Conditional Differential Entropy, Information theory in ML

Introduction

In this exercise we show an alternative proof of *Chow-Liu* algorithm for maximizing "tree distribution" we have seen in class. The exercise is based on a lecture note given by Prof. Weissman from Stanford. links: {lecture, course}.

Exercise

1. The Chain Rule for Relative Entropy

   (a) **Rule Statement**

   Definition 1. Conditional Relative Entropy. Given two conditional PMFs $P_{X|Y}$ and $Q_{X|Y}$, the conditional relative entropy is:

   $$D(P_{X|Y}||Q_{X|Y}|P_Y) = \sum_y D(P_{X|Y=y}||Q_{X|Y=y})P_Y(y)$$

   Exercise 1 Prove the following:

   i. The Chain Rule for Relative Entropy (Two Variables):

   $$D(P_{X,Y}||Q_{X,Y}) = D(P_X||Q_X) + D(P_{Y|X}||Q_{Y|X}|P_X)$$

   ii. The Chain Rule for Relative Entropy (Multiple Variables):

   $$D(P_{X_1,..,X_n}||Q_{X_1,..,X_n}) = \sum_{i=1}^{n} D(P_{X_i|X^{i-1}}||Q_{X_i|X^{i-1}}|P_{X^{i-1}})$$

   (b) **Applications of the Chain Rule**

   i. Rewriting Mutual Information
   Exercise 2 Prove:

   $$I(X;Y) = D(P_{Y|X}||P_Y|P_X)$$

ii. Minimizing Conditional Relative Entropy:
Exercise 3 Prove:

$$\min_{Q_Y} D(P_{Y|X}||Q_Y|P_X) = D(P_{Y|X}||P_Y|P_X) = I(X;Y)$$

iii. "Mutual Information" of three variables:

$$\begin{aligned} H(X)+H(Y)+H(Z)-H(X,Y,Z) &= I(X;Y)+I(X;Z)+I(Y;Z|X) \\ &= I(Y;Z)+I(Y;X)+I(X;Z|Y) \\ &= I(Z;X)+I(Z;Y)+I(X;Y|Z) \end{aligned}$$

The quantity $H(X)+H(Y)+H(Z)-H(X,Y,Z)$ is sometimes thought of as the "Mutual Information" between X, Y and Z. Recall the mutual information between two variables is $I(X;Y) = H(X)+H(Y)-H(X,Y)$.
Exercise 4 Prove

$$H(X)+H(Y)+H(Z)-H(X,Y,Z) = I(X;Y)+I(X;Z)+I(Y;Z|X)$$

2. Markov properties: Let $X-Y-Z$ be a Markov triplet. This means $p(x,y|z) = p(x|y)p(y|z)$, $p(x|y,z) = p(x|y)$, $p(z|y,x) = p(z|y)$. Intuitively, it means that X and Y are more closely related than X and Z.

   (a) $H(X|Y) = X(X|Y,Z)$
   (b) $H(Z|Y) = H(Z|X,Y)$
   (c) $H(X|Y) \leq H(X|Z)$
   (d) $I(X;Y) \geq I(X;Z)$
   (e) $I(Y;Z) \geq I(X;Z)$

   Exercise 5 Prove (a)-(e).

3. Tree distribution
   Recall from class $P_{x^n}(a) = \frac{N(a|x^n)}{n}$, i.e. empirical distribution. Suppose $(X_i, Y_i, Z_i) \sim$ iid $Q_{X,Y,Z}$, then

$$Q_{X,Y,Z}(x^n, y^n, z^n) = 2^{-n[H(P_{x^n,y^n,z^n})+D(P_{x^n,y^n,z^n}||Q_{X,Y,Z})]}. \quad (1)$$

   First, our goal is to find $Q_{X,Y,Z}$ corresponding to the fixed tree $Y-X-Z$ which maximize (1), i.e., to minimize $D(P_{x^n,y^n,z^n}||Q_{X,Y,Z})$. From now on we will use the notation $P_{X,Y,Z}$ instead of P_{x^n,y^n,z^n}.
   Exercise 6 Prove

$$\min_{Q_{X,Y,Z}:Y-X-Z} D(P_{X,Y,Z}||Q_{X,Y,Z}) = I(Z;Y|X).$$

   Hints:

   - Use exercises 1 and 3.

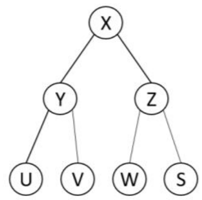


Figure 1: This tree corresponds to $P(x, y, z, u, v, w, s) = p(x)p(y|x)p(z|x)p(u|y)p(v|y)p(w|z)p(s|z)$

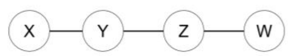


Figure 2: This tree corresponds to Markov 4-tuple $X - Y - Z - W$

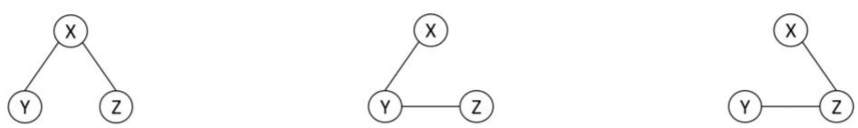
- Remember that P is known, i.e. $P_X, P_Y, P_{X|Z}, ...$ are known (usually we use the empirical distribution).

Exercise 7 Use the results from exercises 4 and 6 to show

$$\max_{Q_{X,Y,Z}: Y-X-Z} Q_{X,Y,Z}(x^n, y^n, z^n) = 2^{-n[H(X)+H(Y)+H(Z)]} 2^{n[I(X;Y)+I(X;Z)]}$$

*Note that $Q_{X,Y,Z}(x^n, y^n, z^n)$ is the likelihood of the data based on the distribution Q. Maximizing it is similar to Maximum-Likelihood criteria.

In this case we have 3 random variables and we want to model the data with a tree distribution. We have three options - $X - Y - Z$, $Z - X - Y$, $Y - X - Z$ (see figure below).



Exercise 8 Find a criteria for choosing the tree with the Maximal likelihood, i.e.:

$$\max_{Q_{X,Y,Z}: \{Y-X-Z \text{ or } Z-Y-X \text{ or } X-Z-Y\}} Q_{X,Y,Z}(x^n, y^n, z^n).$$