What does classification tell us about the brain? Statistical inference through machine learning

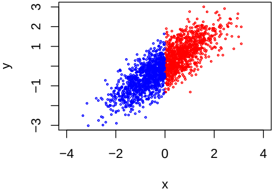
Charles Zheng

Stanford University

October 10, 2016

(Joint work with Yuval Benjamini.)

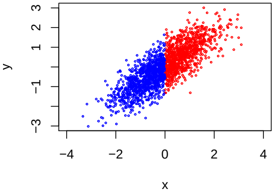
Dependence, information



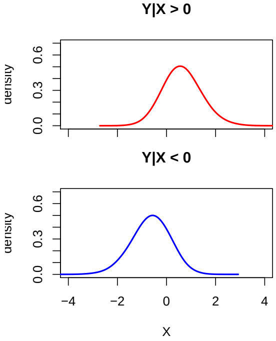
- X is dependent of Y.
- X and Y have mutual information:

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

Dependence, information, and classification



Classifying Sign(X) from Y



- X is dependent of Y.
- X and Y have mutual information:

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

Bayes accuracy $= 0.795$.

Bayes accuracy

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

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

- Bayes accuracy:

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

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

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

(if Y is uniformly distributed)

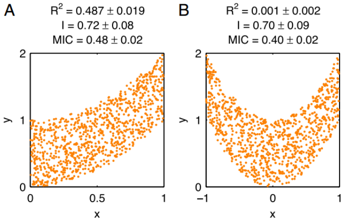
Mutual information

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

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

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

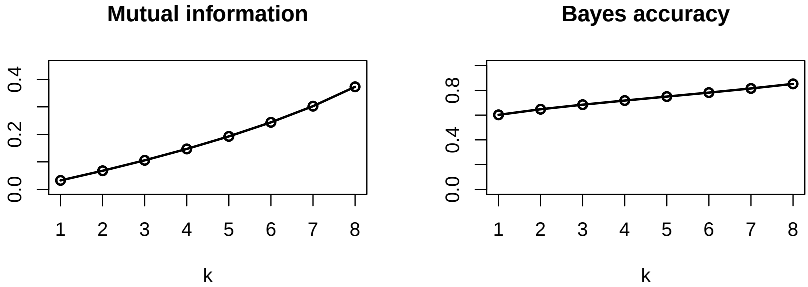
  $$I((X_1, X_2);(Y_1, Y2)) = I(X_1;Y_1) + I(X_2;Y_2).$$

Image credit Kinney et al. 2014.

Informativity of predictor sets

Consider predicting binary Y with:

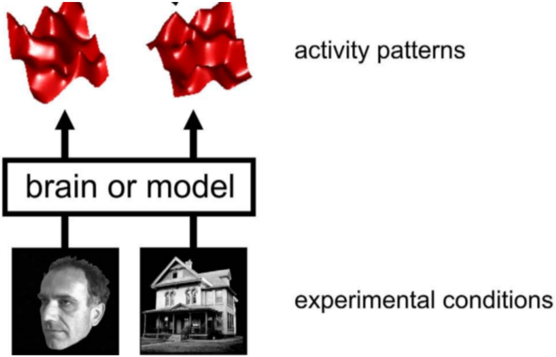
- X_1 only
- X_1 and X_2
- $X_1, ..., X_k$



Mutual information vs Bayes accuracy

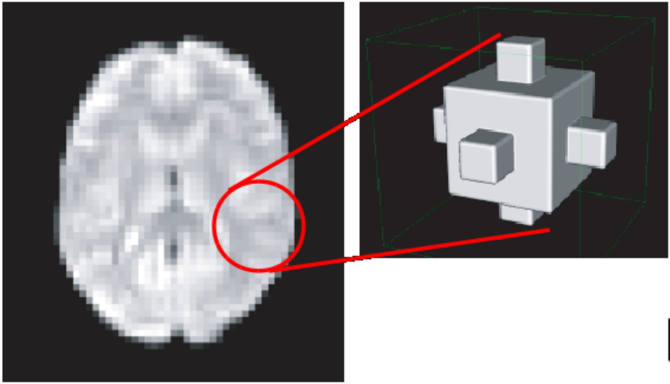
- Both are measures of “informativity”.
- Due to its properties, mutual information is easier to interpret.
- Both are intractable to estimate in high dimensions.
- However, Bayes accuracy has a tractable *lower bound*: the generalization error of *any* classifier.

