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. *Recognition systems* employ such classifiers in order to recognize people, spoken words, or chemicals: it is often of interest to know how many species the system can be trained to recognize before dropping below a minimum accuracy threshold. However, before such systems are deployed, typically only a small number of species are available for testing the system. Can we predict how well the recognition system will scale with an increased number of classes? We distinguish between two types of multi-class classifiers: *separable* classifiers, which include k-nearest neighbors, multinomial regression, one-vs-one and one-vs-all classifiers, *non-separable* classifiers, such as deep neural networks and decision trees. For recognition systems based on separable classifiers, the problem of predicting scalability reduces to the problem of estimating the higher-order moments of a *conditional accuracy distribution*, which in turn can be estimated from data.

1 Introduction

Object recognition, face recognition (or more generally person recognition) and language are a few of the cognitive building blocks which are fundamental to human cognition, and which can be understood as examples of generalized classification tasks. Machine classification can be employed to mimic this power of recognition. A robot equipped with a camera can algorithmically segment its input image into objects, and to learn to recognize unique objects and people which regularly appear in its environment. A general approach to implement such a recognition ability starts by employing some parametric featurization of the object to be identified. For example, for the task of face recognition, one might define features such as the proportions between the eyes and the relative position and size of the nose. The full domain of the recognition task is a collection of *measurements* (e.g. photographs of faces) which are associated to different *species*. The recognition system can be implemented by training a multi-class classifier to assign measurements to their corresponding species. While the system is in deployment, new species may be added to the system: when this happens, the classifier is retrained (or updated) using training data from the species to be added.

A limitation to such recognition systems, whether they be natural or artificial, is that the performance of the system (in terms of correct classification) can degrade if there are too many species. A face

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

recognition algorithm can have very high success rate if it only needs to distinguish between 100 different faces, but its identifications may be less reliable when it needs to distinguish between 10000 different faces. The consequences of such errors may be severe: Cole (2005) lists 22 cases of fingerprint misindentification in criminal trials.

Therefore, in the case of engineering recognition systems, it is of much practical interest to be able to evaluate the reliability of such systems before they are deployed. Yet, during the development phase, typically data from only a fraction of the species in the domain are available. From the empirical performance of the classifier on this initial subset of species, can we predict the performance of the system on a larger subset of species (or the entire domain)?

A related problem arises in studying *naturally occurring* recognition systems: for instance, human memory. Neuroscientists may be directly interested in the number of different cues which can be recalled by a subject. A similar dilemma arises, where data from only small number of species can be obtained, due to experimental constraints. From this data, can we predict the number of species which can be distinguished by the recognition system, above a minimum accuracy threshold?

We address this problem of *performance extrapolation* under the assumption that the initial subset of species is an i.i.d. sample from a larger population of species. Section 2 formalizes the recognition problem, and section 3 formalizes the problem of performance extrapolation. For a restricted family of classifiers–*separable* classifiers, we show that the problem of performance extrapolation reduces to a problem of nonparametric moment estimation. But this is still a difficult problem, and in section 4 we find limitations to traditional maximum likelihood and Bayesian inference techniques. We propose a method called *moment-constrained maximum pseudolikelihood*, and demonstrate that it has reasonable performance in simulated and real data examples in section 5; section 6 concludes.

2 The recognition problem

In the recognition problem, there exists a large or infinite population of *species*, e.g. individuals, and an infinite population of *measurements*, and each measurement is associated with a single species. A recognition system is an algorithm which contains a database of measurements from a number of species: each measurement is represented within the recognition system by a real vector $y \in \mathbb{R}^p$ which lies in the space $\mathcal{Y}$, and labeled with an integer-valued ID key. The ID key is the recognition system's internal representation of the species associated to the measurements: we assume that the database has the property of *integrity*, where any two measurements in the database share the same ID key if and only if they are associated with the same species. Given a query measurement, the system answers with an ID number: this answer is correct if and only if the query measurement and the database measurements labeled with the same ID are associated to the same species. Without loss of generality, take the ID keys to be $\{1, 2, \ldots, k\}$, and let $y^{(i),j}$ denote the jth measurement associated with ID key i in the database; let n_i be the number of measurements associated to the ith ID key, and $n = \sum_{i=1}^{k} n_i$ be the total number of measurements.

