

What does classification tell us about the brain?

Statistical inference through machine learning

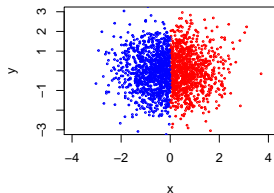
Charles Zheng

Stanford University

October 7, 2016

(Joint work with Yuval Benjamini.)

Dependence, distance, and information



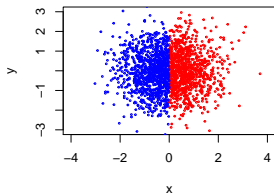
- X is independent of Y :

$$X \perp Y$$

- X and Y have no mutual information:

$$I(X; Y) = 0$$

Dependence, distance, and information



- X is independent of Y :

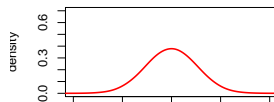
$$X \perp Y$$

- X and Y have no mutual information:

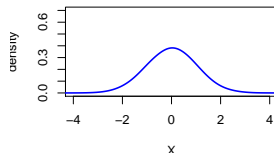
$$I(X; Y) = 0$$

Classifying $\text{Sign}(X)$ from Y

$$Y|X > 0$$



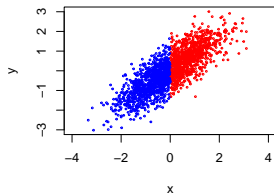
$$Y|X < 0$$



$$KL(Y|X > 0, Y|X < 0) = 0.$$

$$\text{Bayes accuracy} = 0.5.$$

Dependence, distance, and information



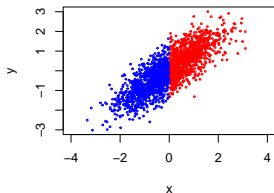
- X is dependent of Y :

$$\text{Cor}(X, Y) = 0.8.$$

- X and Y have mutual information:

$$I(X; Y) = 0.51.$$

Dependence, distance, and information



- X is dependent of Y :

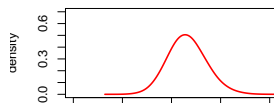
$$\text{Cor}(X, Y) = 0.5.$$

- X and Y have mutual information:

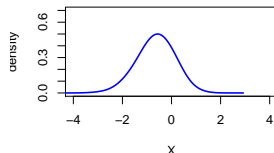
$$I(X; Y) = 0.51.$$

Classifying $\text{Sign}(X)$ from Y

$Y|X > 0$



$Y|X < 0$



$$KL(Y|X > 0, Y|X < 0) = 1.64$$

$$\text{Bayes accuracy} = 0.795.$$

Bayes accuracy

- Discrete $Y \in \{1, \dots, k\}$, continuous or discrete X .
- A classifier is a function f mapping x to a label in $\{1, \dots, k\}$
- Generalization accuracy of the classifier:

$$\text{GA}(f) = \Pr[Y = f(x)]$$

- Bayes accuracy:

$$\text{BA} = \sup_f \Pr[Y = f(x)] = \Pr[Y = \operatorname{argmax}_{i=1} p(X|Y = i)]$$

- Since random guessing is correct with probability $1/k$,

$$\text{BA} \in [1/k, 1]$$

(if Y is uniformly distributed)

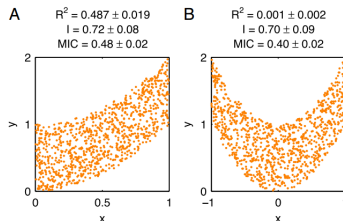
Mutual information

- Invented by Claude Shannon; central to *information theory*.
- Given (X, Y) with joint density $p(x, y)$,

$$I(X; Y) = \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy$$

where $p(x)$ and $p(y)$ are marginal densities.

Mutual information



- $I(X; Y) \in [0, \infty]$. (0 if $X \perp Y$, ∞ if $X = Y$ and X continuous.)
- Symmetry: $I(X; Y) = I(Y; X)$.
- Data-processing inequality

$$I(X; Y) \geq I(\phi(X); \psi(Y))$$

equality for ϕ, ψ bijections

- Additivity. If $(X_1, Y_1) \perp (X_2, Y_2)$, then

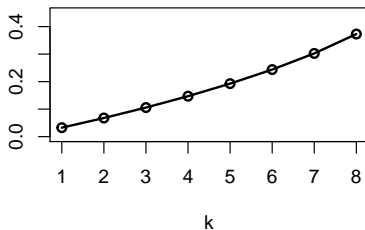
$$I((X_1, X_2); (Y_1, Y_2)) = I(X_1; Y_1) + I(X_2; Y_2).$$