Studying the neural code



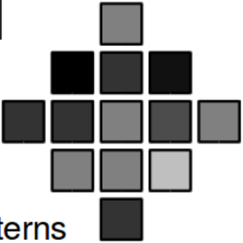
Present the subject with visual stimuli, pictures of faces and houses. Record the subject's brain activity in the fMRI scanner.

Searchlight analysis



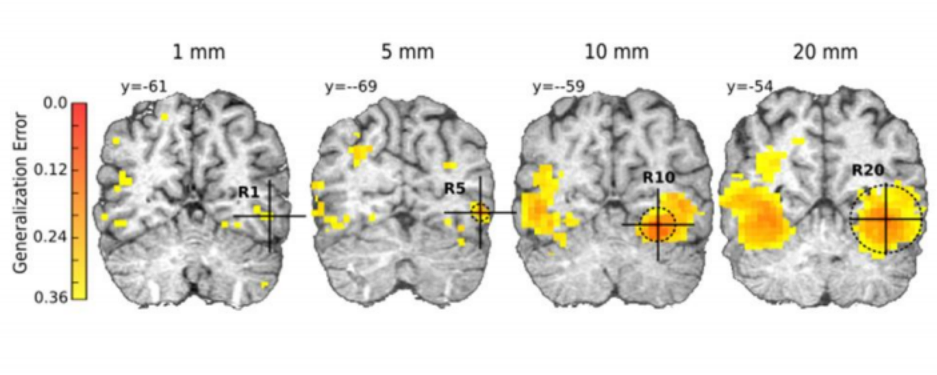
BOLD image

Pull out a local neighbourhood



Look at the patterns in that neighbourhood

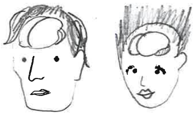
Searchlight analysis



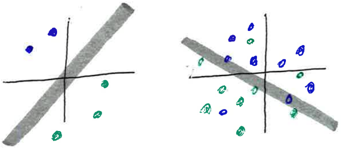
Produces a map of “informative” regions of the brain (as measured by generalization accuracy).

ISSUES W/ TEST ACCURACY

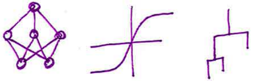
1. Subject dependence



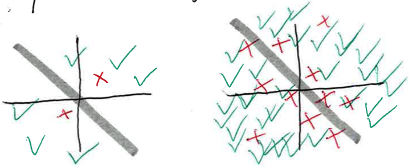
2. Dependence on Training Data



3. Dependence on Classifier

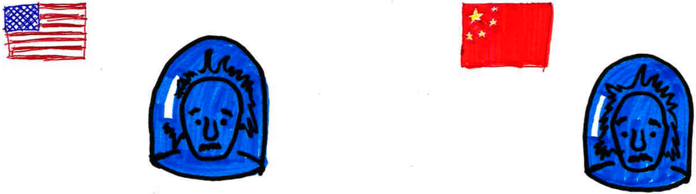


4. Variability due to finite Test Data

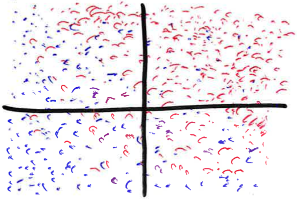


IDEAL WORLD

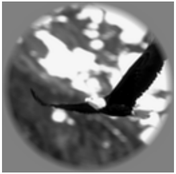
1. Every lab owns a clone of Einstein

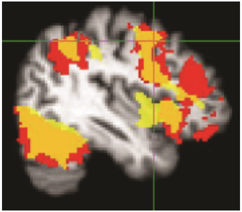


2. Infinite training & test data (⇒ we can obtain Bayes accuracy)



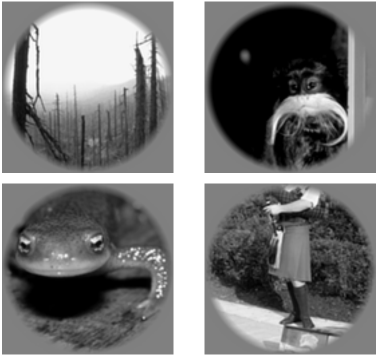
Fixed classification task

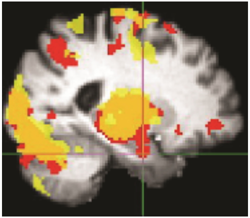




- Different stimuli sets lead to different *Bayes accuracy*.

