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?

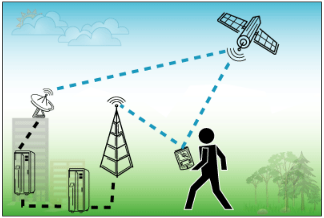
Theory

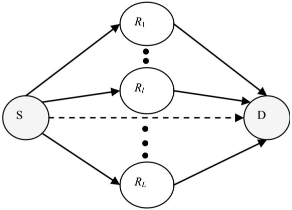
- Gaussian example.
- Low-SNR universality.

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.

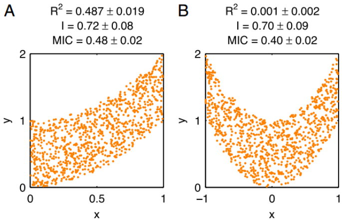
Entropy and mutual information

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:

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

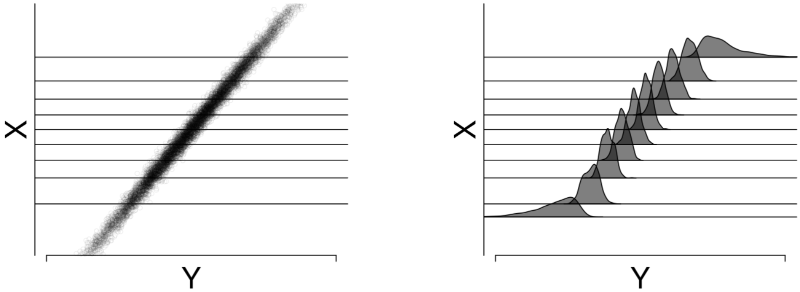
- Fano's inequality:

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

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.

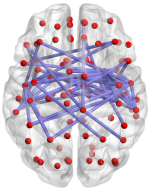


$I(X;Y) = 2.3038$

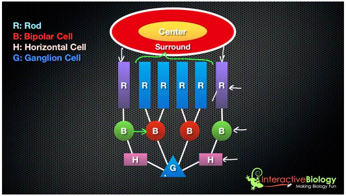
$\ln 10 = 2.3025$

Motivation: the neural code

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



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

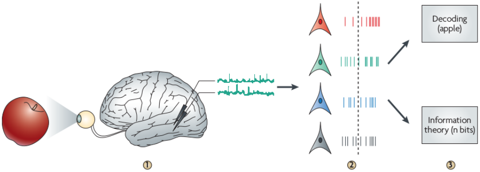


Organization of human retina

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

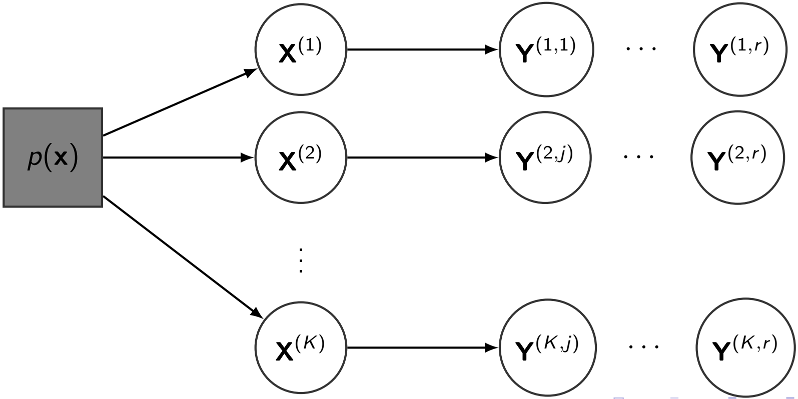
Why do we care?

- 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 $I(\mathbf{X};\mathbf{Y})$.
- So while the general problem of information estimation is nearly impossible in high dimensions, the problem might become tractable if we can exploit known structure in the problem!

Interesting connection to machine learning literature

While we are considering

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

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

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

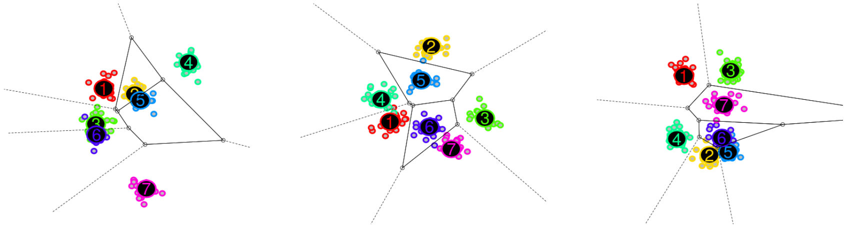
Questions

- 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?