Informativity of predictor sets

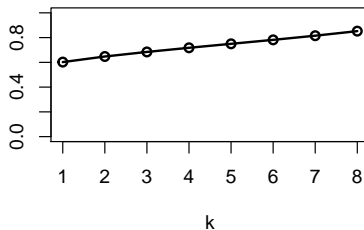
Consider predicting binary Y with:

- X_1 only
- X_1 and X_2
- X_1, \dots, X_k

Mutual information



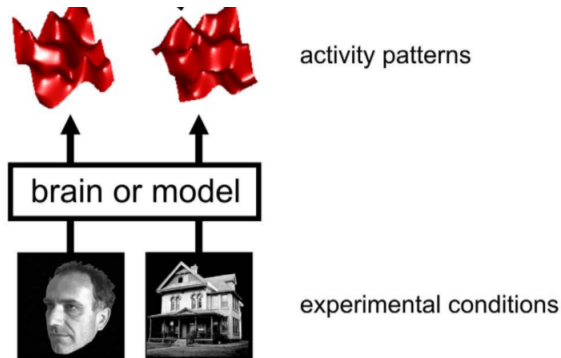
Bayes accuracy



Mutual information vs Bayes accuracy

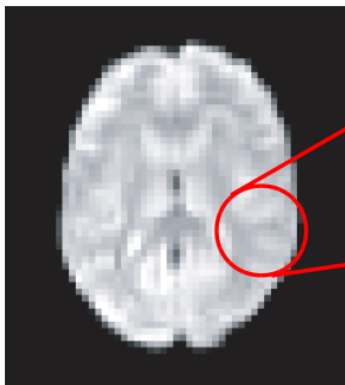
- Both are measures of “informativity”.
- Due to its properties, mutual information is easier to interpret.
- Both are intractable to estimate in high dimensions.
- However, Bayes accuracy has a tractable *lower bound*: the generalization error of *any* classifier.

Studying the neural code

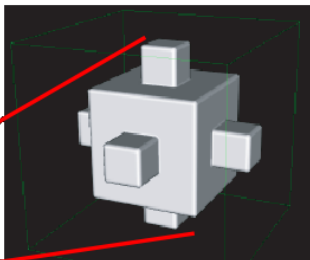


Present the subject with visual stimuli, pictures of faces and houses.
Record the subject's brain activity in the fMRI scanner.

Searchlight analysis



BOLD image

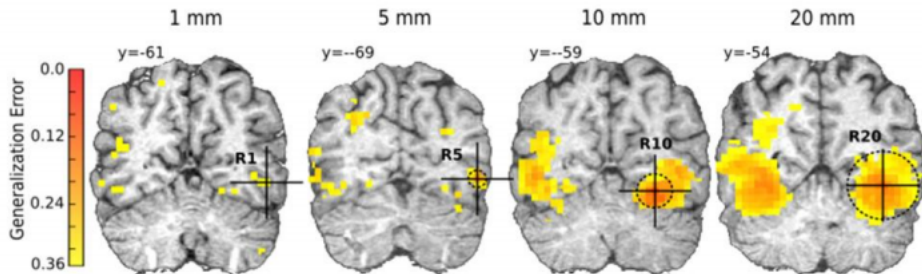


Pull out a local
neighbourhood



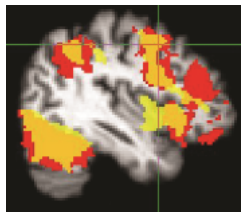
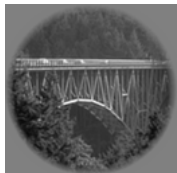
Look at the patterns
in that neighbourhood

Searchlight analysis



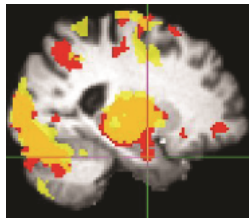
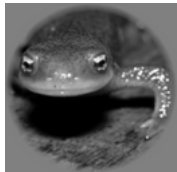
Produces a map of “informative” regions of the brain (as measured by generalization accuracy).

Fixed classification task



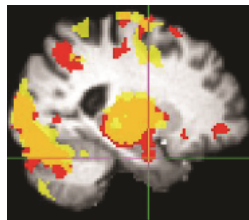
- Experimenter chooses k stimuli.
- Generalization accuracy depends on size of training set, classifier, and choice of stimuli.