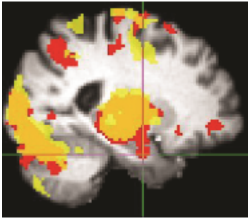
Fixed classification task





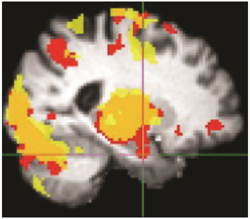
- Different stimuli sets lead to different *Bayes accuracy*.
- Results are incomparable, even in the large-sample limit.

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...

Section 2

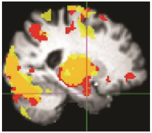
Randomized classification and Average Bayes accuracy

Randomized classification

1. Population of stimuli $p(x)$

2. Subsample k stimuli

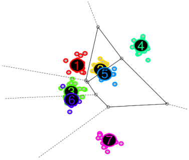
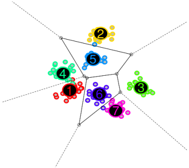
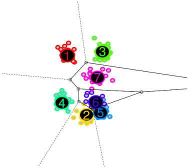
3. Data



4. Train a classifier

5. Estimate generalization accuracy (which is lower bound for the *random* Bayes accuracy BA_k)

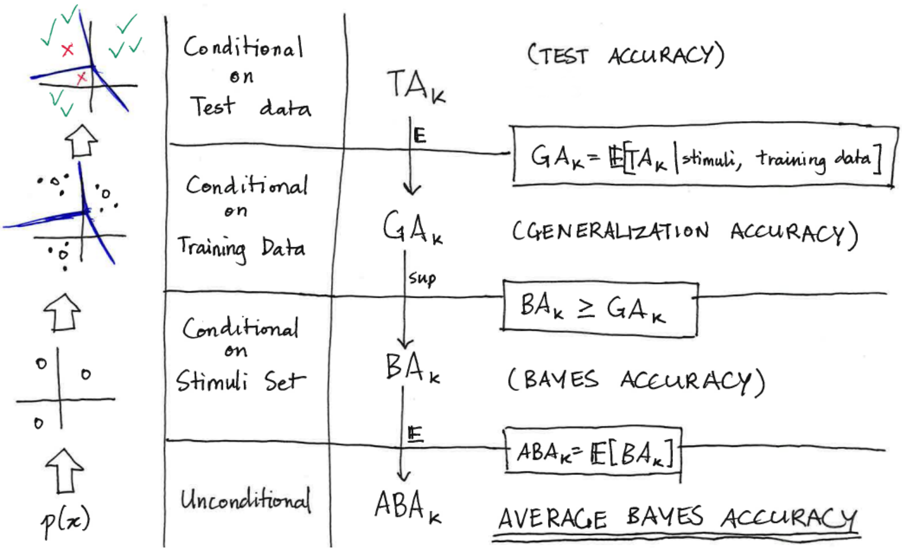
Average Bayes error

	Experiment 1	Experiment 2	Experiment 3
			
Bayes accuracy	0.55	0.65	0.52

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

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

Average Bayes accuracy



Conditional on Test data | TA_K | (TEST ACCURACY)

$\downarrow \mathbb{E}$

$GA_K = \mathbb{E}[TA_K \mid \text{stimuli, training data}]$

Conditional on Training Data | GA_K | (GENERALIZATION ACCURACY)

$\downarrow \sup$

$BA_K \geq GA_K$

Conditional on Stimuli Set | BA_K | (BAYES ACCURACY)

$\downarrow \mathbb{E}$

$ABA_K = \mathbb{E}[BA_K]$

Unconditional | ABA_K | AVERAGE BAYES ACCURACY

Inferring average Bayes accuracy

- $\text{BA}_k \overset{def}{=} \text{BA}(X_1, .., X_k)$ is a good proxy for ABA_k, if

  $$\text{Var}[\text{BA}(X_1, ..., X_k)]$$

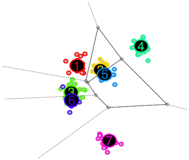
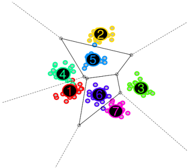
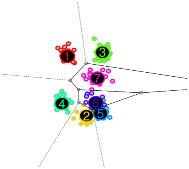
  is small..
- *Theoretical result*. Maximal variability is of order $1/k$.
- Therefore, it is feasbile to get a good idea of ABA_k by choosing a sufficiently large sample size k.

Two intuitions for variability result

Why does variability decrease with k?

