

Upper bounds for average Bayes accuracy in terms of mutual information

Charles Zheng and Yuval Benjamini

September 9, 2016

These are preliminary notes.

1 Introduction

Suppose X and Y are continuous random variables (or vectors) which have a joint distribution with density $p(x, y)$. Let $p(x) = \int p(x, y)dy$ and $p(y) = \int p(x, y)dx$ denote the respective marginal distributions, and $p(y|x) = p(x, y)/p(x)$ denote the conditional distribution.

Mutual information is defined

$$I[p(x, y)] = \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy.$$

ABE_k , or k -class Average Bayes accuracy is defined as follows. Let X_1, \dots, X_K be iid from $p(x)$, and draw Z uniformly from $1, \dots, k$. Draw $Y \sim p(y|X_Z)$. Then, the average Bayes accuracy is defined as

$$ABA_k[p(x, y)] = \sup_f \Pr[f(x_1, \dots, x_k, y) = Z]$$

where the supremum is taken over all functions f . A function f which achieves the supremum is

$$f_{Bayes}(x_1, \dots, x_k, y) = \operatorname{argmax}_{z \in \{1, \dots, k\}} p(y|x_z),$$

where an arbitrary rule can be employed to break ties. Such a function f_{Bayes} is called a *Bayes classification rule*. It follows that ABA_k is given explicitly

by

$$\text{ABA}_k = \frac{1}{k} \int \left[\prod_{i=1}^k p(x_i) dx_i \right] \int dy \max_i p(y|x_i).$$

2 Problem formulation

Let \mathcal{P} denote the collection of all joint densities $p(x, y)$ on finite-dimensional Euclidean space. For $\iota \in [0, \infty)$ define $C_k(\iota)$ to be the largest k -class average Bayes error attained by any distribution $p(x, y)$ with mutual information not exceeding ι :

$$C_k(\iota) = \sup_{p \in \mathcal{P}: I[p(x, y)] \leq \iota} \text{ABA}_k[p(x, y)].$$

A priori, $C_k(\iota)$ exists since ABA_k is bounded between 0 and 1. Furthermore, C_k is nondecreasing since the domain of the supremum is monotonically increasing with ι .

It follows that for any density $p(x, y)$, we have

$$\text{ABA}_k[p(x, y)] \leq C_k(I[p(x, y)]).$$

Hence C_k provides an upper bound for average Bayes error in terms of mutual information.

Conversely we have

$$I[p(x, y)] \geq C_k^{-1}(\text{ABA}_k[p(x, y)])$$

so that C_k^{-1} provides a lower bound for mutual information in terms of average Bayes error.

On the other hand, there is no nontrivial *lower* bound for average Bayes error in terms of mutual information, nor upper bound for mutual information in terms of average Bayes error, since

$$\inf_{p \in \mathcal{P}: I[p(x, y)] \leq \iota} \text{ABA}_k[p(x, y)] = \frac{1}{k}.$$

regardless of ι .

The goal of this work is to attempt to compute or approximate the functions C_k and C_k^{-1} .

3 Special case

We work out the special case where $p(x, y)$ lies on the unit square, and $p(x)$ and $p(y)$ are both the uniform distribution. Let \mathcal{P}^{unif} denote the set of such distributions, and

$$C_k^{unif}(\iota) = \sup_{p(x,y) \in \mathcal{P}^{unif} : I[p] \leq \iota} \text{ABA}_k[p].$$

In this case, letting $X_1, \dots, X_k \sim \text{Unif}[0, 1]$, and $Y \sim \text{Unif}[0, 1]$ define $Z_i(y) = p(y|X_i)$. We have $\mathbf{E}(Z(y)) = 1$ and,

$$I[p(x, y)] = \mathbf{E}(Z(Y) \log Z(Y))$$

while

$$\text{ABA}_k[p(x, y)] = k^{-1} \mathbf{E}(\max_i Z_i(Y)).$$

Letting g_y be the density of $Z(y)$, we have

$$I[p(x, y)] = \mathbf{E}(-H[g_Y])$$

and

$$\text{ABA}_k[p(x, y)] = \mathbf{E}(\psi_k[g_Y])$$

where

$$H[g] = - \int g(x) x \log x dx$$

and

$$\psi_k[g] = \int x g(x) G(x)^{k-1} dx$$

for $G(x) = \int_0^x g(t) dt$. Additionally g_y satisfies the constraint $\int x g(x) dx = 1$ since $\mathbf{E}[Z(y)] = 1$.

