How many neurons does it take to classify a lightbulb?

Charles Zheng

Stanford University

December 14, 2015

(Joint work with Yuval Benjamini)

Overview

Introduction

- Review of information theory
- Study of neural coding

Related work

- Estimating mutual information between stimulus and response.
- Can we use machine learning methods to estimate MI?

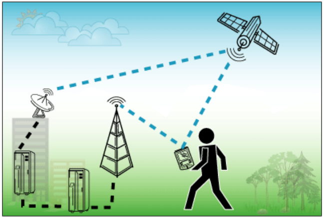
Methods

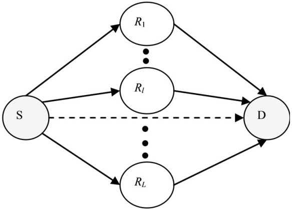
- Setup
- Gaussian example
- Using Fano's inequality
- Using low-SNR universality

Results

Information theory

The high performance and reliability of modern communications system is made possible by information theory, founded by Shannon in 1948.





A information-processing network can be analyzed in terms of interactions between its components (which are viewed as random variables.)

Image credit CartouCHe, Aziz et al. 2011

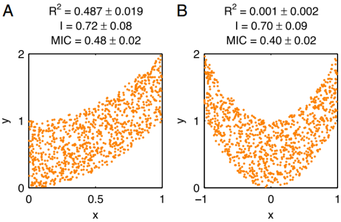
Entropy and mutual information

X and Y have joint density $p(x, y)$ with respect to μ.

Quantity	Definition	Linear analogue
Entropy	$H(X) = -\int(\log p(x))p(x)\mu_X(dx)$	$\text{Var}(X)$
Conditional entropy	$H(X\|Y) = \mathbf{E}[H(X\|Y)]$	$\mathbf{E}[\text{Var}(X\|Y)]$
Mutual information	$I(X;Y) = H(X) - H(X\|Y)$	$\text{Cor}^2(X, Y)$

The above definition includes both *differential* entropy and *discrete* entropy.
Information theorists tend to use log base 2, we will use natural logs in this talk.

Properties of mutual information



- $I(X;Y) \in [0,\infty]$. (0 if $X \perp Y$, ∞ if $X = Y$ and X continuous.)
- Symmetric: $I(X;Y) = I(Y;X)$
- Bijection-invariant: $I(\phi(X);\psi(Y)) = I(\psi(Y);\phi(X)$.
- Additivity. If $(X_1, Y_1) \perp (X_2, Y_2)$, then

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

- Relation to KL divergence $\mathbb{D}$.

$$\mathbb{D}(p(x,y)||p(x)p(y)) = I(X;Y)$$

Image credit Kinney et al. 2014

Relationship between mutual information and classification

- Suppose X and Y are discrete random variables, and X is uniformly distributed over its support.
- Classify X given Y. The optimal rule is to guess

$$\hat{X} = \text{argmax}_x \; p(Y|X = x)$$

- Bayes error:

$$p_e = \Pr[X \neq \hat{X}]$$

- Fano's inequality:

$$I(X;Y) \geq (1 - p_e) \ln K - p_e \ln(p_e) - (1 - p_e) \ln(1 - p_e)$$

where K is the size of the support of X.

Relationship between mutual information and classification

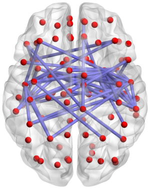
A nice interpretation of $I(X; Y)$ for *continuous* random variables:

- Bin the continuous X into K equal-probability bins.
- $I(X; Y)$ approx. measures how finely we can bin X so that Y can still distinguish the bins with high probability!

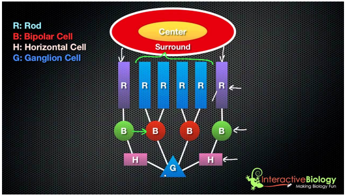
This is another way of seeing why $I(X; Y)$ is bijection-invariant!