- 1. Bayes accuracy behaves like an average of k i.i.d random variables. (Also gives correct $1/k$ rate.)
- 2. Bayes accuracy behaves like a max of k i.i.d. random variables.

Intuition 1: averaging

	Experiment 1	Experiment 2	Experiment 3
			
Bayes accuracy	0.55	0.65	0.52

Average of k gaussian probability integrals... (which are asympt. uncorrelated.)

Intuition 2: An identity

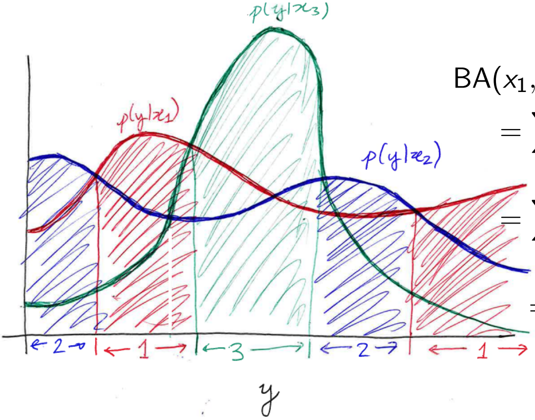
- It is a well-known result from Bayesian inference that the optimal classifier f is defined as

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

  since the prior class probabilities are uniform.
- Therefore,

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

Intuition behind identity


$p(y|x_3)$
$p(y|x_1)$
$p(y|x_2)$
$\leftarrow 2 \rightarrow$
$\leftarrow 1 \rightarrow$
$\leftarrow 3 \rightarrow$
$\leftarrow 2 \rightarrow$
$\leftarrow 1 \rightarrow$
y

$$\begin{aligned}
&\text{BA}(x_1, x_2, x_3)\\
&\quad = \sum_i \Pr[x_i] \Pr_{Y \sim p(y|x_i)}[Y \in \text{zone } i]\\
&\quad = \sum_i \frac{1}{k} \text{Area under curve } i \text{ in zone } i\\
&\quad = \frac{1}{k} \text{Area under } \max_{i=1}^{k} p(y|x_i)
\end{aligned}$$

Variability of Bayes accuracy

Theoretical result. In the max formulation of BA_k, we can apply Efron-Stein inequality to get

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

Empirical results. (searching for worst-case stimuli).

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 take the form

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

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

- Sampling k items from d with replacement; BA_k is the number of unique items divided by k.
- According to Birthday paradox,

$$\text{ABA}_k \approx (1 - e^{-d/k})$$

and

$$\text{Var}(\text{BA}_k) \approx \frac{1}{d} e^{-d/k} (1 - e^{-d/k})$$

- "Discreteness" of the distribution seems to maximize variance?
- If we could prove that this is indeed the worst case, then we have a better constant for variance bound.

Inferring average Bayes error

For now, return to the world of finite data...

1. *Experimental design*: draw k stimuli $X_1, ..., X_k$ iid from $p(x)$. Then collect data (X_i, Y_i^j).
2. *Supervised learning*: train a classifier and obtain a test accuracy TA_k.
3. *Generalization accuracy*: if n_{test} is the size of the test set,

   $$\underline{\mathrm{GA}}_k = \mathrm{TA}_k - \frac{z_{\alpha/2}\sqrt{\mathrm{TA}_k(1 - \mathrm{TA}_k)}}{\sqrt{n_{test}}}$$

   is a lower confidence bound for GA_k
4. *Bayes accuracy*:

   $$\underline{\mathrm{BA}}_k = \underline{\mathrm{GA}}_k$$

   is a lower confidence bound for BA_k
5. *Average Bayes accuracy*

   $$\underline{\mathrm{ABA}}_k = \underline{\mathrm{BA}}_k - \frac{1}{2\sqrt{\alpha k}}$$

   is a lower confidence bound for ABA_k.

Section 3

Relationship between mutual information and average Bayes accuracy

Two measures of informativity: ABA and mutual information

Both are:

- measures of informativity between X and Y
- invariant to bijective transformations of either X or Y
- defined with reference to a *population* of stimuli and either a single subject or population of subjects

Comparison of ABA and mutual information

ABA_k advantages:

- intuitive to understand "classification performance".
- easy to average over a *population* of subjects.
- closer to what you can measure: (generalization accuracy).

ABA_k disadvantages:

- Not symmetric with respect to X and Y. Have to choose one as predictor and one as response.
- Dependent on k, the number of classes.
- Problem of *saturation*. If k is too large, ABA_k gets close to chance accuracy. If k is too large, ABA_k gets close to 1.

