Extrapolating Multi-class classification curves

Charles Zheng and Yuval Benjamini

April 2, 2016

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. Fundamentally, humans learn to recognize objects, persons, or linguistic entities and learn equivalence relationships between perceived objects, words, or individuals. A child will encounter their neighbor's dog for the first time, and at a later date, encounter the same dog in a different time or place. But they will recognize that the dog is the same dog they met before.

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. Such features can be estimated from the video input, and while the estimated features might vary slightly depending on the configuration of the object (e.g. posture of the body, emotional expression of the face, opening/closing of the mouth) and while some features might be unobserved because of occlusion, the regularity of the the identified features can still be used as a basis for recognition. In the simplest case, a feature vector can be extracted from a given object from the input data, and this feature vector can be evaluated against a library of learned models for the distribution of the feature vector for various objects. In other words,

given a method for extracting objects from input and producing feature vectors for those objects, one can employ multi-class classification to identify the object corresponding to the feature vector. In our discussion thus far, we neglected to mention how to learn object categories for novel objects, but this falls outside of the scope of the paper.

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 categories. A face 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. In humans, it is known that repetition learning is hampered when there are too many similar concepts to be learned.

The tradeoff between performance and the number of categories can be summarized by a *classification curve* while plots the accuracy versus the number of categories. An interesting question is whether it is possible to *extrapolate* a classification curve: i.e. given the performance with up to N classes, can one predict the performance when the number of classes is increased to $2N$?

In the most general setting, only fairly trivial worst-case guarantees are possible. Suppose a system has an accuracy of p given the N-class problem, i.e., given a random feature vector, the system correctly identifies the class with probability p. Then, what can we say about the performance when N additional classes are added? In a very bad (not necessarily the 'worst') case, the N additional classes are identical copies or the original N classes, except we still consider it an error if a feature vector from one of the new classes is identified as coming from the corresponding class in the original set, or vice versa. In that case, a reasonable system would be the original system, but adding a coin flip to choose between the original classes or the copied new classes, and the resulting accuracy would be reduced from p to $p/2$ due to the coin flip. In general, we can state that when going from N classes to kN classes, the Bayes accuracy might be as low as p/k.

We can obtain more meaningful bounds if we are willing to assume more structure to the problem. In many applications, such as face recognition or word recognition, a reasonable model is to assume that there exists an infinite population of classes, and that given a classification problem with N classes, that those N classes were drawn i.i.d. from this infinite population. This is 'random classification model' is the setting we consider in this paper.

2 Setup

Let $\mathcal{X}$ be a space of objects to be recognized, and let the objects be uniquely parameterized by a real vector $x \in \mathcal{X}$. Let $y \in \mathcal{Y}$ be a possible feature vector for an object in $\mathcal{X}$. Let $p(x)$ be a distribution of random objects, and let a conditional feature vector distribution $p(y|x)$ be defined for every $x \in \mathcal{X}$.

The k-class random classification task is defined as follows. First, the k classes are drawn iid from $\mathcal{X}$, represented by vectors $x_1, \ldots, x_k$. One obtains training data consisting of independent (x, y) pairs, where x is equal to one of $x_1, \ldots, x_k$ and y is drawn from $p(y|x)$. Based on the training data, one constructs a classification rule f which maps a feature vector to one of $\{x_1, \ldots, x_k\}$. The accuracy of the classification rule is

$$\Pr[f(y) = x | x \sim \text{Unif}\{x_1, \ldots, x_k\}].$$

The accuracy depends on what model is used to learn the classification rule. Hence, discussing the properties of the accuracy for finite sample sizes and specific classification rules is an immense task, falling outside of the scope of the paper. However, it is vastly easier to work with the accuracy of the *Bayes rule*, which is the optimal classification rule. This is because the Bayes rule f^* can be written explicitly as

$$f^*(y) = \text{argmax}_{x_1,\ldots,x_k} p(y|x).$$

Since there exist consistent estimators of the Bayes rule (under mild regularity conditions,) the theory we develop for the accuracy of the Bayes rule can be viewed as a large-sample approximation of the original problem. Henceforth we work with the accuracy of the Bayes rule for the k-class classification problem, P_k, where

$$P_k = \Pr[\text{argmax}_{x_1,\ldots,x_k} p(y|x) = x | x \sim \text{Unif}\{x_1, \ldots, x_k\}].$$

Notice that P_k is a random variable, since it varies depending on the sampling of $x_1, \ldots, x_k$. Hence the average Bayes accuracy p_k is defined as

$$p_k = E[P_k]$$

where the expectation is taken over the sampling distribution of $x_1, \ldots, x_k$. With these definitions, we can finally state the specific questions to be addressed in the paper.

