Boulder Theoretical Biophysics 2019

Neuroscience Mini-course: Exercise Set 4

Some problems for this lecture are adapted from Cover and Thomas, Chapter 10.

Rate distortion theory: Consider a source, X, that is subject to encoding, passage through a noisy channel, and decoding that results in a mapping, $\hat{X}(X)$, that distorts the source. In this notation, the output of the decoder is $\hat{X}$. We seek a mapping from X to $\hat{X}$ that keeps the average of the distortion, d, of the source over all $\{x, \hat{x}\}$ less than or equal to a maximal distortion that is set to some value, D, while minimizing the bit-rate of the channel, R. Lowercase d denotes the distortion for particular values x and $\hat{x}$, drawn from the distribution, $P(X, \hat{X})$. The rate-distortion theorem states that this target minimal rate, subject to a constraint on the average distortion, is equal to the minimal mutual information between X and $\hat{X}$, subject to the same constraint on the distortion,

$$R(D) = \min_{p(\hat{x}|x):\sum_{x,\hat{x}} p(x)p(\hat{x}|x)d(x,\hat{x})\leq D} I(X;\hat{X}). \tag{1}$$

Squared-error distortion: Consider a continuous random variable, X, with mean zero and variance σ^2, passed through a noisy channel and decoded. The decoding performance is measured with squared-error distortion. That means that the distortion function, d, is the mean squared-error between X and $\hat{X}$, $\langle d(x,\hat{x})\rangle = \langle (x-\hat{x})^2\rangle_{p(x)p(\hat{x}|x)}$.

1. Show that

$$S(X) - \frac{1}{2}\log(2\pi e D) \leq R(D),$$

...and show that

$$R(D) \leq \frac{1}{2}\log\frac{\sigma^2}{D}.$$

You will find it useful to remember the formula for the entropy of a Gaussian distribution, and recall that a Gaussian distribution is the maximum entropy distribution for constrained variance. You should also note that a conditional entropy is always smaller than or equal to an unconditioned entropy, i.e. $S(X) \geq S(X|Y)$.

For this problem, consider the mapping (encoder: X, a noisy channel, decoder: $\hat{X}$)

$$\hat{X} = \frac{\sigma^2 - D}{\sigma^2}(X + Z),$$

where Z is a Gaussian variable with zero mean and variance $\frac{D\sigma^2}{\sigma^2 - D}$. X and Z are independent.

2. With this decoder, check that $\langle d(x,\hat{x})\rangle = \langle (x-\hat{x})^2\rangle_{p(x)p(\hat{x}|x)} = D$.

3. Are Gaussian random variables harder or easier to 'describe' than other random variables with the same variance? Meaning, do you have to use a higher $R(D)$ to read out a Gaussian source with the same distortion, D, than any other source? Hint: Consider the case in this problem where the source is Gaussian, with the same mean and variance as stated here.

The Information bottleneck: A method for solving for the rate-distortion function, $R(D)$, is to rewrite the constraint in equation 1 using the method of Lagrange multipliers,

$$(2) \qquad \min_{p(\hat{x}|x)} \mathcal{L} = I(X;\hat{X}) - \beta \langle d(x,\hat{x}) \rangle_{p(x,\hat{x})} - \sum_x \lambda(x) \left(\sum_{\hat{x}} p(\hat{x}|x) - 1 \right),$$

where the third term enforces the normalization of $p(\hat{x}|x)$. In the information bottleneck approach, we derive a particularly interesting choice of the distortion function, $d = I(X;Y)$, where Y is a variable that describes what we define as the 'relevant' information in X. The parameter β sets the tradeoff between compressing (minimizing $I(X;\hat{X})$), and retaining relevant information,(maximizing $I(\hat{X};Y)$). Equation 2 then becomes

$$(3) \qquad \min_{p(\hat{x}|x)} \mathcal{L} = I(X;\hat{X}) - \beta I(\hat{X};Y) - \sum_x \lambda(x) \left(\sum_{\hat{x}} p(\hat{x}|x) - 1 \right).$$

First note that

$$p(y|\hat{x}) = \sum_x p(y|x)p(x|\hat{x}),$$

which follows from the fact that we infer X from $\hat{X}$, and X carries information about Y. You might also like to use the identities,

$$p(\hat{x}) = \sum_x p(\hat{x}|x)p(x),$$

and

$$p(\hat{x}|y) = \sum_x p(\hat{x}|x)p(x|y).$$

4. Use these equations to derive expressions for

$$\frac{\delta p(\hat{x})}{\delta p(\hat{x}|x)}$$

and

$$\frac{\delta p(\hat{x}|y)}{\delta p(\hat{x}|x)}$$

5. Use the two expressions above (check with me if you are not sure of your answers) to simplify an expression for

$$\frac{\delta \mathcal{L}}{\delta p(\hat{x}|x)}.$$

Remember Bayes' Rule, and use it to simplify and rearrange your terms. Introduce the following change of variables for λ,

$$\tilde{\lambda} = \frac{\lambda(x)}{p(x)} + \beta \sum_y p(y|x) \log \left[\frac{p(y|x)}{p(y)} \right],$$

in which you should note that the second term only depends on x, not on $\hat{x}$, which is why we can absorb it into the Lagrange multiplier, λ. Set the derivative of $\mathcal{L}$ to zero and obtain an expression for $p(\hat{x}|x)$ in terms of the D_{KL} between $p(y|x)$ and $p(y|\hat{x})$. Hint: You should obtain

$$p(\hat{x}|x) \propto p(\hat{x}) \exp\left(-\beta D_{\mathrm{KL}}\left[p(y|x)||p(y|\hat{x})\right]\right).$$