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

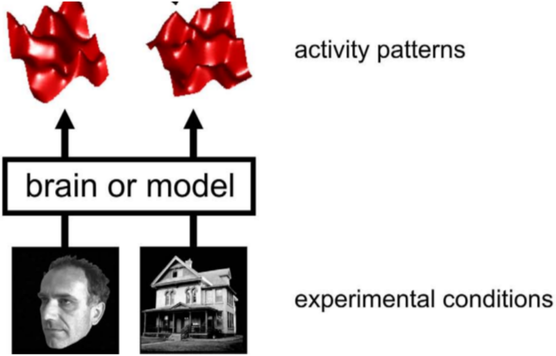
Charles Zheng

Stanford University

October 4, 2016

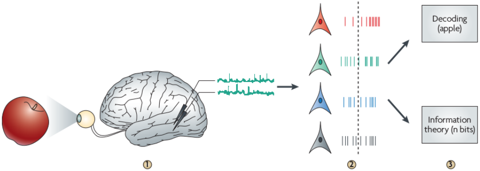
(Joint work with Yuval Benjamini.)

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

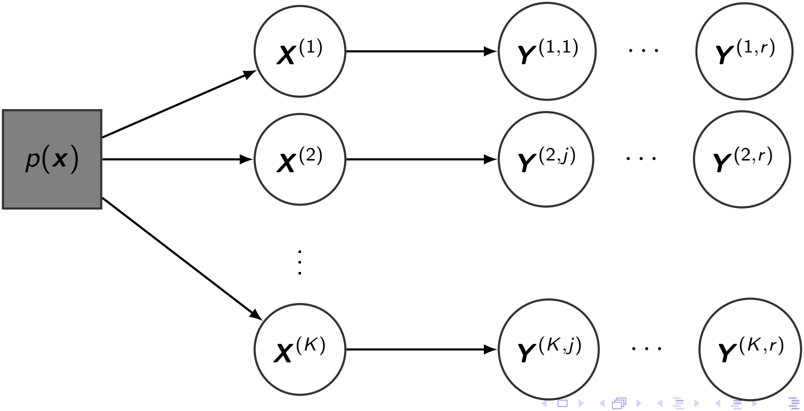


- Let $\mathcal{X}$ define a class of stimuli (faces, objects, sounds.)
- Stimulus $\boldsymbol{X} = (X_1, \ldots, X_p)$, where X_i are features (e.g. pixels.)
- Present $\boldsymbol{X}$ to the subject, record the subject's brain activity using EEG, MEG, fMRI, or calcium imaging.
- Recorded response $\boldsymbol{Y} = (Y_1, \ldots, Y_q)$, 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 $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(K)}$ randomly from some distribution $p(\boldsymbol{x})$ (e.g. an image database). Record r responses from each stimulus.



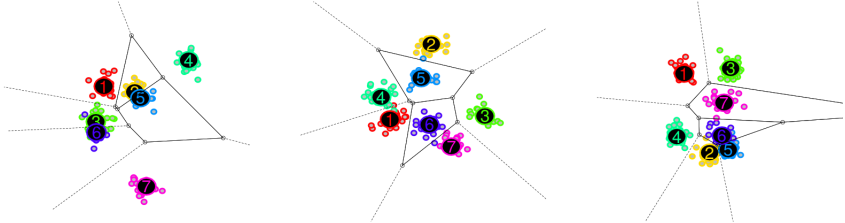
Analyzing the data using machine learning

- 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 $\boldsymbol{X} \sim N(0, I_d)$ and $\boldsymbol{Y}|\boldsymbol{X} \sim N(\boldsymbol{X}, \sigma^2 I_d)$.
- We draw stimuli $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(K)} \sim N(0, I_d)$ i.i.d.
- For each stimulus $\boldsymbol{x}^{(i)}$, we draw observations $\boldsymbol{y}^{(i,j)} = \boldsymbol{x}^{(i)} + \epsilon^{(i,j)}$, 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: Generalizing to identical replicate

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 $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(K)}$!

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

$$\text{ABA}_k = \mathbf{E}[\text{BA}_k]$$

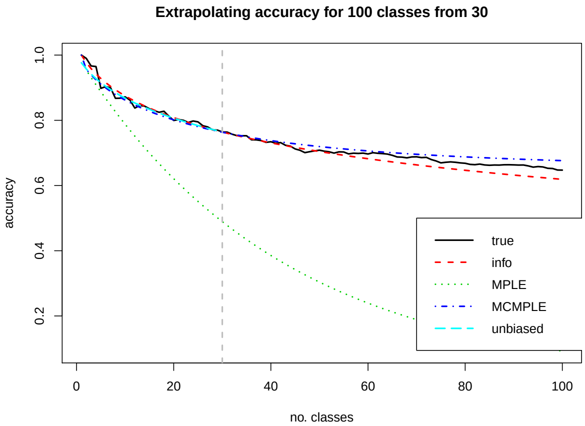
Converting a LCB for BA_k to an LCB on ABA_k boils down to the following problem: *What is the variability of* BA_k?

Concentration of Bayes accuracy

We have both theoretical and empirical results on $\mathrm{Var}[\mathrm{BA}_k]$.

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 Benjamnini (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

Inferring mutual information from classification accuracy

Outline

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

Outline

- Step 1: Apply machine learning to obtain *test accuracy* TA_k
- Step 2: Use TA_k to infer the generalization accuracy GA_k

$$\mathrm{GA}_k \geq \mathrm{TA}_k - z_\alpha \sqrt{\frac{\mathrm{TA}_k(1-\mathrm{TA}_k)}{n_{test}}} \text{ with probability } \geq 1-\alpha$$

- Step 3: The generalization accuracy is a lower bound on the Bayes accuracy,

$$\mathrm{GA}_k \leq \mathrm{BA}_k$$

- Step 4: Use BA_k to infer the average Bayes accuracy

$$\mathrm{ABA}_k \geq \mathrm{BA}_k - \frac{1}{2\sqrt{\alpha k}} \text{ with probability } \geq 1-\alpha$$

References

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