Fixed classification task



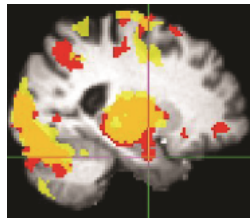
- Different stimuli sets not only lead to different generalization accuracy maps, but also different *Bayes accuracy*.
- Results are incomparable, even in the large-sample limit.

Generalizing beyond the design



Scientists are not innately interested in the Bayes accuracy of a *particular* stimuli set, which is often chosen arbitrarily...

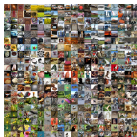
Generalizing beyond the design



But it would be more interesting to be able to make inferences from the data about a *larger* class of stimuli...

Randomized classification

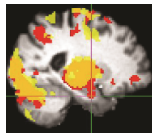
1. Population of stimuli $p(x)$



2. Subsample k stimuli

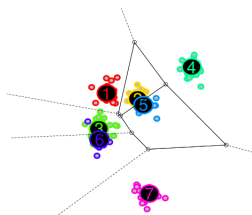


3. Data

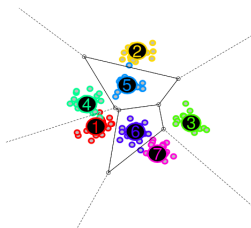


Randomized classification: gaussian example

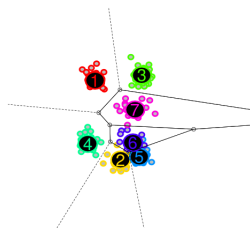
Experiment 1



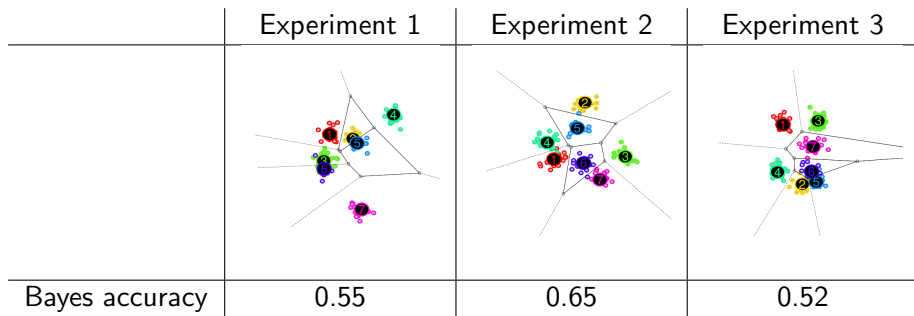
Experiment 2



Experiment 3



Randomized classification: gaussian example



- Bayes accuracy depends on the stimuli drawn.
- Therefore, define *k*-class *average Bayes error* as the expected Bayes error for $X_1, \dots, X_k \stackrel{iid}{\sim} p(x)$.

$$ABA_k = \mathbf{E}[BA(X_1, \dots, X_k)]$$

Analyzing the data using machine learning

- Now we have data consisting of (stimulus, response) pairs.
- Can we classify the response using the stimulus? What is the confusion matrix?

Motivation for my research

Ultimately, the goal of these experiments is to understand the dependence between X (stimulus) and Y (the brain response).

Possible goals for statistical methodology (which currently don't exist):

- 1 What can be inferred from the classification accuracy?
- 2 Can we predict what the result (classification accuracy) would be in a similar (but possibly larger or smaller) experiment?
- 3 Can we *summarize* the total information content contained in Y about X ?
- 4 Can we *decompose* the total information contained in Y about X ? (Something like a nonlinear ANOVA decomposition?)

Motivation 1: What can be inferred from the classification accuracy?

- The achieved classification accuracy is an estimate of *generalization accuracy*...
- which in turn lower bounds on the generalization error of the best classifier, the *Bayes accuracy*.
- But the Bayes accuracy varies depending on the stimuli set $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(K)}$!