Define the set $D = \{(\alpha, \beta)\}$ as the set of possible values of $(-H[g], \psi_k[g])$ taken over all distributions g supported on $[0, \infty)$ with $\int x g(x) dx = 1$. Next, let $\mathcal{C}(D)$ denote the convex hull of D . It follows that $(I[p], \text{ABA}_k[p]) \in \mathcal{C}(D)$ since the pair is obtained via a convex average of points $(-H[g_y], \psi_k[g_y])$.

Define the upper envelope of D as the curve

$$d_k(\alpha) = \sup\{\beta : (\alpha, \beta) \in D\}.$$

We make the claim (to be shown in the following section) that $d_k(\alpha)$ is convex in α . As a result, the upper envelope of D is also the upper envelope of $\mathcal{C}(D)$. This in turn implies that $C_k^{unif}(\iota) = d_k(\iota)$. We establish these results, along with a open-form expression for C_k^{unif} , in the following section.

3.1 Variational methods

Consider the quantile function $Q(t) = G^{-1}(t)$. $Q(t)$ must be a continuous function from $[0, 1]$ to $[0, \infty)$. We can rewrite the moment constraint $\mathbf{E}[g] = 1$ as

$$\int_0^1 Q(t) dt = 1.$$

Meanwhile, $\beta = \psi_k[g]$ takes the form

$$\beta = \int_0^1 Q(t) x^{k-1} dt.$$

and $\alpha = -H[g]$ takes the form

$$\alpha = \int_0^1 Q(t) \log Q(t) dt.$$

To find the upper envelope, it will be useful to write the Langrangian

$$\begin{aligned} \mathcal{L}[g] &= \lambda \int_0^1 Q(t) dt + \mu \int_0^1 Q(t) x^{k-1} dt + \lambda \int_0^1 Q(t) \log Q(t) dt \\ &= \int_0^1 Q(t) (\lambda + \mu x^{k-1} + \nu \log Q(t)) dt. \end{aligned}$$

In order for a quantile function $Q(t)$ to be on the upper envelope, it must be a local maximum of $-H$ with respect to small perturbations. Therefore, consider the functional derivative

$$D[\xi] = \lim_{\epsilon \rightarrow 0} \frac{\mathcal{L}[g + \epsilon \xi] - \mathcal{L}[g]}{\epsilon}.$$

We have

$$D[\xi] = \int_0^1 \xi(t) (\lambda + \nu + \mu x^{k-1} + \nu \log Q(t)) dt.$$

Now consider the following three cases:

- $Q(t)$ is strictly monotonic, i.e. $Q'(t) > 0$.
- $Q(t)$ is differentiable but not strongly monotonic:
- $Q(t)$ is not strongly monotonic: there exist intervals $A_i = [a_i, b_i)$ such that $Q(t)$ is constant on A_i , and isolated points t_i where $Q'(t_i) = 0$.

Strictly monotonic case. Because Q is defined on a closed interval, strict monotonicity further implies the property of *strong monotonicity* where $\inf_{[0,1]} Q'(t) > 0$. Therefore, for any differentiable perturbation $\xi(t)$ with $\sup |\xi'(t)| < \infty$, and further imposing that $\xi(0) \geq 0$ in the case that $Q(0) = 0$, there exists some $\epsilon > 0$ such that $(Q + \epsilon\xi)(t)$ is still a valid quantile function. Therefore, in order for $Q(t)$ to be a local maximum, we must have

$$0 = \lambda + \nu + \mu x^{k-1} + \nu \log Q(t)$$

for $t \in [0, 1]$. This implies that

$$Q(t) = c_0 e^{-c_1 x^{k-1}}$$

for some $c_0, c_1 \geq 0$.