Studying the neural code

The brain is the *most complex* information processing system we know!



Neural network inferred from data (Hong et al.)

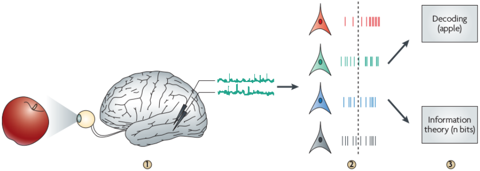


Organization of human retina

How do neurons encode, process, and decode sensory information?

Image credit: Hong et al., Interactive Biology

Studying the neural code: data



- Let $\mathcal{X}$ define a class of stimuli (faces, objects, sounds.)
- Stimulus $\mathbf{X} = (X_1, \ldots, X_p)$, 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 $\mathbf{Y} = (Y_1, \ldots, Y_q)$, where Y_i are single-cell responses, or recorded activities in different brain region.

Image credits: Quiroga et al. (2009)

Problem statement

Given stimulus-reponse data $(\mathbf{X}, \mathbf{Y})$, can we estimate the mutual information $I(\mathbf{X}; \mathbf{Y})$?

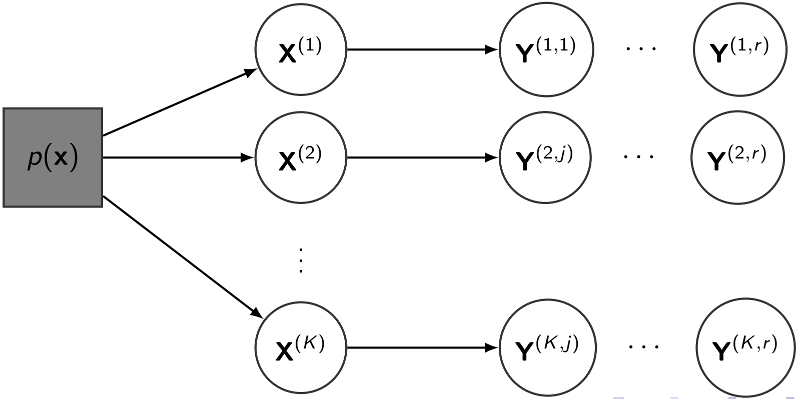
Why do we care?

- Selecting the correct model for neural encoding
- Assessing the *efficiency* of the neural code
- Measuring the *redundancy* of a population of neurons

$$r' = \frac{\sum_{i=1}^{q} I(\mathbf{X}; Y_i) - I(\mathbf{X}; \mathbf{Y})}{\sum_{i=1}^{q} I(\mathbf{X}; Y_i)}$$

Experimental setup

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



Can we learn $I(\mathbf{X};\mathbf{Y})$ from such data?

Answer: yes.

- Let $p^*(\mathbf{x})$ be the uniform distribution over $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(K)}$, and let $\tilde{\mathbf{X}}$ be a random vector with this distribution.
- Let $\tilde{\mathbf{Y}}$ have the distribution

$$p^*(\tilde{\mathbf{y}}) = \frac{1}{K}\sum_{i=1}^{K} p(\mathbf{y}|\mathbf{x}^{(i)})$$

- Then, as $K \to \infty$, $I(\tilde{\mathbf{X}};\tilde{\mathbf{Y}}) \xrightarrow{p} I(\mathbf{X};\mathbf{Y})$, where

$$I(\tilde{\mathbf{X}};\tilde{\mathbf{Y}}) = H(\tilde{\mathbf{Y}}) - \frac{1}{K}\sum_{i=1}^{K} H(\mathbf{Y}|\mathbf{X}^{(i)})$$

- We can apply nonparametric methods to estimate $H(\mathbf{Y}|\mathbf{X}^{(i)})$ for $i = 1, \ldots, K$, and $H(\tilde{\mathbf{Y}})$. Plugging those estimates into the above formula gives a *nonparametric* estimate of $I(\mathbf{X};\mathbf{Y})$.

References

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