The recognition system maps queries y to an ID key; let us refer to this mapping as $f : \mathcal{Y} \to \{1, \ldots, k\}$. New data can be added to the recognition system over time; we assume it is done in a manner so that the integrity of the database is preserved. Let $t = 1, 2, \ldots$ refer to discrete time stages, and let k_t denote the number of classes in the database at time t, and f_t denote the mapping from queries to answers at time t; also define $n_{i,t}$ analogously.

2.1 Multi-class classification

A natural approach for implementing a recognition system is through *supervised learning*: the data is input into a supervised learning algorithm in order to *learn* a classification rule $f(y)$. Examples of multi-class classifiers include k-nearest neighbors, multinomial logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), decision trees, and random forests, as well as the two 'divide and conquer' approaches, one-vs-one (OVO) and one-vs-all (OVA) (Friedman et al, 2008.) A unified definition of a classifier is as follows: a k-class classifier $\mathcal{C}$ is a function which takes $k + 2$ arguments: the first k arguments are *probability distributions* $F_1, \ldots, F_k$, the $(k + 1)$st argument is a vector of prior probabilities $\vec{\pi}$, and the $(k + 2)$nd argument is a *query* y [1]. The level

[1] We treat randomized classifiers (such as random forests) as a probability distribution over deterministic classifiers.

sets $\{y : \mathcal{C}(F_1, \ldots, F_k, y) = k\}$ are called *decision regions.* We say the classifier is *continuous* if and only if the *decision regions* of $\mathcal{C}$ are continuous in the first k arguments with respect to the topology of weak convergence[2]. Most of the commonly used classifiers satisfy this definition of continuity: an exception is k-nearest neighbors with fixed k, but then again, k-nearest neighbors with $\lim_{n\to\infty} k/n \in (0, 1)$ is continuous. *Training a classifier* refers to defining a classification rule

$$f(y) = \mathcal{C}(\hat{F}_1, \ldots, \hat{F}_k, \vec{\pi}, y)$$

where $\hat{F}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \delta_{y^{(i),j}}$ is the empirical distribution, and where $\vec{\pi}$ is commonly set to either be the uniform distribution on k elements, or the empirical proportions of the classes, but can be adjusted in order to favor certain classes over others.

For theoretical purposes, let us assume that each species can be *uniquely* parameterized by a parameter vector x, and let $\mathcal{X}$ denote the space of parameter vectors. Note that the parameter x may or may not observed by the recognition system: in the latter case, the system only has access to ID keys which are non-informative of $p(y|x)$. Let $p(x)$ be the population distribution of species, and assume that a conditional feature vector distribution $p(y|x)$ be defined for every $x \in \mathcal{X}$. We require that $p(x)$ be a density with respect to Lesbegue measure, but $p(y|x)$ is allowed to include Dirac delta components. The database of the classifier therefore consists of pairs $\{(x^{(i)}), y^{(i),j})\}$ where $x^{(i)}$ is the parameter of the ith species in the database.

Whenever new data is added, the classifier is retrained. Depending on which multi-class classifier is used, we can categorize the recognition system as either a *separable* system, or a *non-separable* system.

2.2 Separability

The property of separability captures the intuitive notion that *information is not shared between classes*. To formalize the notion of separability, we begin with the class of *scoring rule*-based classifier. A scoring rule $\mathcal{Q}$ is a real-valued function which takes three arguments: a distribution F on $\mathcal{Y}$, a prior probability π, and a query measurement $y \in \mathcal{Y}$. A scoring-based classifier defines

$$\mathcal{C}(F_1, \ldots, F_k, \vec{\pi}, y) = \text{argmax}_i \mathcal{Q}(F_i, y, \vec{\pi}_i).$$

For notational convenience, we assume that ties occur with probability zero: that is, $p(x, y)$ and $\mathcal{Q}$ jointly satisfy the *tie-breaking* property: for any x and $\pi \in [0, 1]$, letting $F_n(x)$ be the empirical distribution of n points drawn from $p(y|x)$, and $Y, Y' \overset{iid}{\sim} p(y|x)$, we have

$$\Pr[\mathcal{Q}(F_n(x), \pi, Y) = \mathcal{Q}(F_n(x), \pi, Y')] = 0. \tag{1}$$

Quadratic discriminant analysis and Naive Bayes are two examples of scoring-based classifiers. For QDA, the scoring rule is given by

$$\mathcal{Q}_{QDA}(F, \pi, y) = -(y - \mu(F))^T \Sigma(F)^{-1} (y - \mu(F)) - \log\det(\Sigma(F)) + 2\log\pi$$

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

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

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

A *separable* classifier is a classifier which can be *approximated* by a scoring-based classifier in a large-sample limit. Let $\hat{F}_{i,t}$ denote the empirical distribution of the measurements of species i at time t, and $\hat{F}_t$ denote the empirical distribution of all measurements at time t.