Comparison of ABA and mutual information

Mutual information advantages:

- already has a tradition of usage in neuroscience.
- symmetric between X and Y: X is equally informative of Y as Y is of X.
- doesn't depend on k, the number of stimuli.
- additional theoretical properties like independent additivity.

Mutual information disadvantages:

- not robust: $I(X;Y)$ becomes unbounded if $p(x,y)$ contains singularities.
- does it make sense to take the average mutual information across subjects?

Question

Given that mutual information and average Bayes error have complementary advantages and disadvantages, can we “convert” one to the other?

Related work

- Classically, *Fano's inequality* obtains a lower bound for mutual information from *Bayes accuracy*. (We do the same, but for *average Bayes error*).
- Treves (1997) proposes using the *confusion matrix* obtained from classification to estimate mutual information. This has been a popular approach; see Quiroga (2009).
- Gastpar et al (2010) develop *nonparametric* estimators of mutual information for the randomized classification setup (but does not involve using supervised learning.)

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 $1/k$?

Functional formulation

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

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

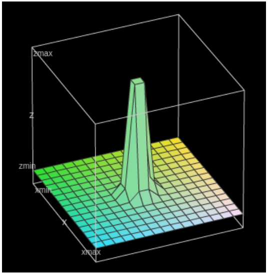
$$\mathrm{I}[p(x,y)] = \int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy.$$

Does I close to 0 imply ABA$_k$ close to $1/k$?

Answer is yes, since $\mathrm{I}[p(x,y)] = 0$ 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.



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

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

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

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

$$\mathrm{I}[p(x,y)] \to \infty, \;\; \mathrm{ABA}_k[p(x,y)] \to \frac{1}{k}.$$

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 an “extension” of Fano’s inequality. Unlike in Fano’s inequality,

$$\mathrm{ABA}_k \to 1$$

implies

$$\mathrm{I}[p(x, y)] \to \infty.$$

Problem formulation

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

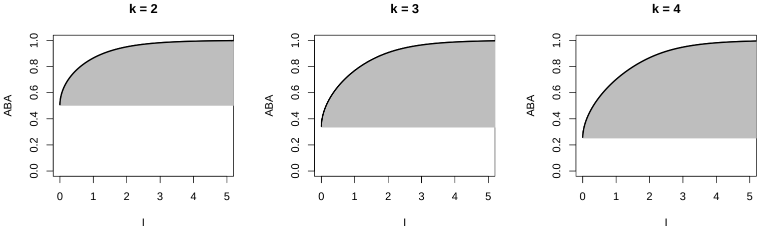
$$\mathrm{I}[p(x, y)] \leq \iota,$$

then can we find an upper bound on $\mathrm{ABA}_k[p(x, y)]$?
In other words, can we compute the value of

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

Preview

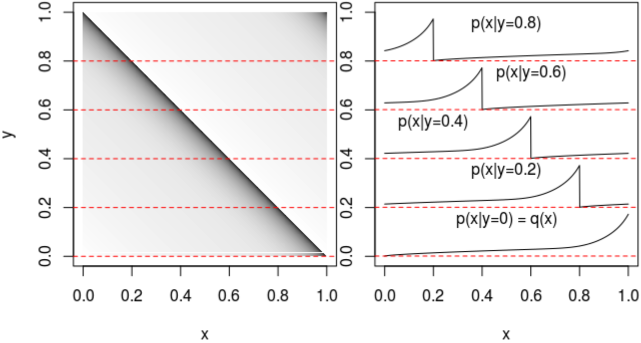
Yes we can, and this is what the resulting function $C_k(\iota)$ looks like:



As information increases, the maximal average Bayes accuracy goes to 1.

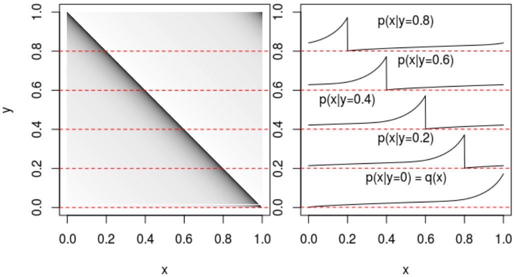
Reduced Problem

Rather than show the whole proof, we consider a simplified problem to illustrate the methods.



Actually, the simplified problem is equivalent to the full problem and we get the same answer (but this is non-trivial).