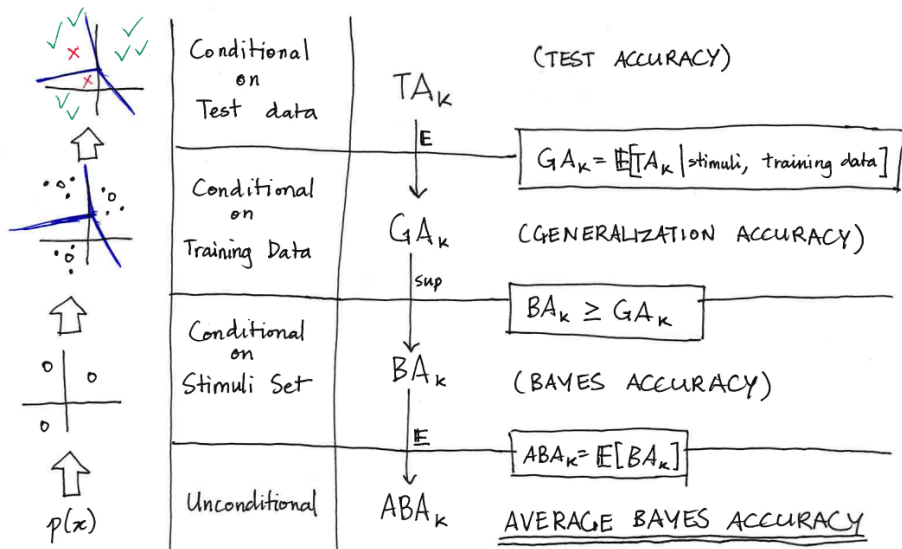
$$\text{BA}(X_1, \dots, X_k)$$

Define *average Bayes accuracy* as the expected Bayes accuracy

$$\text{ABA}_k = \mathbf{E}[\text{BA}(X_1, \dots, X_k)]$$

where expectation is taken over sampling X_1, \dots, X_k from $p(x)$.

Average Bayes accuracy



Inferring average Bayes accuracy

- We cannot observe either ABA_k , or even BA_k .
- However, we can obtain a *lower confidence bound* for BA_k , since the generalization accuracy is an *underestimate* of BA_k
- But we actually want a lower confidence bound for ABA_k !

Concentration of Bayes accuracy

Recall that

$$ABA_k = \mathbf{E}[BA_k]$$

Converting a LCB for BA_k to an LCB on ABA_k boils down to the following problem:

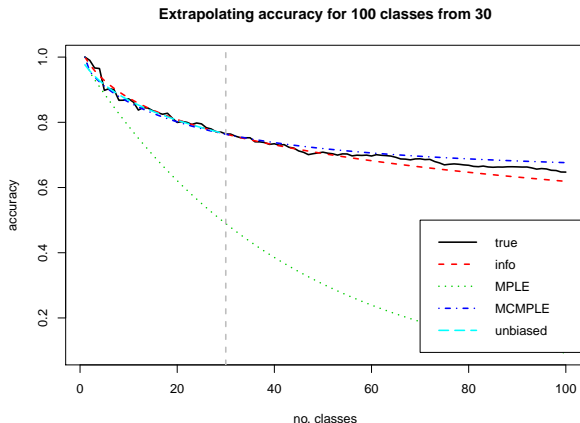
What is the variability of BA_k ?

We will discuss this later in the talk!

Motivation 2: Generalizing to similar designs

Define ABA_k as the average Bayes accuracy for k classes.

Can we predict ABA_{100} given data from 30 classes? See Z., Achanta and Benjamini (2016).



Motivation 3 and 4: Quantifying and decomposing information

Can we summarize the total information content contained in Y about X ?
Can we decompose the total information contained in Y about X ?

The answer is yes, and the solution was provided by Claude Shannon.
Mutual information measures the information contained in Y about X (or vice versa) in a nonlinear way.

Section 2

Variability of Bayes accuracy

Definitions

- Suppose (X, Y) have a joint density $p(x, y)$,
- The Bayes accuracy is a function of the stimuli set x_1, \dots, x_k ,

$$\text{BA}(x_1, \dots, x_k)$$

- Draw $Z \sim \{1, \dots, k\}$, and draw $Y \sim p(y|x_z)$.
- Let f (the *classifier*) that associates a label $\{1, \dots, k\}$ to each possible value of y :

$$f : \mathcal{Y} \rightarrow \{1, \dots, k\}$$

- Define

$$\text{BA}(x_1, \dots, x_k) = \sup_f \Pr[f(Y) = Z | x_1, \dots, x_k]$$

where the probability is over the joint distribution (Y, Z) defined above. Notice we condition on the particular stimuli set x_1, \dots, x_k .