Definition 2.1. (Separability) Suppose that for time stages $t = 1, 2, \ldots$, $k_t = O(t)$ and $n_{i,t} = O(t)$, with $x^{(1)}, x^{(2)}, \ldots$ drawn i.i.d. from $p(x)$, and $y^{(i),1}, \ldots$ drawn i.i.d. from $p(y|x^{(i)})$ for each $i = 1, 2, \ldots$. A recognition system (characterized by mappings f_t) is considered *separable* if and only there exists a scoring rule $\mathcal{Q}$ and probability vector $\tilde{\pi}$ such that defining

$$\tilde{f}_t(y) = \text{argmax}_{i=1}^{k_t} \mathcal{Q}(\hat{F}_{i,t}, \tilde{\pi}_i, y)$$

[2]We say that a randomized classifier is continuous if and only if it is continuous with probability one.

we have

$$\lim_{t\to\infty} \frac{1}{k_t} \sum_{i=1}^{k_t} \Pr[f_t(Y) = \tilde{f}_t(Y) | Y \sim p(y|x^{(i)})] \to 1.$$

We will show that recognition systems based on certain implementations of k-nearest neighbors, LDA, one-vs-one, or one-vs-all classifiers satisfy this definition of separability.

Definition 2.2.(i) Define a binary classifier $\mathcal{B}$ as a binary-valued mapping with four arguments: distributions F_0, F_1, prior probability π_0 and a query y. A one-vs-one (OVO) recognition system is defined by

$$f_t(y) = \text{argmax}_{i=1}^{k_t} \sum_{j\neq i} I\left(\mathcal{B}(\hat{F}_{i,t}, \hat{F}_{j,t}, \frac{n_{i,t}}{n_{i,t}+n_{j,t}}, y) = 0\right),$$

resolving ties arbitrarily.

(ii) Define a binary scoring rule $\mathcal{D}$ as a real-valued mapping with four arguments: distributions F_0, F_1, prior probability π_0, and a query y. A one-vs-all (OVA) recognition system is defined by

$$f_t(y) = \text{argmax}_{i=1}^{k_t} \mathcal{D}\left(\hat{F}_{i,t}, \sum_{j\neq i} \frac{n_{j,t}}{n_t - n_{i,t}} \hat{F}_{j,t}, \frac{n_{i,t}}{n_t}, y\right).$$

(iii) Let d be a distance metric on $\mathcal{Y}$. Let $D_t(y)$ denote the induced distribution of $d(Y, y)$ when $Y \sim \hat{F}_t$, and let $d_{\alpha,t}$ denote the α-quantile of $D_t(y)$. A kNN recognition system with neighborhood size α is defined by

$$f_t(y) = \text{argmax}_{i=1}^{k_t} \Pr[d(y, Y) < d_{\alpha,t} | Y \sim \hat{F}_{i,t}].$$

(iv) Assume WLOG that $y_1 = 1$ for all $y \in \mathcal{Y}$, and let B^t be a $p \times k_t$ matrix which minimizes the log-likelihood

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

A multinomial logistic regression recognition system is defined by

$$f_t(y) = \text{argmax}_{i=1}^{k_t} \langle y, B_i^t\rangle.$$

(v) An LDA recognition system is defined by

$$f_t(y) = \text{argmax}_{i=1}^{k_t} 2\log\vec{\pi}_i - (y - \mu(\hat{F}_{i,t}))^T \Sigma(\hat{F}_t)^{-1}(y - \mu(\hat{F}_{i,t})).$$

for arbitrary $\vec{\pi}$.