Other cases. (TODO) We have to show that these cannot be local maxima.

4 General case

We claim that the constants $C_k^{unif}(\iota)$ obtained for the special case also apply for the general case, i.e.

$$C_k(\iota) = C_k^{unif}(\iota).$$

We make use of the following Lemma:

Lemma. *Suppose X, Y, W, Z are continuous random variables, and that $W \perp Y|Z$, $Z \perp X|Y$, and $W \perp Z|(X, Y)$. Then,*

$$I[p(x, y)] = I[p((x, w), (y, z))]$$

and

$$ABA_k[p(x, y)] = ABA_k[p((x, w), (y, z))].$$

Proof. Due to conditional independence relationships, we have

$$p((x, w), (y, z)) = p(x, y)p(w|x)p(z|y).$$

It follows that

$$\begin{aligned}
\mathbb{I}[p((x, w), (y, z))] &= \int dx dw dy dz \, p(x, y) p(w|x) p(z|w) \log \frac{p((x, w), (y, z))}{p(x, w) p(y, z)} \\
&= \int dx dw dy dz \, p(x, y) p(w|x) p(z|w) \log \frac{p(x, y) p(w|x) p(z|y)}{p(x) p(y) p(w|x) p(z|y)} \\
&= \int dx dw dy dz \, p(x, y) p(w|x) p(z|w) \log \frac{p(x, y)}{p(x) p(y)} \\
&= \int dx dy \, p(x, y) \log \frac{p(x, y)}{p(x) p(y)} = \mathbb{I}[p(x, y)].
\end{aligned}$$

Also,

$$\begin{aligned}
\text{ABA}_k[p((x, w), (y, z))] &= \int \left[\prod_{i=1}^k p(x_i, w_i) dx_i dw_i \right] \int dy dz \, \max_i p(y, z|x_i, w_i). \\
&= \int \left[\prod_{i=1}^k p(x_i, w_i) dx_i dw_i \right] \int dy \, \max_i p(y|x_i) \int dz \, p(z|y). \\
&= \int \left[\prod_{i=1}^k p(x_i) dx_i \right] \left[\prod_{i=1}^k \int dw_i p(w_i|x_i) \right] \int dy \, \max_i p(y|x_i) \\
&= \text{ABA}_k[p(x, y)].
\end{aligned}$$

□

Next, we use the fact that for any $p(x, y)$ and $\epsilon > 0$, there exists a discrete distribution $p_\epsilon(\tilde{x}, \tilde{y})$ such that

$$|\mathbb{I}[p(x, y)] - \mathbb{I}[p_\epsilon(\tilde{x}, \tilde{y})]| < \epsilon,$$

where for discrete distributions, one defines

$$\mathbb{I}[p(x, y)] = \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x) p(y)}.$$

We require the additional condition that the marginals of the discrete distribution are close to uniform: that is, for some $\delta > 0$, we have

$$\sup_{x, x': p_\epsilon(x) > 0 \text{ and } p_\epsilon(x') > 0} \frac{p_\epsilon(x)}{p_\epsilon(x')} \leq 1 + \delta.$$

and likewise

$$\sup_{y, y': p_\epsilon(y) > 0 \text{ and } p_\epsilon(y') > 0} \frac{p_\epsilon(y)}{p_\epsilon(y')} \leq 1 + \delta.$$

To construct the discretization with the required properties, choose a regular rectangular grid Λ over the domain of $p(x, y)$ sufficiently fine so that partitioning X, Y into grid cells, we have

$$|I[p(x, y)] - I[\tilde{p}(\tilde{x}, \tilde{y})]| < \epsilon.$$

[NOTE: to be written more clearly] Next, define