An identity

- It is a well-known result from Bayesian inference that the optimal classifier f is defined as

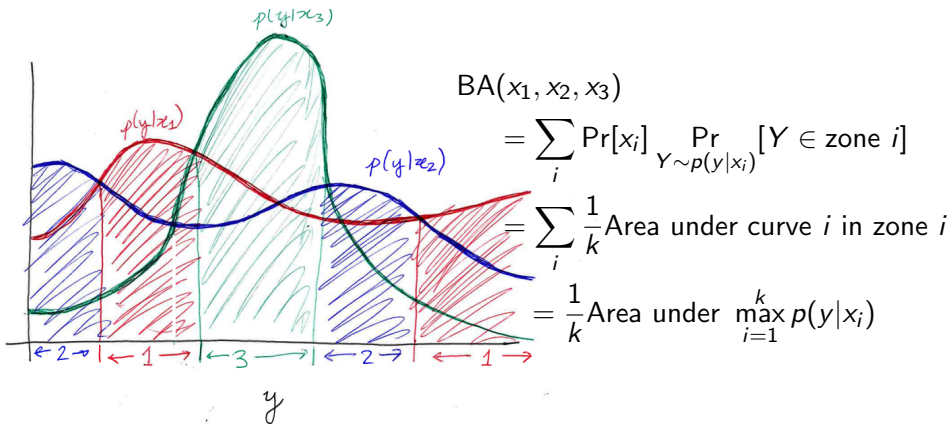
$$f(y) = \operatorname{argmax}_{i=1}^k p(y|x_i),$$

since the prior class probabilities are uniform.

- Therefore,

$$\begin{aligned} \text{BA}(x_1, \dots, x_k) &= \Pr[\operatorname{argmax}_{i=1}^k p(y|x_i) = Z | x_1, \dots, x_k] \\ &= \frac{1}{k} \int \max_{i=1}^k p(y|x_i) dy. \end{aligned}$$

Intuition behind identity



Efron-Stein lemma

- We have

$$\text{ABA}_k = \mathbf{E}[\text{BA}(X_1, \dots, X_k)]$$

where the expectation is over the independent sampling of X_1, \dots, X_k from $p(x)$.

- According to the Efron-Stein lemma,

$$\text{Var}[\text{BA}(X_1, \dots, X_k)] \leq \sum_{i=1}^k \mathbf{E}[\text{Var}[\text{BA}|X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_k]].$$

which is the same as

$$\text{Var}[\text{BA}(X_1, \dots, X_k)] \leq k \mathbf{E}[\text{Var}[\text{BA}|X_1, \dots, X_{k-1}]].$$

- The term $\text{Var}[\text{BA}|X_1, \dots, X_{k-1}]$ is the variance of $\text{BA}(X_1, \dots, X_k)$ conditional on fixing the first $k - 1$ curves $p(y|x_1), \dots, p(y|x_{k-1})$ and allowing the final curve $p(y|X_k)$ to vary randomly.

Efron-Stein lemma



$$\text{Var}[\text{BA}(X_1, \dots, X_k)] \leq k \mathbf{E}[\text{Var}[\text{BA}|X_1, \dots, X_{k-1}]].$$

- Note the following trivial results

$$-p(y|x_k) + \max_{i=1}^k p(y|x_i) \leq \max_{i=1}^{k-1} p(y|x_i) \leq \max_{i=1}^k p(y|x_i)$$

- This implies

$$\text{BA}(X_1, \dots, X_k) - \frac{1}{k} \leq \frac{k-1}{k} \text{BA}(X_1, \dots, X_{k-1}) \leq \text{BA}(X_1, \dots, X_k).$$

i.e. conditional on (X_1, \dots, X_{k-1}) , BA_k is supported on an interval of size $1/k$.

- Therefore,

$$\text{Var}[\text{BA}|X_1, \dots, X_{k-1}] \leq \frac{1}{4k^2}$$

since $\frac{1}{4c^2}$ is the maximal variance for any r.v. with support of length c .

Variance bound

Therefore, Efron-Stein bound gives

$$\text{sd}[\text{BA}_k] \leq \frac{1}{2\sqrt{k}}$$

Compare this with empirical results (searching for worst-case distributions):

k	2	3	4	5	6	7	8
$\frac{1}{2\sqrt{k}}$	0.353	0.289	0.250	0.223	0.204	0.189	0.177
Worst-case sd	0.25	0.194	0.167	0.150	0.136	0.126	0.118

Improving the variance bound?

