EE376A/STATS376A Information Theory Lecture 20 (Final Lecture) - 03/15/2018

Lecture 20: Conditional Differential Entropy, Info. Theory in ML

Lecturer: Tsachy Weissman *Scribe: Colleen Josephson, William Clary, Scott Ray*

1 The Chain Rule for Relative Entropy

1.1 Rule Statement

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

$$D(P_{X|Y}||Q_{X|Y}) \triangleq \sum_{y} D(P_{X|Y=y}||Q_{X|Y=y})P_Y(y) \tag{1}$$

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

$$D(P_{X,Y}||Q_{X,Y}) = D(P_X||Q_X) + D(P_{Y|X}||Q_{Y|X}|P_X) \tag{2}$$

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

$$D(P_{X_1,\ldots,X_n}||Q_{X_1,\ldots,X_n}) = \sum_{i=1}^{n} D(P_{X_i|X^{i-1}}||Q_{X_i|X^{i-1}}|P_{X^{i-1}}) \tag{3}$$

Where X^{i-1} denotes the variables $X_1, \ldots, X_{i-1}$.

Exercise 2. Prove the Chain Rule for Relative Entropy.

1.2 Applications of the Chain Rule

1. Rewriting Mutual Information:

$$I(X;Y) = D(P_{X,Y}||P_X \times P_Y) \tag{4}$$

$$= D(P_X||P_X) + D(P_{Y|X} + D(P_{Y|X}||P_Y|P_X) \tag{5}$$

$$= D(P_{Y|X}||P_Y|P_X) \tag{6}$$

Line 5 is obtained by applying the Chain Rule for Relative Entropy.

2. Minimizing Conditional Relative Entropy:

$$\min_{Q_Y} D(P_{Y|X}||Q_Y|P_X) = D(P_{Y|X}||P_Y|P_X) = I(X;Y) \tag{7}$$

Exercise 3. Prove that $\min_{Q_Y} D(P_{Y|X}||Q_Y|P_X) = D(P_{Y|X}||P_Y|P_X) = I(X;Y)$.

3. "Mutual Information" of three variables:

$$\begin{aligned} H(X)+H(Y)+H(Z)-H(X,Y,X) &= I(X;Y)+I(X;Z)+I(Y;Z|X) && (8)\\ &= I(Y;Z)+I(Y;X)+I(X;Z|Y) && (9)\\ &= I(Z;X)+I(Z;Y)+I(X;Y|Z) && (10)\end{aligned}$$

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

2 Method of types review

We briefly restate our notation and results from the method of types. For a specific n-tuple, $x^n = (x_1, x_2, ..., x_n) \quad x_i \in \mathcal{X}$, we denote the empirical distribution of x^n to be P_{x^n}.

For random variables $X_i \quad \text{iid} \sim Q_X$ the probability of a specific n-tuple can be written:

$$Q_X(x^n) = