\hat{F}) > y + H(y) \\ 0 & \text{if } y + H(y) > 1 \text{ and } \mu(\hat{F}) \in [H(y) - y, y] \\ 1 + \mu(\hat{F}) - y & \text{if } \mu(\hat{F}) \in [y, y + H(y)] \\ 1 + y + \mu(\hat{F}) & \text{if } \mu(\hat{F}) + H(y) > 1 \text{ and } \mu(\hat{F}) \in [0, H(y) - y]. \end{cases}$$

Let $\theta \sim \text{Uniform}[0, 1]$, and define $F \sim \mathbb{F}$ by $F = \delta_\theta$, and also $\hat{F} = F$. A straightforward calculation yields that $\nu = \nu'$. □

On the other hand, we can obtain a positive result if we assume that the classifier approximates a *Bayes classifier.* Assuming that F is absolutely continuous with respect to Lebesgue measure Λ with probability one, a Bayes classifier results from assuming perfect sampling ($\hat{F} = F$) and taking $\mathcal{Q}(\hat{F}, y) = \frac{dF}{d\Lambda}(y)$. Theorem 3.4. states that for a Bayes classifier, ν has a density $\eta(u)$ which is monotonically increasing. Since a 'good' classifier approximates the Bayes classifier, we intuitively expect that a monotonically increasing density η is a good model for the conditional accuracy distribution of a 'good' classifier.

Theorem 3.4. *Assume the conditions of theorem 3.1, and further suppose that $\hat{F} = F$, F is absolutely continuous with respect to Λ with probability one, and that $\mathcal{Q}(\hat{F}, y) = \frac{dF}{d\Lambda}(y)$. Let ν denote the law of U. Then ν has a density $\eta(u)$ on $[0,1]$ which is monotonic in u.*

Proof. It suffices to prove that

$$\nu([u, u+\delta]) < \nu([v, v+\delta])$$

for all $0 < u < v < 1$ and $0 < \delta < 1 - v$. Let $\mathcal{P}_{ac}(\mathcal{Y})$ denote the space of distributions supported on $\mathcal{Y}$ which are absolutely continuous with respect to p-dimensional Lebesgue measure Λ. For $F \in \mathcal{P}_{ac}(\mathcal{Y})$, let $f = \frac{dF}{d\Lambda}$. Define the set

$$J_F(A) = \{y \in \mathcal{Y} : u(F, y) \in A\} = \left\{y \in \mathcal{Y} : \Pr_{Y \sim F}[f(y) > f(Y)] \in A\right\}.$$

for all $A \subset [0,1]$. One can verify that for all $F \in \mathcal{P}_{ac}\mathcal{Y}$,

$$F(J_F([u, u+\delta])) \leq F(J_F([v, v+\delta])).$$

Yet, since

$$\Pr[U \in [u, u+\delta]] = \Pr_{F \sim \mathbb{F}}[Y \in J_F([u, u+\delta])] = \mathbf{E}_{F \sim \mathbb{F}}[F(J_F([u, u+\delta]))]$$

$$\Pr[U \in [v, v+\delta]] = \Pr_{F \sim \mathbb{F}}[Y \in J_F([v, v+\delta])] = \mathbf{E}_{F \sim \mathbb{F}}[F(J_F([v, v+\delta]))]$$

we obtain

$$\Pr[U \in [u-\delta, u+\delta]] \leq \Pr[U \in [v-\delta, v+\delta]].$$

Taking $\delta \to 0$, we conclude the theorem. □

4 Nonparametric Estimation

Let us assume that U has a density $\eta(u)$. While $U = u(\hat{F}, 0, Y)$ cannot be directly observed, we can estimate $u(\hat{F}_i, 0, y^{(i), r_1 + j})$ for any $1 \leq i \leq k$, $1 \leq j \leq r_2$ from the data.

At a high level, we have a hierarchical model where U is drawn from a density $\eta(u)$ on $[0,1]$ and then $V_{i,j} \sim \text{Binomial}(k, U)$; therefore the marginal distribution of $V_{i,j}$ can be written

$$\Pr[V_{i,j} = \ell] = \binom{k}{\ell} \int_0^1 u^\ell (1-u)^{k-\ell} \eta(u) du.$$

However, the observed $\{V_{i,j}\}$ do *not* comprise an i.i.d. sample.

We discuss the following three approaches for estimating $p_t = \mathbf{E}[U^{t-1}]$ based on $V_{i,j}$. The first is *unbiased estimation* based on binomial U-statistics, which is discussed in Section 4.1. The second is the *psuedolikelihood* approach. In problems where the marginal distributions are known, but the dependence structure between variables is unknown, the *psuedolikelihood* is defined as the product of the marginal distributions. For certain problems in time series analysis and spatial statistics, the maximum psuedolikelihood estimator (MPLE) is proved to be consistent (CITE). We discuss psuedolikelihood-based approaches in Sections 4.2 and 4.3.

4.1 Unbiased estimation

If $V \sim \text{Binomial}(k, \eta)$, then an unbiased estimator $f_t(V)$ of $\eta^(t-1)$ exists if and only if $0 \leq t \leq k$.

