

# What does classification tell us about the brain?

## Statistical inference through machine learning

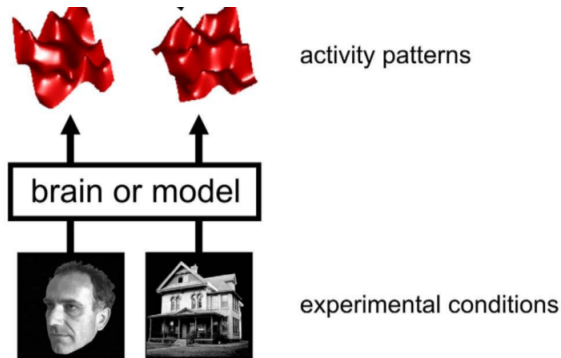
Charles Zheng

Stanford University

October 4, 2016

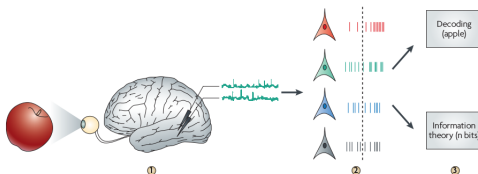
(Joint work with Yuval Benjamini.)

# 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.

# Studying the neural code: data

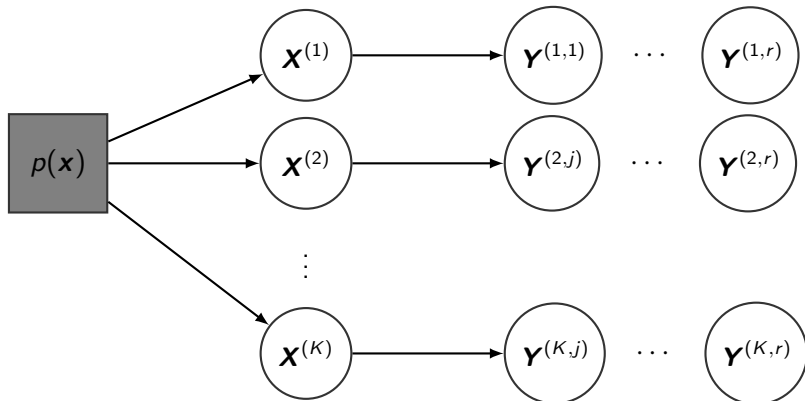


- Let  $\mathcal{X}$  define a class of stimuli (faces, objects, sounds.)
- Stimulus  $\mathbf{X} = (X_1, \dots, X_p)$ , where  $X_i$  are features (e.g. pixels.)
- Present  $\mathbf{X}$  to the subject, record the subject's brain activity using EEG, MEG, fMRI, or calcium imaging.
- Recorded response  $\mathbf{Y} = (Y_1, \dots, Y_q)$ , where  $Y_i$  are single-cell responses, or recorded activities in different brain region.

Image credits: Quiroga et al. (2009).

# Experimental design

- How to make inferences about the population of stimuli in  $\mathcal{X}$  using finitely many examples?
- *Randomization*. Select  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(K)}$  randomly from some distribution  $p(\mathbf{x})$  (e.g. an image database). Record  $r$  responses from each stimulus.



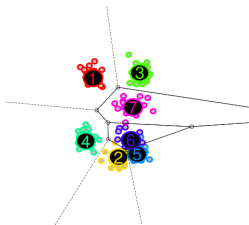
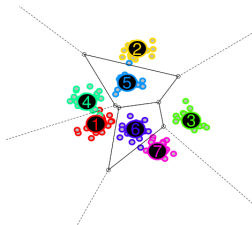
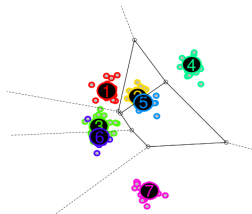
# Analyzing the data using machine learning

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

# Gaussian example

To help think about these problems, consider a concrete example:

- Let  $\mathbf{X} \sim N(0, I_d)$  and  $\mathbf{Y}|\mathbf{X} \sim N(\mathbf{X}, \sigma^2 I_d)$ .
- We draw stimuli  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(K)} \sim N(0, I_d)$  i.i.d.
- For each stimulus  $\mathbf{x}^{(i)}$ , we draw observations  $\mathbf{y}^{(i,j)} = \mathbf{x}^{(i)} + \epsilon^{(i,j)}$ , where  $\epsilon^{(i,j)} \sim N(0, \sigma^2 I_d)$ .