Reduced Problem



- $p(x, y)$ on unit square with uniform marginals.
- 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.

Simplified formulae

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

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

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

Simplified formulae

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

$$\mathrm{I}[q(x)] \overset{def}{=} \int_0^1 q(x) \log q(x) dx$$

$$\mathrm{ABA}_k[q(x)] \overset{def}{=} \frac{1}{k} \int_{[0,1]^k} \max_{i=1}^{k} q(x_i) dx_1 \cdots dx_k$$

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

Simplified formulae

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 $\mathrm{ABA}_k[q(x)]$ subject to $\mathrm{I}[q(x)] \leq \iota$?

- *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 $[0,1] \to [0,\infty)$ which are increasing in x.
- *Constraints*: We have two remaining constraints, $\mathrm{I}[q(x)] \leq \iota$ and $\int_0^1 q(x)dx = 1$.

Hence the problem is

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

Optimization problem

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

- Does a solution exist? *Yes*, because the space of measures with density $q(x)$ satisfying $\text{I}[q(x)] \leq \iota$ 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

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

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

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

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

Functional derivatives

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

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

- 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

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

Technical sidenote

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.

Result

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

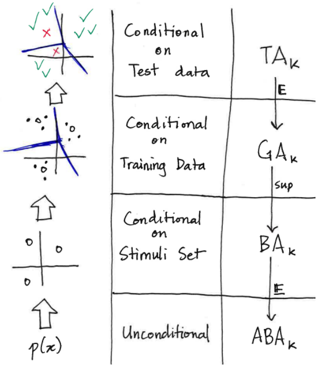
$$\int_0^1 q_{\beta_\iota}(t) \log q_{\beta_\iota}(t) dt = \iota.$$

Then,

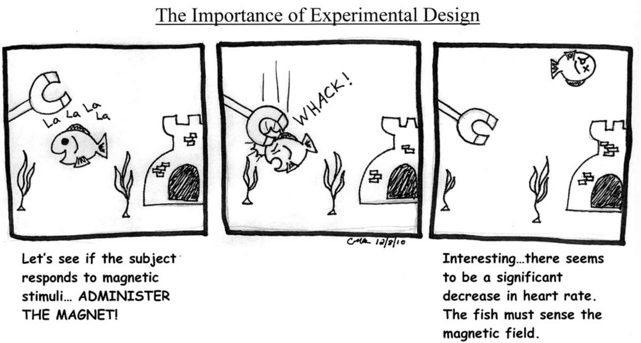
$$C_k(\iota) = \int_0^1 q_{\beta_\iota}(t) t^{k-1} dt.$$

Conclusion: Inferring mutual information from randomized classification

- Step 1: Apply machine learning to obtain *test accuracy* TA_k.
- Step 2: Obtain lower confidence bound $\underline{\text{ABA}_k}$.
- Step 3: Obtain a lower confidence bound on $I(X;Y)$ from $\underline{\text{ABA}_k}$.



The end


The Importance of Experimental Design
La La La La
Let's see if the subject responds to magnetic stimuli... ADMINISTER THE MAGNET!
WHACK!
CMA 12/8/10
Interesting...there seems to be a significant decrease in heart rate. The fish must sense the magnetic field.

(credit C. Ambrosino)

Fun fact: “variational” proof of Fano’s inequality

$X \sim \text{Unif}\{1, ..., k\}$, $Y \sim \text{Unif}[0, 1]$.

$$\text{I}(X; Y) = \frac{1}{k} \sum_x \int p(y|x) \log p(y|x) dy,$$

$$\text{BA} = \frac{1}{k} \int \max_x p(y|x) dy.$$

reduces to

$$\text{maximize}_{q_i \geq 0} \ \max_{i=1}^k q_i$$

$$\text{s.t. } \sum_{i=1}^k q_i = 1 \text{ and } \log(k) + \sum_{i=1}^k q_i \log q_i \leq \iota.$$

Fun fact: "variational" proof of Fano's inequality

Optimum takes the form

$$q_1 = \beta,\ q_2 = \cdots = q_k = (1-\beta)/(k-1).$$

where $\mathrm{BA} = \beta$. Hence,

$$\begin{aligned} \mathrm{I}(X;Y) \leq \iota &= \log(k) + \beta\log(\beta) + (1-\beta)\log((1-\beta)/(k-1)) \\ &= \log(k) - H(\mathrm{BA}) - (1-\mathrm{BA})\log(k-1), \end{aligned}$$

which is Fano's inequality.