The theory of U-statistics provides the minimal variance unbiased estimator for $\eta^(t-1)$:

$$\eta^t = \mathbf{E}\left[\frac{\binom{V}{t}}{\binom{k}{t}}\right].$$

This result can be immediately applied to yield an unbiased estimator of p_t, when $t \leq k$:

$$\hat{p}_t^{UN} = \mathbf{E}\left[\frac{1}{kr_2}\sum_{i=1}^{k}\sum_{j=1}^{r_2}\frac{\binom{V_{i,j}}{t}}{\binom{k}{t}}\right]. \tag{3}$$

The problem of *extrapolation* concerns the case $t > k$, in which the expression (3) is undefined. Still, the estimator (3) is worthy of study, since it has close to optimal performance for the case $t \leq k$.

4.2 Maximum pseudo-likelihood

The psuedolikelihood is defined as

$$\ell_t(\eta) = \sum_{i=1}^{k}\sum_{j=1}^{r_1}\log\left(\int u^{V_{i,j}}(1-u)^{k-V_{i,j}}\eta(u)du\right), \tag{4}$$

and a maximum psuedolikelihood estimator (MPLE) is defined as any density $\hat{\eta}$ such that

$$\ell(\hat{\eta}_{MPLE}) = \sup_{\eta}\ell_t(\eta).$$

The motivation for $\hat{\eta}_{MPLE}$ is that it consistently estimates η in the limit where $k \to \infty$.

Theorem 4.2. *For given $\mathbb{F}$ and scoring rule $\mathcal{Q}$, assume that U as defined in Theorem 3.1 has a density $\eta(u)$ and that $\mathcal{Q}$ satisfies the tie-breaking property* (1)*, and also that $r_2 \geq 1$. For $t = 1, 2, \ldots,$ let $\hat{\eta}_t$ be any MPLE for ℓ_t. As $k_t \to \infty$, $\hat{\eta}_t$ weakly converges to η.*

However, in finite samples, $\hat{\eta}_{MPLE}$ is not uniquely defined, and if we define the plug-in estimator

$$\hat{p}_t^{MPLE} = \int u^{t-1}\hat{\eta}_{MPLE}(u)du,$$

$\hat{p}_t^{MPLE}$ can vary over a large range, depending on which $\hat{\eta} \in \operatorname{argmax}_\eta \ell_t(\eta)$ is selected. These shortcomings motivate the adoption of additional constraints on the estimator $\hat{\eta}$.

4.3 Constrained pseudo-likelihood

Theorem 3.2. motivates the *monotonicity constraint* that $\frac{d\hat{\eta}}{du} > 0$, hence we define $\hat{\eta}_{INC}$ as a solution to

$$\text{maximize } \ell_t(\eta) \text{ subject to } \frac{d\hat{\eta}}{du} > 0.$$

An alternative strategy is to directly attack the variability is $\hat{p}_t$ due to non-uniqueness of $\hat{\eta}$. Therefore, we define $\hat{\eta}_{MC}$ (where MC stands for moment-constrained) as

$$\text{maximize } \ell_t(\eta) \text{ subject to } \int u^{k-1}\eta(u)du = \hat{p}_k^{UN}.$$

Thirdly, we can combine both the moment constraint and the monotonicity constraint, yielding $\hat{\eta}_{COM}$, which is obtained by solving

$$\text{maximize } \ell_t(\eta) \text{ subject to } \int u^{k-1}\eta(u)du = \hat{p}_k^{UN} \text{ and } \frac{d\hat{\eta}}{du} > 0.$$

Unfortunately, none of the three density estimators are uniquely defined. An easy way to see this is to transform the parameterization of $\eta(u)$, defining

$$\eta(u) = \int_0^u \xi(u)du;$$

the monotonicity constraint is equivalent to the condition that $\xi > 0$, and the moment condition translates into a linear equality constraint on ξ.

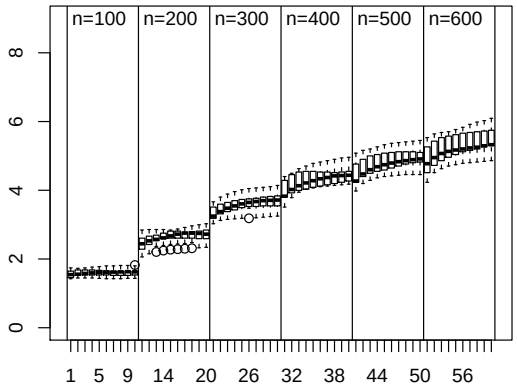

Figure 1: Extrapolation classification performance for CIFAR data. (This simulation needs to be fixed later.) PMLE: maximum psuedolikelihood. MCPMLE: Moment-constrained max psuedolikelihood. Info: Zheng and Benjamini's info-theoretic method. Unbiased: U-statistic (cannot be used to extrapolate.)

5 Results

6 Discussion

Acknowledgments

CZ is supported by an NSF graduate research fellowship.

References

[X] Ng, Andrew Y., and Michael I. Jordan. "On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes." (2002).

[X] Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage*, 56(2), 400-410.

[X] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning.* Vol. 1. Springer, Berlin: Springer series in statistics, 2008.