- All of the worst-case distributions found so far have the following simple form:

$$\mathcal{Y} = \mathcal{X} = \{1, \dots, d\} \text{ for some } d$$

$$p(y|x) = \frac{1}{d} I\{x = y\}$$

Can we prove this rigorously?

- Recalling that

$$\text{BA}(X_1, \dots, X_k) - \frac{1}{k} \leq \frac{k-1}{k} \text{BA}(X_1, \dots, X_{k-1}) \leq \text{BA}(X_1, \dots, X_k).$$

it is worth noting that distributions of this type actually concentrate on the two endpoints of the bound, thus in some sense “maximizing” the variance.

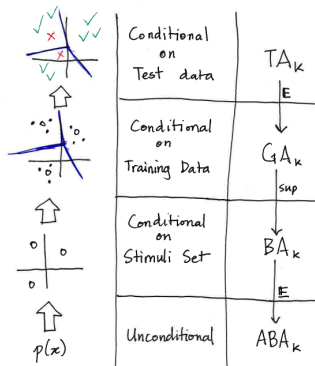
Section 3

Inferring mutual information from classification accuracy

- We observe (X, Y) pairs from the random-stimulus repeated-sampling design.
- Goal is to infer $I(X; Y)$, also written $I[p(x, y)]$.

Outline

- Step 1: Apply machine learning to obtain *test accuracy* TA_k
- Step 2: Infer ABA_k from TA_k
- Step 3: Obtain a lower bound on $I(X; Y)$ from ABA_k !



We already know how to do steps 1 and 2; now we discuss step 3.

Comparison of ABA and I

Average Bayes accuracy $\text{ABA}_k[p(x, y)]$ and mutual information $I[p(x, y)]$ are both *functionals* of $p(x, y)$.

$$\text{ABA}_k[p(x, y)] = \frac{1}{k} \int p_X(x_1) \dots p_X(x_k) \max_{i=1}^k p(y|x_i) dx_1 \dots dx_k dy.$$

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

Natural questions

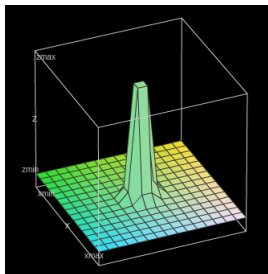
- Does ABA_k close to 1 imply I large?
- Does ABA_k close to $1/k$ imply I close to 0?
- Does I large imply ABA_k close to 1?
- Does I close to 0 imply ABA_k close to $1/k$?

Does I close to 0 imply ABA_k close to $1/k$?

Answer is yes, since $I[p(x, y)] = 0$ implies that X is independent of Y . And when $X \perp Y$, the best classifier does not better than random guessing.

Does I large imply ABA_k close to 1?

Answer is **no**... per the following counterexample.



$$X \in [0, 1], \quad Y \in [0, 1]$$

$$p(x, y) \propto (1 - \alpha) + \alpha \left(\frac{e^{-\frac{x^2 + y^2}{2\sigma^2}}}{2\pi\sigma^2} \right)$$

$$I[p(x, y)] \approx \alpha \left(\frac{1}{2} \log \frac{1}{\sigma^2} - 1 - \log(2\pi) \right)$$

Taking $\alpha \rightarrow 0$ and $\sigma^2 \leq e^{-\frac{1}{\alpha^2}}$, we get

$$I[p(x, y)] \rightarrow \infty, \quad ABA_k[p(x, y)] \rightarrow \frac{1}{k}.$$

This also answers “Does ABA_k close to $1/k$ imply I close to 0?” (Also no.)

Natural questions

- Does ABA_k close to $1/k$ imply I close to 0? **No.** (counterexample)
- Does I large imply ABA_k close to 1? **No.** (counterexample)
- Does I close to 0 imply ABA_k close to $1/k$? **Yes.**

The only remaining question is:

Does ABA_k close to 1 imply I large?