Theorem 2.1. *(i) an OVO recognition systems equipped with a continuous binary classifier is separable; (ii) an OVA recognition systems equipped with a continuous binary scoring rule is separable; (iii) a kNN recognition system with fixed neighborhood size $\alpha \in (0, 1)$ is separable; (iv) multinomial logistic regression recognition systems are separable; (v) LDA recognition systems are separable.*

3 Prediction Extrapolation

3.1 Problem formulation

Recall the notation from section 2.1. In order to formalize the problem of *extrapolation*, we adopt the language of stochastic processes. Let $(\Omega, \mathcal{F}, \mathbb{P})$ define a probability space. Let $t = 1, 2, 3, \ldots$ index discrete time steps; each time step is associated with a filtration $\mathcal{F}_t$, with $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}$. At time zero, we have not observed any data. At time k, we sample a new species $x^{(k)}$ from the population distribution $p(x)$, and also observe r replicates $y^{(i),1}, \ldots, y^{(i),r}$ from the conditional distribution $p(y|x^{(i)})$. Choose some $r_1 < r$: this determines the number of training observations $S(x^{(i)}) = \{y^{(i),1}, \ldots, y^{(i),r_1}\}$ from each class. The filtration at time t is defined as the σ-algebra induced by observations from species $x^{(1)}, \ldots, x^{(k)}$, hence $\mathcal{F}_t = \sigma\{(x^{(i)}, y^{(i),j})\}_{i=1,j=1}^{t,r}$.

The classifier also changes with time. At time t, $f^{(t)}$ is (random) function from $\mathcal{Y}$ to $\{x^{(1)}, \ldots, x^{(k)}\}$. The randomness is due to the variability of the training set, hence $f^{(t)}$ is independent of the test data $\{y^{(i),r_1+1}, \ldots, y^{(i),r}\}$ for $i = 1, \ldots, t$. Furthermore, we assume that the classifier $f^{(t)}$ is constructed in the following way. The user chooses an algorithm $\mathcal{Q}$ which constructs scoring functions $q^{(i)}$ from the training data for the ith class:

$$q^{(i)} = \mathcal{Q}(x^{(i)}, S(x^{(i)})).$$

Each $q^{(i)}$ is a function from $\mathcal{Y}$ to $\mathbb{R}$. The classifier $f^{(t)}$ is defined

$$f^{(t)}(y) = \text{argmax}_i q^{(i)}(y).$$

The generalization accuracy at time t is defined

$$\text{acc}^{(t)} = \frac{1}{k} \sum_{i=1}^{k} \Pr[f^{(t)}(y) = i | y \sim p(y|x^{(i)})].$$

The extrapolation problem is the problem of predicting $\text{acc}^{(K)}$ using only information known at time $k < K$.

3.2 Conditional accuracy

The optimal predictor of the generalization error (in mean square) is the conditional expected generalization error, $\mathbf{E}[\text{acc}^{(t)}|\mathcal{F}_k]$. However, it is easier to work with the unconditional expected generalization error $p_t \stackrel{def}{=} \mathbf{E}[\text{acc}^{(t)}]$.

Define the *conditional accuracy* function $u(x, y, S(x))$ which maps a data triple consisting of a species x, a *test* observation y, and a set of r_1 training replicates $S(x) = \{y^1, \ldots, y^{r_1}\}$ to a real number in $[0, 1]$. The conditional accuracy gives the probability that for independently drawn $(X, S(X))$ such that $X \sim p(x)$, and $S(X) = \{Y^1, \ldots, Y^{r_1}\}$ with $Y^i \stackrel{iid}{\sim} p(y|X)$, that the scoring function $\mathcal{Q}(x, S(x))$ will give a higher score to y than the scoring function $\mathcal{Q}(X, S(X))$:

$$u(x, y, S(x)) = \Pr[(\mathcal{Q}(x, S(x)))(y) > (\mathcal{Q}(X, s(X)))(y)].$$

Define the *conditional accuracy* distribution μ as the law of $u(X, Y, S(X))$ when $X \sim p(x)$, $Y \sim p(y|X)$, and $S(X)$ is drawn as specified above. 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 U be defined as the random variable*

$$U = u(X, Y, S(X))$$

for X, Y drawn from $p(x, y) = p(x)p(y|x)$, and $S(X) = \{Y^1, \ldots, Y^{r_1}\}$ with $Y^i \stackrel{iid}{\sim} p(y|X)$ Then $p_k = \boldsymbol{E}[U^{k-1}]$.

Proof. Write $q^{(i)} = \mathcal{Q}(x^{(i)}, S(x^{(i)}))$ Note that by using conditioning and conditional independence, p_k can be written

$$\begin{aligned}
p_k &= \mathbf{E}\left[\frac{1}{k}\sum_{i=1}^{k} \Pr[q^{(i)}(Y) > \max_{j \neq i} q^{(j)}(Y)]\right] \\
&= \mathbf{E}\left[\Pr[q^{(1)}(Y) > \max_{j \neq 1} q^{(j)}(Y)]\right] \\
&= \mathbf{E}[\Pr[q^{(1)}(Y) > \max_{j \neq 1} q^{(j)}(Y) | X^{(1)}, Y, S(X^{(1)})]] \\
&= \mathbf{E}[\Pr[\cap_{j>1} q^{(1)}(Y) > q^{(j)}(Y) | X^{(1)}, Y, S(X^{(1)})]] \\
&= \mathbf{E}[\prod_{j>1} \Pr[q^{(1)}(Y) > q^{(j)}(Y) | X^{(1)}, Y, S(X^{(1)})]] \\
&= \mathbf{E}[\Pr[q^{(1)}(Y) > q^{(2)}(Y) | X^{(1)}, Y, S(X^{(1)})]^{k-1}] \\
&= \mathbf{E}[u(X^{(1)}, Y, S(X^{(1)}))^{k-1}].
\end{aligned}$$

□

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 $e^{(t)}$.

3.3 Properties of the conditional accuracy distribution

The conditional error distribution μ is determined by $p(x, y)$ and $\mathcal{Q}$. What can we say about the the conditional accuracy distribution without making any assumptions on either $p(x, y)$ or $\mathcal{Q}$? The answer is: not much–for an arbitrary probability measure ν on $[0, 1]$, one can construct $p(x, y)$ and $\mathcal{Q}$ such that $\mu = \nu$.

Theorem 3.2. *Let U be defined as in Theorem 2.1, and let μ denote the law of U. Then, for any probability distribution ν on $[0, 1]$, one can construct a joint distribution $p(x, y)$ and a scoring rule $\mathcal{Q}$ such that $\mu = \nu$.*

Proof. Let X and Y have the degenerate joint distribution $X = Y \sim Unif[0, 1]$. 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

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

One can verify that $u(x, x, S(x)) = H(u)$. Therefore, the cdf of U is equal to G, as needed. □

In practice, however, the scoring rule $\mathcal{Q}$ must approximate a monotonic function of the conditional density $p(y|x)$ in order to yield an effective classifier. It is therefore notable that in the case that (X, Y) have a density with respect to Lesbegue measure, taking an *optimal* scoring rule, with the property that $\mathcal{Q}(x, y, S(x)) = g(p(y|x))$ for monotonic g, the distribution of U has a monotonically increasing density.

Theorem 3.1. *Let U be defined as in Theorem 2.2, and let μ denote the law of U. Suppose (X, Y) has a density $p(x, y)$ with respect to Lebesgue measure on $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{Q}(x, S(x))$ satisfies the property of monotonicity*

$$p(y|x) > p(y'|x) \textit{ implies } \mathcal{Q}(x, S(x))(y) > \mathcal{Q}(x, S(x))(y')$$

and the property of tie-breaking (1)*, then μ has a density $\eta(u)$ on $[0, 1]$ which is monotonic in u.*

Proof. Choose $0 < u < v < 1$ and $0 < \delta < \min(u, 1 - v, v - u)$. For $x \in \mathcal{X}$, define the set

$$\underline{J}_x = \{y \in \mathcal{Y} : \int_{\mathcal{Y}} I(p(y|x) > p(w|x))p(w|x)dw \in [u - \delta, u + \delta]\}$$

and

$$\bar{J}_x = \{y \in \mathcal{Y} : \int_{\mathcal{Y}} I(p(y|x) > p(w|x))p(w|x)dw \in [v - \delta, v + \delta]\}$$

One can verify that for all $x \in \mathcal{X}$,

$$\int_{\underline{J}_x} p(y|x)dy \leq \int_{\bar{J}_x} p(y|x)dy.$$

Yet, since

$$\Pr[U \in [u - \delta, u + \delta]] = \Pr[\cup_{\mathcal{X}} x \times \underline{J}_x]$$
$$\Pr[U \in [v - \delta, v + \delta]] = \Pr[\cup_{\mathcal{X}} x \times \underline{J}_x].$$

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 $U = u(x, y)$ cannot be directly observed, we can estimate $u(x^{(i)}, y^{(i),r_1+j}, S(x^{(i)}))$ for any $1 \le i \le k$, $1 \le j \le r_2$, where $r_2 = r - r_1$ is the number of testing repeats.

Theorem 4.1 *Assume that $\mathcal{Q}$ satisfies the tie-breaking property* (1)*. Define*

$$V_{i,j} = \sum_{i=1}^{k} I(q^{(i)}(y^{(i),j}) > q^{(j)}(y^{(i),j})).$$

Then

$$V_{i,j} \sim \textit{Binomial}(k, u(x^{(i)}, y^{(i),j}, S(x^{(i)}))).$$

□

The proof of Theorem 4.1 follows from the same conditioning argument used in Theorem 3.1.

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. A third approach is an adaptation of the *mutual information estimator* developed by (Anon 2016). Anon 2016 develop an asymptotic theory which relates the Bayes error to the mutual information $I(X; Y)$ and vice versa. This allows us to estimate "information" from classification error p_k, and then predict the generalization error p_t from the estimated information: details are given in Section 4.4.

4.1 Unbiased estimation

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

The theory of U-statistics provides the minimal variance unbiased estimator for η^t:

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

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

$$p_t = \mathbf{E}\left[\frac{1}{kr_2}\sum_{i=1}^{k}\sum_{j=1}^{r_2}\frac{\binom{V_{i,j}}{t}}{\binom{k}{t}}\right]. \tag{2}$$

The problem of *extrapolation* concerns the case $t > k$, in which the expression (2) is undefined. Still, the estimator (2) is worthy of study, since it has close to optimal performance for the case $t \le k$.

4.2 Maximum pseudo-likelihood

For fixed j the quantities $\{V_{1,j}, \ldots, V_{k,j}\}$ are mutually independent, and one can write a nonparametric likelihood for the density $\eta(u)$ of U:

$$\mathcal{L}_j(\eta) = \prod_{i=1}^{k} \binom{k}{V_{i,j}} \eta(u)^{V_{i,j}} (1 - \eta(u))^{k - V_{i,j}}.$$

However, it is not possible to write a likelihood for $\eta(u)$ depending on all the terms $V_{i,j}$,

4.3 Constrained pseudo-likelihood

4.4 Information-based methodology

[Mostly copy and paste, needs fixing]

We start by restating the results of ZB 2016. The asymptotic regime considered is a sequence of joint distributions $p(x, y)$ where the dimensionality of x goes to infinity. A specific example of a sequence in this regime is one where X is d-dimensional multivariate normal with covariance identity I_d, and $Y = X + E$, where E is an independent multivariate normal with covariance cdI_d, for some fixed constant $c > 0$.

Theorem 2. *Let $p^{[d]}(x, y)$ be a sequence of joint densities for $d = 1, 2, \ldots$ as given above. Further assume that*

A1. $\lim_{d\to\infty} I(X^{[d]}; Y^{[d]}) = \iota < \infty$.

A2. There exists a sequence of scaling constants $a_{ij}^{[d]}$ and $b_{ij}^{[d]}$ such that the random vector $(a_{ij}\ell_{ij}^{[d]} + b_{ij}^{[d]})_{i,j=1,\ldots,K}$ converges in distribution to a multivariate normal distribution.

A3. There exists a sequence of scaling constants $a^{[d]}$, $b^{[d]}$ such that

$$a^{[d]}u(X^{(1)}, Y^{(2)}) + b^{[d]}$$

converges in distribution to a univariate normal distribution.

A4. For all $i \neq k$,

$$\lim_{d\to\infty} Cov[u(X^{(i)}, Y^{(j)}), u(X^{(k)}, Y^{(j)})] = 0.$$

Then for p_K as defined above, we have

$$\lim_{d\to\infty} 1 - p_K = \pi_K(\sqrt{2\iota})$$

where

$$\pi_K(c) = 1 - \int_{\mathbb{R}} \phi(z - c)\Phi(z)^{K-1}dz$$

where ϕ and Φ are the standard normal density function and cumulative distribution function, respectively.

By combining Theorems 1 and 2, we immediately compute the limiting distribution of P in the given regime **Corollary.***Let $p^{[d]}(x, y)$ be a sequence of joint densities satisfying A1-A4 as stated in Theorem 2. For any d, let $P^{[d]}$ as defined in Theorem 1. Then $P^{[d]}$ converges in distribution to P, where the cdf of P is given by*

$$\Pr[P < t] = \int_0^t \frac{\phi(\Phi^{-1}(u) - \sqrt{2\iota})}{\phi(\Phi^{-1}(u))}du.$$

Proof. By Theorem 1, the moments of $P^{[d]}$ are given by

$$\mathbf{E}[P^{[d]k-1}] = p_k^{[d]}$$

and meanwhile, Theorem 2 implies that

$$\lim_{d\to\infty} p_k^{[d]} = \int_{\mathbb{R}} \phi(z - \sqrt{2\iota})\Phi(z)^{k-1}dz.$$

Let Z be a normal $N(\sqrt{2\iota}, 1)$ variate, and define $P = \bar{\Phi}(Z)$. Then it is clear that

$$\lim_{d\to\infty} \mathbf{E}[P^{[d]k-1}] = \int_{\mathbb{R}} \phi(z - \sqrt{2\iota})\Phi(z)^{k-1}dz = \mathbf{E}[P^{k-1}]$$

for all k. Since both $P^{[d]}$ and P lie in the compact interval $[0, 1]$, the fact that the moments of $P^{[d]}$ converge to the moments of P implies that the distribution of $P^{[}d]$ converges to the distribution of P. $\square$.

The corollary identifies a parametric family of distributions $\mathcal{P} = \{P_\iota\}$ indexed by the mutual information ι. For given ι, the density of P_ι if given by

$$g_\iota(u) = \frac{\phi(\Phi^{-1}(u) - \sqrt{2\iota})}{\phi(\Phi^{-1}(u))}.$$

Note the special case $\iota = 0$, which yields $P_0 = U$, the uniform distribution on $[0, 1]$. This implies that in the special case that X is independent of Y, and hence optimal classification does no better than random guessing, $p_k = \frac{1}{k}$, which indeed matches the moments of the uniform distribution

$$\mathbf{E}[U^{k-1}] = \int_0^1 u^{k-1} du = \frac{1}{k}.$$

We see that for any given finite-dimensional joint distribution $p(x, y)$, if the distribution of P lies close to a member of the parametric family $\mathcal{P}$, the information-theoretic methodology for estimating p_N from p_k will be accurate.

Acknowledgments

CZ is supported by an NSF graduate research fellowship.

References

[X] Anonymous, A. "High-dimensional estimation of mutual information via classification error."

[X] Gastpar, M. Gill, P. Huth, A. Theunissen, F. "Anthropic Correction of Information Estimates and Its Application to Neural Coding." *IEEE Trans. Info. Theory*, Vol 56 No 2, 2010.

[X] A. Borst and F. E. Theunissen, "Information theory and neural coding" Nature Neurosci., vol. 2, pp. 947?957, Nov. 1999.

[X] L. Paninski, "Estimation of entropy and mutual information," Neural Comput., vol. 15, no. 6, pp. 1191?1253, 2003.

[X] I. Nelken, G. Chechik, T. D. Mrsic-Flogel, A. J. King, and J. W. H. Schnupp, "Encoding stimulus information by spike numbers and mean response time in primary auditory cortex," J. Comput. Neurosci., vol. 19, pp. 199?221, 2005.

[X] Cover and Thomas. Elements of information theory.

[X] Muirhead. Aspects of multivariate statistical theory.

[X] van der Vaart. Asymptotic statistics.

[X] F. E. Theunissen and J. P. Miller, "Representation of sensory information in the cricket cercal sensory system. II. information theoretic calculation of system accuracy and optimal tuning-curve widths of four primary interneurons," J. Neurophysiol., vol. 66, no. 5, pp. 1690?1703, 1991. [8]

[X] De Campos, Luis M. "A scoring function for learning Bayesian networks based on mutual information and conditional independence tests." The Journal of Machine Learning Research 7 (2006): 2149-2187.

[X] Linsker, Ralph. "An application of the principle of maximum information preservation to linear systems." Advances in neural information processing systems. 1989.

[X] Speed, Terry. "A correlation for the 21st century." Science 334.6062 (2011): 1502-1503.

[X] Beirlant, J., Dudewicz, E. J., Gy orfi, L., & der Meulen, E. C. (1997). Nonparametric Entropy Estimation: An Overview. International Journal of Mathematical and Statistical Sciences, 6, 17–40. doi:10.1.1.87.5281

[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.

[X] Tse, David, and Pramod Viswanath. *Fundamentals of wireless communication.* Cambridge university press, 2005.

[X] Banerjee, Arpan, Heather L. Dean, and Bijan Pesaran. "Parametric models to relate spike train and LFP dynamics with neural information processing." *Frontiers in computational neuroscience* 6 (2011): 51-51.