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?

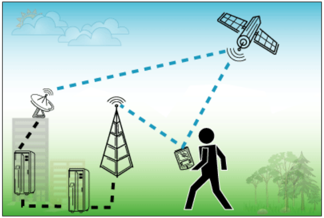
Methods

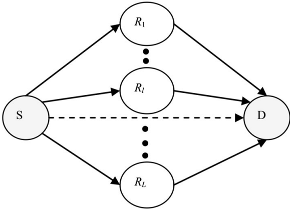
- Gaussian example.
- Using Fano's inequality.
- Using low-SNR universality.

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

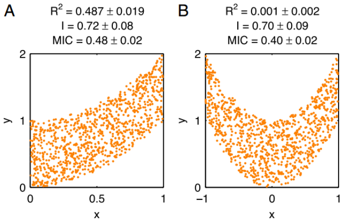
X and Y have joint density $p(x, y)$ with respect to μ.

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

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.)
- Symmetry: $I(X;Y) = I(Y;X)$.
- 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:

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

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:

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

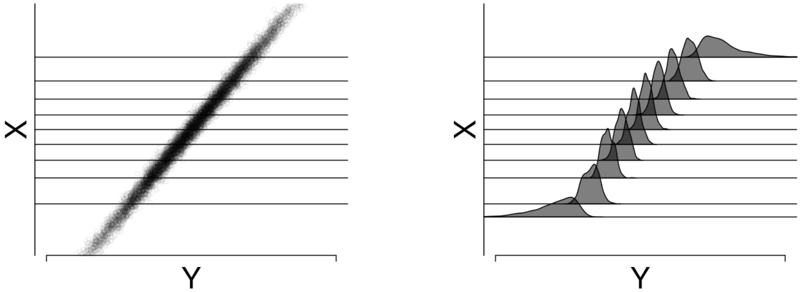
- Fano's inequality:

$$I(X; Y) \geq (1 - p_e) \ln K - \text{const}.$$

where K is the size of the support of X.

Nice interpretation of $I(X; Y)$ for continuous rvs

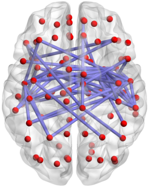
- If we bin the continuous X into $K \approx e^{I(X;Y)}$ equal-probability bins, we can reliably guess the bin given Y.
- Heuristic is more accurate if $I(X; Y)$ is large, due to Shannon's noisy channel theorem.



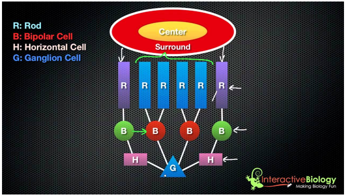
$I(X; Y) = 2.3038$ $\qquad$ $\ln 10 = 2.3025$

Motivation: the neural code

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



Neural network inferred from data. (Hong et al.)

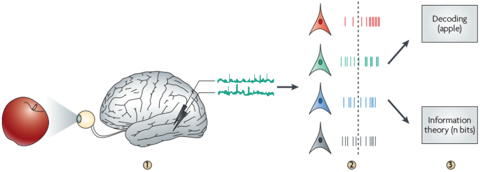


Organization of human retina

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 $I(\mathbf{X}; \mathbf{Y})$?

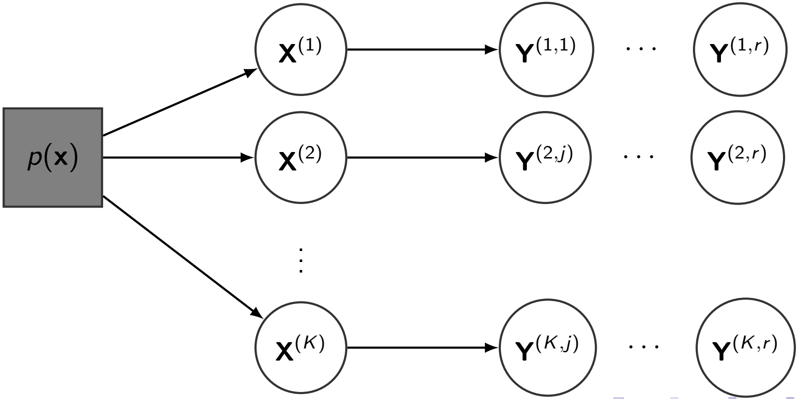
Why do we care?