The answer is yes and provides the desired lower bound. In fact,

$$ABA_k \rightarrow 1$$

implies

$$I[p(x, y)] \rightarrow \infty.$$

Problem formulation

Take $\iota > 0$, and fix $k \in \{2, 3, \dots\}$. Let $p(x, y)$ be a joint density (where (X, Y) could be random vectors of any dimensionality.) Supposing

$$I[p(x, y)] \leq \iota,$$

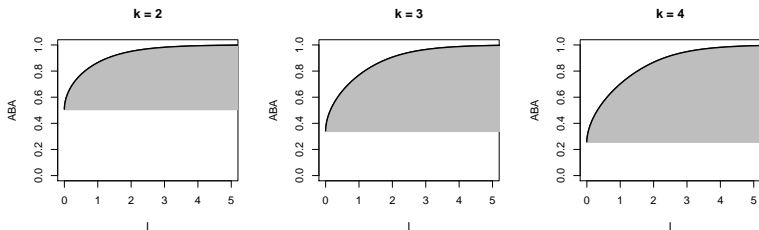
then can we find an upper bound on $ABA_k[p(x, y)]$?

In other words, can we compute the value of

$$C_k(\iota) = \sup_{p(x, y): I[p(x, y)] < \iota} ABA_k[p(x, y)]?$$

Preview

Yes we can, and this is what the resulting function $C_k(\iota)$ looks like:



As information increases, the maximal average Bayes accuracy goes to 1. We find these curves using *variational calculus*.

Reduced Problem

Consider a special class of densities with the following properties:

- $p(x, y)$ is on the unit square
- The marginal distributions of x and y are both uniform
- The conditional distributions $p(x|y)$ are just “shifted” copies of a common density, $q(x)$, on $[0, 1]$

$$p(x|y) = q(x - y + I\{x < y\})$$

- Furthermore, $q(x)$ is increasing in x .

We claim (but will not show) that this special class of densities achieves the maximal average Bayes error. Therefore, it suffices to search over this special class of densities to compute $C_k(\iota)$.

The information and average Bayes error can be written in terms of $q(x)$.

$$I[p(x, y)] = \int_0^1 q(x) \log q(x) dx$$

$$\text{ABA}_k[p(x, y)] = \int_{[0,1]^k} \max_{i=1}^k q(x_i) dx_1 \cdots dx_k$$

Overload the notation and “redefine” information and average Bayes error as functionals of $q(x)$.

$$I[q(x)] \stackrel{\text{def}}{=} \int_0^1 q(x) \log q(x) dx$$
$$\text{ABA}_k[q(x)] \stackrel{\text{def}}{=} \frac{1}{k} \int_{[0,1]^k} \max_{i=1}^k q(x_i) dx_1 \cdots dx_k$$

Simplified formulae

We can simplify the expression for ABA_k even more.

Observe that since $q(x)$ is increasing,

$$\max_{i=1}^k q(x_i) = q\left(\max_{i=1}^k x_i\right)$$

Therefore,

$$\begin{aligned}\text{ABA}_k[q(x)] &= k^{-1} \int_{[0,1]^k} \max_{i=1}^k q(x_i) dx_1 \cdots dx_k \\ &= k^{-1} \int_{[0,1]^k} q\left(\max_{i=1}^k x_i\right) dx_1 \cdots dx_k \\ &= k^{-1} \mathbf{E}\left[q\left(\max_{i=1}^k X_i\right)\right] = k^{-1} \mathbf{E}[q(M)]\end{aligned}$$

where $X_1, \dots, X_k \stackrel{iid}{\sim} \text{Unif}[0, 1]$ and $M = \max_{i=1}^k X_i$.

Recall that the max of k iid uniforms has density

$$f(m) = km^{k-1}.$$

Therefore,

$$\text{ABA}_k[q(x)] = k^{-1} \mathbf{E}[q(M)] = \int_0^1 q(t) t^{k-1} dt.$$

Optimization problem

We now pose the question: how do we find $q(x)$ which maximizes $\text{ABA}_k[q(x)]$ subject to $\mathcal{I}[q(x)] \leq \iota$?

- *Domain of the optimization:* Recall that $q(x)$ satisfies $q(x) \geq 0$, $\int_0^1 q(x)dx = 1$, and is increasing in x . Let \mathcal{Q} denote the space of functions on $[0, 1] \rightarrow [0, \infty)$ which are increasing in x .
- *Constraints:* We have two remaining constraints, $\mathcal{I}[q(x)] \leq \iota$ and $\int_0^1 q(x)dx = 1$.

Hence the problem is

$$\text{maximize}_{q(x) \in \mathcal{Q}} \text{ABA}_k[q(x)] \text{ subject to } \int_0^1 q(x)dx = 1 \text{ and } \mathcal{I}[q(x)] \leq \iota.$$

Optimization problem

maximize $_{q(x) \in \mathcal{Q}}$ $\text{ABA}_k[q(x)]$ subject to $\int_0^1 q(x)dx = 1$ and $I[q(x)] \leq \iota$.