# 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  $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(K)}$ !

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

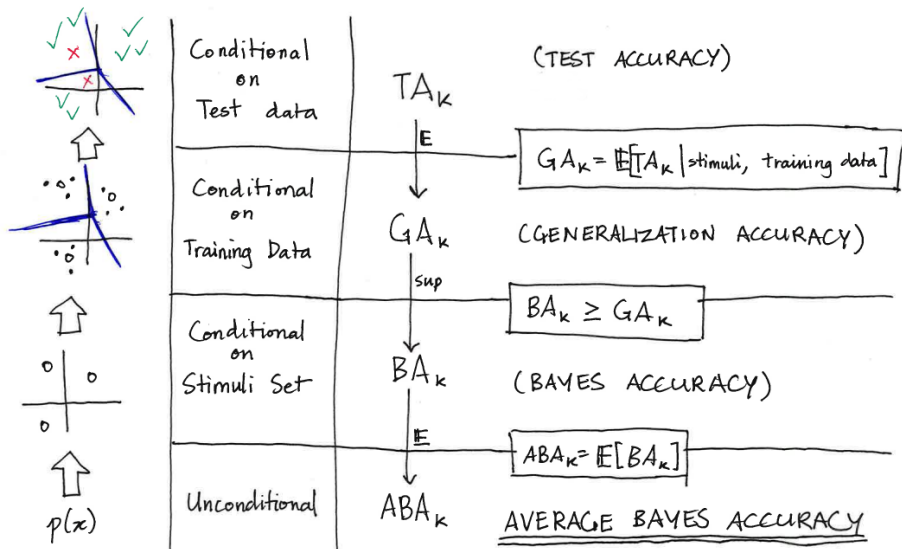
Define *average Bayes accuracy* as the expected Bayes accuracy

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

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:

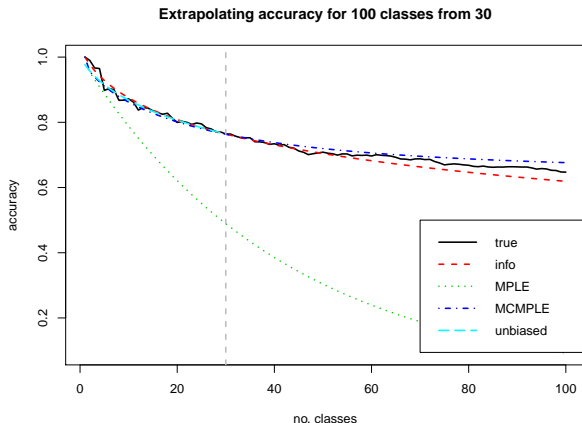
*What is the variability of  $BA_k$ ?*

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.

# Mutual information

## 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

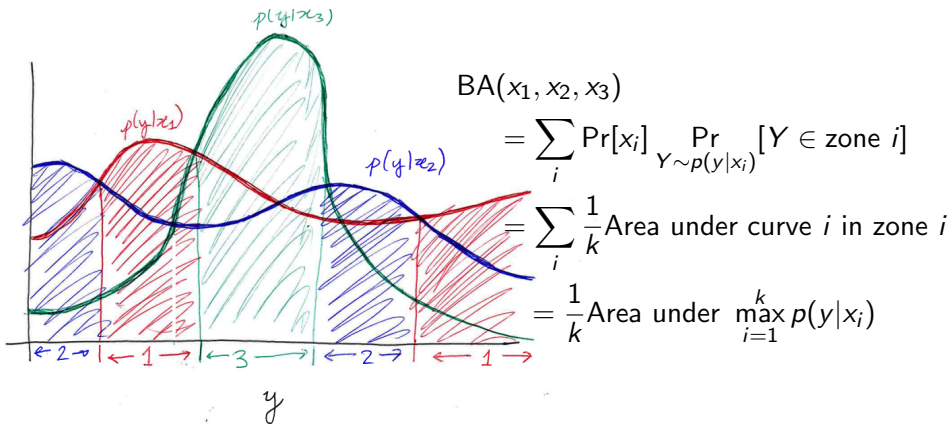
$$f(y) = \operatorname{argmax}_{i=1}^k p(y|x_i),$$

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  $(X_1, \dots, X_{k-1})$ ,  $\text{BA}_k$  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

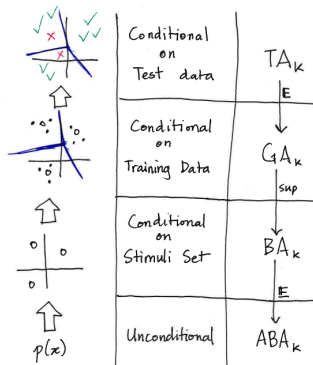
# Inferring mutual information from classification accuracy

- 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

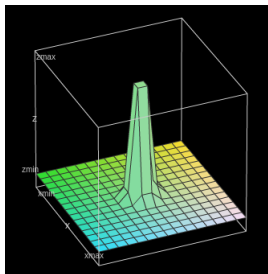
- Does  $ABA_k$  close to 1 imply  $I$  large?
- Does  $ABA_k$  close to  $1/k$  imply  $I$  close to 0?
- Does  $I$  large imply  $ABA_k$  close to 1?
- Does  $I$  close to 0 imply  $ABA_k$  close to  $1/k$ ?