- 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 design

- 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.

- We have $I(\mathbf{X};\mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X})$.
- We can estimate $H(\mathbf{Y})$ from the data
- We can estimate $H(\mathbf{Y}|\mathbf{x}^{(i)})$ from the data, and define

$$\hat{H}(\mathbf{Y}|\mathbf{X}) = \frac{1}{K}\sum_{i=1}^{K} \hat{H}(\mathbf{Y}|\mathbf{X}^{(i)})$$

- As K and r both tend to infinity,

$$\hat{I}(\mathbf{X};\mathbf{Y}) = \hat{H}(\mathbf{Y}) - \hat{H}(\mathbf{Y}|\mathbf{X})$$

is consistent for $I(\mathbf{X};\mathbf{Y})$.

Limitations with the 'naive' approach

Naive estimator:

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

- If K is small, the naive estimator may be quite biased, even for low-dimensional problems. Gastpar et al. (2010) introduced an *antropic correction* to deal with the small-K bias.
- Difficult to estimate differential entropies $H(\mathbf{Y})$, $H(\mathbf{Y}|\mathbf{x}^{(i)})$ in high dimensions. Best rates are $O(1/\sqrt{n})$ for $d \leq 3$ dimensions. Convergence rates for $d > 3$ unknown!

Can we use machine learning to deal with dimensionality?

- Supervised learning becomes an extremely common approach for dealing with high-dimensional data, for numerous reasons!
- Perhaps we can use supervised learning to estimate $I(\mathbf{X}; \mathbf{Y})$ as well.

Procedure.

- Fix $r_{train} < r$. Let $r_{test} = r - r_{train}$.
- Use $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i,j)}) : i = 1, \ldots, K, j = 1, \ldots, r_{train}\}$ as training data to learn a classifier $\hat{\mathbf{x}}$.
- Compute the confusion matrix, normalized so that each row adds to $1/K$:

$$C(i,j) = \frac{1}{K^2 r_{test}} \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{\ell = r_{train}+1}^{r} I(\hat{\mathbf{x}}(\mathbf{y}^{(i,\ell)}) = \mathbf{x}^{(j)}).$$

Normalized this way, $C(i,j)$ gives the empirical joint distribution

$$C(i,j) = \hat{\Pr}[\mathbf{X} = \mathbf{x}^{(i)}, \hat{\mathbf{X}} = \mathbf{x}^{(j)}].$$

Can we use machine learning to deal with dimensionality?

- Treves et al. (1997) suggest computing the mutual information from the confusion matrix, i.e.

$$\hat{I}(\mathbf{X};\mathbf{Y}) \approx \sum_{i=1}^{K}\sum_{j=1}^{K} C(i,j) \ln\left(\frac{C(i,j)}{\left(\sum_{\ell=1}^{K} C(i,\ell)\right)\left(\sum_{\ell=1}^{k} C(j,\ell)\right)}\right)$$

- Quiroga (2009) review the applications of this approach, and note sources of bias or "information loss."

Why use supervised learning to estimate $I(\mathbf{X}; \mathbf{Y})$?

- Successful supervised learning exploits structure in the data, which *nonparametric methods ignore.*
- Using supervised learning to estimate mutual information can be viewed as *using prior information* to improve the estimate of $I(\mathbf{X}; \mathbf{Y})$.

Interesting connection to machine learning literature

While we are considering

$$\text{supervised learning} \rightarrow \text{estimate mutual information},$$

a vast literature exists on applications of mutual information (as the 'infomax criterion') for feature selection, training objectives, i.e.

$$\text{estimate mutual information} \rightarrow \text{supervised learning}.$$

Questions

- How much could we potentially gain in estimating $I(\mathbf{X};\mathbf{Y})$ by using supervised learning, compared to nonparametric approaches?
- Is the Bayes confusion matrix sufficient for consistently estimating $I(\mathbf{X};\mathbf{Y})$?
- Is the Bayes error sufficient for consistently estimating $I(\mathbf{X};\mathbf{Y})$?
- In practice, we cannot obtain the Bayes error due to:
  - Model mispecification.
  - Finite training data to fit the model (even if correctly specified).
  - Finite test data to estimate the generalization error.

  How sensitive is our estimator to these issues?

References

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