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 $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)} \sim N(0, I_d)$ i.i.d.
- For each stimulus $\mathbf{x}^{(i)}$, we draw observations $\mathbf{y}^{(i,j)} = \mathbf{x}^{(i)} + \epsilon^{(i,j)}$, where $\epsilon^{(i,j)} \sim N(0, \sigma^2 I_d)$.



Gaussian example

The mutual information is given by

$$I(\mathbf{X};\mathbf{Y}) = \frac{d}{2}\log(1+\frac{1}{\sigma^2}).$$

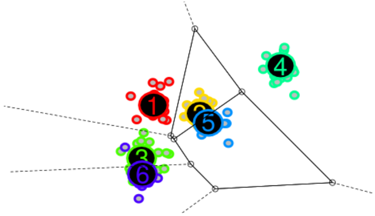
The Bayes rule takes the form

$$\hat{\mathbf{X}}(\mathbf{Y}) = \text{argmax}_{\mathbf{x}^i} \underbrace{-\frac{1}{2\sigma^2}||\mathbf{Y}-\mathbf{x}||^2}_{Z_i}.$$

Notation: Without loss of generality, assume $\mathbf{Y}$ belongs to the centroid $\mathbf{x}^{(K)}$. Then write $\mathbf{x}^* = \mathbf{x}^{(K)}$ and $Z_* = Z_K$.
The average Bayes error can be written

$$\text{ABE} = \Pr[Z_* < \max_{i=1}^{K-1} Z_k].$$

Gaussian example: Average Bayes error



- Conditional on the centroid locations $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)}$, the Bayes misclassification error can be computed by integrating p-dimensional Gaussian density over Voronoi polytopes.
- This is pretty much intractable in high dimensions!
- The average Bayes error (ABE) is an average of an intractable quantity. Luckily, *taking averages makes the problem tractable*.

Gaussian example: Average Bayes risk

- To make the problem even easier, we use another time-honored technique: the central limit theorem.
- Letting $p \to \infty$, the scores

$$Z_i = -\frac{1}{2\sigma^2}||\mathbf{Y} - \mathbf{x}||^2$$

have a normal joint distribution in the limit:

$$\begin{bmatrix} Z_* \\ Z_1 \\ \vdots \\ Z_{K-1} \end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix} -\frac{d}{2} \\ -\frac{d}{2} - \frac{1}{\sigma^2} \\ \vdots \\ -\frac{d}{2} - \frac{1}{\sigma^2} \end{bmatrix}, \begin{bmatrix} \frac{d}{2} & \frac{d}{2} & \cdots & \frac{d}{2} \\ \frac{d}{2} & \frac{d}{2} + \frac{2}{\sigma^2} & \cdots & \frac{d}{2} + \frac{1}{\sigma^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d}{2} & \frac{d}{2} + \frac{1}{\sigma^2} & \cdots & \frac{d}{2} + \frac{2}{\sigma^2} \end{bmatrix} \right).$$

Gaussian example: Average Bayes risk

Assume $(Z_*, Z_1, \ldots, Z_{K-1})$ have a normal distribution with the given moments.
We can compute

$$\text{ABE} = \Pr[Z_* < \max_{i=1}^{K-1} Z_i]$$

by writing

$$Z_i = \frac{\text{Cov}(Z_*, Z_i)}{\text{Var}(Z_*)}(Z_* - \mathbf{E}Z_*) + \sqrt{\text{Var}(Z_i) - \frac{\text{Cov}(Z_*, Z_i)^2}{\text{Var}(Z_*)}} W_i,$$

where W_i are i.i.d. standard normal.
This yields

$$\Pr[Z_* < \max_{i=1}^{K-1} Z_i] = \Pr[N(\mu, \nu^2) < \max_{i=1}^{K-1} W_i]$$

where

$$\mu = \frac{\mathbf{E}[Z_* - Z_i]}{\sqrt{\frac{1}{2}\text{Var}(Z_i - Z_j)}}, \quad \nu^2 = \frac{\text{Cov}(Z_* - Z_i, Z_* - Z_j)}{\frac{1}{2}\text{Var}(Z_i - Z_j)}$$

for $i \neq j \neq K$.