- Suppose that for randomly drawn $x_1, \ldots, x_k$, one is given knowledge of $p(y|x_1), \ldots, p(y|x_k)$. Can one estimate p_N for some $N > k$?
- Suppose one is given $p_2, \ldots, p_k$. Can one estimate p_N?

The average Bayes accuracy is the complement of average Bayes error, introduced in Zheng and Benjamini 2016. The paper showed that in a particular asymptotic regime, one obtained the following relationship between the average Bayes error and the mutual information $I(X;Y)$.

$$1 - p_k \approx \pi_k(\sqrt{2I(X;Y)})$$

where

$$\pi_k(c) = \int \phi(x-c)\Phi(x)^{k-1}dx.$$

In principle, this result allows one to extrapolate from p_k to p_N for any $2 \leq k \leq N$ by the following means: first, obtain $\hat{I}(X;Y) = \frac{1}{2}(\pi_k^{-1}(1-p_k))^2$, secondly, estimate $p_N = \pi_N(\sqrt{2\hat{I}(X;Y)})$. While this method of extrapolation is consistent in the limit of the high-dimensional regime considered in the ZB 2016, it may be quite inaccurate in the finite dimensional cases.

3 Moment equivalence

Our key insight is as follows: the average Bayes accuracy p_k is the $k-1$th moment of a random variable P which takes values in $[0,1]$. Since the moments of P are completely determined by the distribution, the problem of predicting p_N can be reduced to the problem of estimating the distribution of P in a suitable way (such that the moments can be recovered consistently.)

The random variable P is constructed as follows:

- Draw $X \sim p(x)$.
- Draw $Y \sim p(y|x)$.
- Let $P = \int I(p(y|X) > p(y|x))p(x)dx$.

The fact that the average Bayes accuracies correspond to moments of P is stated in the following theorem.

Theorem 1. *Let P be defined as the random variable*

$$\Pr[p(Y|X) > p(Y|X')|X, Y]$$

for X, Y drawn from $p(x, y) = p(x)p(y|x)$, and X' drawn independently from $p(x)$. Then $p_k = \boldsymbol{E}[P^{k-1}]$.

Proof. Note that by using conditioning and conditional independence, p_k can be written

$$\begin{aligned}
p_k &= \mathbf{E}\left[\frac{1}{k}\sum_{i=1}^{k}\Pr[p(Y|X_i) > \max_{j\neq i} p(Y|X_j)]\right] \\
&= \mathbf{E}\left[\Pr[p(Y|X_1) > \max_{j\neq 1} p(Y|X_j)]\right] \\
&= \mathbf{E}[\Pr[p(Y|X_1) > \max_{j\neq 1} p(Y|X_j)|X_1, Y]] \\
&= \mathbf{E}[\Pr[\cap_{j>1} p(Y|X_1) > p(Y|X_j)|X_1, Y]] \\
&= \mathbf{E}[\prod_{j>1}\Pr[p(Y|X_1) > p(Y|X_j)|X_1, Y]] \\
&= \mathbf{E}[\Pr[p(Y|X_1) > p(Y|X_2)|X_1, Y]^{k-1}]
\end{aligned}$$

But observe that the conditional probability $\Pr[p(Y|X_1) > p(Y|X_2)|X_1, Y]$, viewed as a random variable, has the same distribution as P. $\square$.

Before discussing a general method to estimate P, we gain some insight about the 'information-based methodology' of ZB 2016 and possible shortcomings by explicitly calculating the distribution of P in the case when (a) the asymptotic regime considered in ZB 2016 is well-specified and (b) when the model of ZB 2016 is violated in a specific case.

4 Information-based methodology

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* $(a_{ij}\ell_{ij}^{[d]} + b_{ij}^{[d]})_{i,j=1,\dots,K}$ *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 $P^{\lfloor}d]$ 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.

[This allows us to get convergence rates ... to be continued.]

Conversely, what happens if the distribution P is not close to any member of $\mathcal{P}$? Will the information-theoretic method give an overestimate or underestimate of p_N? To address this question, consider the example of a *well-separated mixture*. Suppose $p(x, y)$ is a mixture of joint distributions $q_1(x, y)$ and $q_2(x, y)$ with mixing fraction α,

$$p(x, y) = \alpha q_1(x, y) + (1 - \alpha)q_2(x, y).$$

Let ι_i be the mutual information of X and Y when $X, Y \sim q_i$ for $i = 1, 2$. The well-separated condition is that q_1 and q_2 have disjoint support marginally, that is

$$\int_x \left| \int_y q_1(x, y)dy - \int_y q_2(x, y)dy \right| dx = 2;$$

$$\int_y \left| \int_x q_1(x,y)dx - \int_x q_2(x,y)dx \right| dy = 2.$$

Then it follows that P has the density

$$g(u) = 2\alpha(1-\alpha)\delta_1(u) + \alpha^2 g_{\iota_1}(u) + (1-\alpha)^2 g_{\iota_2}(u).$$