# 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 bound

$$\text{ABA}_k[p(x, y)]?$$

In other words, does there exist a constant

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

# Step 1: Reduction

Let  $\mathcal{P}^{unif}$  be the class of densities  $p(x, y)$  on the unit square, with uniform marginals

$$p(x) = 1, \quad p(y) = 1.$$

We can show that if  $C_k(\iota)$  exists, then we need only to search through distributions in  $\mathcal{P}^{unif}$  to compute it:

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

We omit the proof.



# Step 1: Reduction

Reduction simplifies the formulas. For  $p(x, y) \in \mathcal{P}^{unif}$ ,

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

and

$$ABA_k[p(x, y)] = \frac{1}{k} \int \max_{i=1}^k p(x_i, y) dx_1 \cdots dx_k dy.$$

## Step 2: Conditioning

Due to uniform marginals  $p(x, y) = p(x|y) = p(y|x)$ . Therefore,

$$I[p(x, y)] = \int p(x|y) \log p(x|y) dx dy,$$

and

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

## Step 2: Conditioning

Write both functionals as an nested integral, with the outer integration over  $y$

$$I[p(x, y)] = \int \left[ \int p(x|y) \log p(x|y) dx \right] dy,$$

and

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

Notice that in both cases, the inner integral is a functional of  $p(x|y)$ .

## Step 2: Conditioning

Name the functionals, as *conditional information*

$$I[p(x, y)] = \int \underbrace{\left[ \int p(x|y) \log p(x|y) dx \right]}_{CI[p(x|y)]} dy,$$

and *conditional accuracy*

$$ABA_k[p(x, y)] = \frac{1}{k} \int \underbrace{\left[ \int \max_{i=1}^k p(x_i|y) dx_1 \cdots dx_k \right]}_{CA_k[p(x|y)]} dy.$$

## Step 2: conditioning

Conditional information and conditional accuracies are functionals of one-dimensional densities.

$$\text{CI}[q(x)] = \int \log q(x) dx,$$

$$\text{CA}_k[q(x)] = \frac{1}{k} \int \max_{i=1}^k q(x_i) dx_1 \cdots dx_k.$$

And, we have

$$\text{I}[p(x, y)] = \int \text{CI}[p(x|y)] dy = \mathbf{E}[\text{CI}[p(x|Y)]],$$

and

$$\text{ABA}_k[p(x, y)] = \int \text{CA}_k[p(x|y)] dy = \mathbf{E}[\text{CA}_k[p(x|Y)]].$$

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

## Step 3: convexity

- Given  $p(x, y) \in \mathcal{P}^{unif}$ , define a map  $\phi : [0, 1] \rightarrow \mathbb{R}^2$  by

$$\phi(y) = (\text{CI}[p(x|y)], \text{CA}_k[p(x|y)]).$$

We are mapping observations  $y$  to their associated conditional information and accuracy.

- Consider the image of  $\phi$ :

$$\text{Im}(\phi) = \{(\text{CI}[p(x|y)], \text{CA}_k[p(x|y)]) : y \in [0, 1]\}$$

- We have  $(\text{I}[p(x, y)], \text{ABA}_k[p(x, y)]) = \mathbf{E}[\phi(y)]$ .
- Therefore,

$$(\text{I}[p(x, y)], \text{ABA}_k[p(x, y)]) \in \mathcal{C}(\text{Im}(\phi)),$$

the convex hull of the image of  $\phi$ .

## Step 3: convexity

- Let  $\mathcal{P}^1$  denote the set of all densities  $q(x)$  on  $[0, 1]$ . Define the set  $S$  as

$$S : \{(\text{CI}[q(x)], \text{CA}_k[q(x)]) : q \in \mathcal{P}^1\}$$

- Since all conditional densities  $p(x|y) \in \mathcal{P}^1$ , we have

$$\text{Im}(\phi) \subset S,$$

and therefore

$$\mathcal{C}(\text{Im}(\phi)) \subset \mathcal{C}(S).$$

- Therefore,

$$(\text{I}[p(x, y)], \text{ABA}_k[p(x, y)]) \in \mathcal{C}(S).$$

- Cover and Thomas. Elements of information theory.
- Muirhead. Aspects of multivariate statistical theory.
- van der Vaart. Asymptotic statistics.