Gaussian example: Average Bayes risk

Finally, we get

$$\text{ABE} = \Pr[Z_* < \max_{i=1}^{K-1} Z_i] \to f_K\left(\frac{\sqrt{d}}{\sigma}\right)$$

where

$$f_K(\mu) = 1 - \int_{-\infty}^{\infty} \phi(z-\mu)(1-\Phi(z))^{K-1} dz.$$

Gaussian example: Average Bayes risk

Recall that

$$I(\mathbf{X}; \mathbf{Y}) = \frac{d}{2} \log(1 + \frac{1}{\sigma^2}),$$

while

$$\text{ABE} = f_K(\sqrt{d}/\sigma).$$

Hence ABE is not a function of $I(\mathbf{X}; \mathbf{Y})$!

Gaussian example: Low SNR limit

However, what if we consider a limit where the noise level σ^2 increases with d?
Fix some $\sigma_1^2 > 0$, and let $\sigma_d^2 = d\sigma_1^2$.
Then when d is large,

$$I(\mathbf{X};\mathbf{Y}) = \frac{d}{2}\log(1 + \frac{1}{d\sigma_1^2}) \approx \frac{d}{2}\frac{1}{d\sigma^1} = \frac{1}{2\sigma_1^2}.$$

We get

$$\text{ABE} = f_k(\sqrt{2I(\mathbf{X};\mathbf{Y})})$$

in the limit!

Low SNR limit: generalization

In a sequence of gaussian models of increasing dimensionality with

$$\lim_{d\to\infty} I(\mathbf{X};\mathbf{Y}) \to \iota < 0,$$

we get an exact relationship between the limiting mutual information and the average Bayes error,

$$\text{ABE} = f_K(\sqrt{2\iota}).$$

This limiting relationship holds more generally!

Exponential family sequence model

The Gaussian sequence is a special case of the following class of models. Let $b_X(x)$ and $b_Y(y)$ be probability densities and $u(x, y)$ be an arbitrary smooth function. Let

$$b_\theta(x, y) = \frac{b_X(x) b_Y(y) \exp[\theta u(x, y)]}{\int b_X(x) b_Y(y) \exp[\theta u(x, y)] dxdy}.$$

A corresponding *exponential family sequence model* is a sequence of joint probability distributions $p_d(\mathbf{x}, \mathbf{y})$ given by

$$p_d(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{d} b_{\theta_d}(x_i, y_i)$$

such that

$$\lim_{d \to \infty} d\theta_d = c < \infty.$$

For instance, choosing $b_X(x) = b_Y(y) = \phi(x)$ and $u(x, y) = xy$ yields the Gaussian sequence model (up to marginal scaling.)

Low SNR theorem

Theorem. *Given an exponential family sequence model* $p_d(\mathbf{x}, \mathbf{y})$, *for random variates* $(\mathbf{X}^{[d]}, \mathbf{Y}^{[d]}) \sim p_d(\mathbf{X}, \mathbf{Y})$, *we have*

$$\lim_{d\to\infty} I(\mathbf{X}^{[d]}, \mathbf{Y}^{[d]}) = \iota < \infty$$

for some constant $\iota < \infty$; *and the limiting* K*-class average Bayes error is given by*

$$\lim_{d\to\infty} ABE = f_K(\sqrt{2\iota}).$$

The low-SNR estimator of $I(\mathbf{X}; \mathbf{Y})$

We are willing to bet that the relationship

$$\text{ABE} \approx f_K(\sqrt{2\iota})$$

holds in much greater generality than we managed to prove–namely, whenever $I(\mathbf{X}; \mathbf{Y}) \ll p$, and the scores Z_i are approximately jointly multivariate normal.

Based on these assumptions, our proposed estimator for mutual information is

$$\hat{I}_{ls}(\mathbf{X}; \mathbf{Y}) = \frac{1}{2} f_K^{-1}(\widehat{\text{ABE}})^2$$

where $\widehat{\text{ABE}}$ is the test error of the classifier. (The subscript ls stands for low-SNR.)

References

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