- Does a solution exist? Yes, because the space of measures with density $q(x)$ satisfying $I[q(x)] \leq \iota$ is tight, and both the constraints and objective are continuous wrt to the topology of weak convergence.
- Given a solution $q^*(x)$ exists, there exist Lagrange multipliers $\lambda \in \mathbb{R}$ and $\nu > 0$ such that q^* minimizes

$$\begin{aligned}\mathcal{L}[q(x)] &= -\text{ABA}_k[q(x)] + \lambda \int_0^1 q(x)dx + \nu I[q(x)] \\ &= \int_0^1 (-t^{k-1} + \lambda + \nu \log q(x))q(x)dx.\end{aligned}$$

Functional derivatives

- Functional derivatives are essential to variational calculus.
- Let \mathcal{F} be a *Hilbert space* of functions with domain \mathcal{X} and range \mathbb{R} .
- Suppose F is a functional which maps functions f to the real line. Then the functional derivative $\nabla F[f]$ at f is a function in the space \mathcal{F} such that

$$\lim_{\epsilon \rightarrow 0} \frac{F(f + \epsilon \xi) - F(f)}{\epsilon} = \int_{\mathcal{X}} \nabla F[f](x) \xi(x) dx.$$

for all $\xi \in \mathcal{F}$.

Functional derivatives

- Taylor expansions are a useful trick for computing functional derivatives
- We can compute the functional derivative of $\mathcal{L}[q(x)]$ by writing

$$\begin{aligned}\mathcal{L}[q(x) + \epsilon \xi(x)] &= \int_0^1 (-t^{k-1} + \lambda + \nu \log(q(x) + \epsilon \xi(x)))(q(x) + \epsilon \xi(x)) dx. \\ &\approx \int (q(x) + \epsilon \xi(x))(-t^{k-1} + \lambda + \nu \{\log q(x) + \frac{\epsilon \xi(x)}{q(x)}\}) dx \\ &\approx \mathcal{L}[q(x)] + \int_0^1 (-t^{k-1} + \lambda + \nu(1 + \log q(x))) \epsilon \xi(x) dx.\end{aligned}$$

- Hence

$$\nabla \mathcal{L}[q](x) = -t^{k-1} + \lambda + \nu(1 + \log q(x))$$

Variational magic!

Suppose we set the functional derivative to 0,

$$0 = \nabla \mathcal{L}[q](t) = -t^{k-1} + \lambda + \nu + \nu \log q(t).$$

Then we conclude that the optimal $q^*(t)$ takes the form

$$q^*(t) = \alpha e^{\beta t^{k-1}}$$

for some $\alpha > 0$, $\beta > 0$.

From the constraint $\int q(t) dt = 1$, we get

$$q_\beta(t) = \frac{e^{\beta t^{k-1}}}{\int e^{\beta t^{k-1}} dt}.$$

For the optimal $q(t)$, how do we know $\nabla \mathcal{L}[q](t) = 0$?

- Since \mathcal{Q} has a monotonicity constraint, we cannot simply take for granted that

$$\nabla \mathcal{L}[q^*](t) = 0$$

- However, we can show that assuming

$$\nabla \mathcal{L}[q^*](t) \neq 0$$

on a set of positive measure results in a contradiction.

- The contradiction is achieved by constructing a suitable perturbation ξ which is “localized” around a region where $\mathcal{L}[q^*](t) \neq 0$, such that $q^* + \epsilon \xi \in \mathcal{Q}$ and also so that $\int \xi(t) \nabla \mathcal{L}[q^*](t) dt < 0$. This implies that for ϵ sufficiently small, $\mathcal{L}[q^* + \epsilon \xi] < \mathcal{L}[q^*]$ —a contradiction, since we assumed that q^* was optimal.

Theorem. For any $\iota > 0$, there exists $\beta_\iota \geq 0$ such that defining

$$q_\beta(t) = \frac{\exp[\beta t^{k-1}]}{\int_0^1 \exp[\beta t^{k-1}]},$$

we have

$$\int_0^1 q_{\beta_\iota}(t) \log q_{\beta_\iota}(t) dt = \iota.$$

Then,

$$C_k(\iota) = \int_0^1 q_{\beta_\iota}(t) t^{k-1} dt.$$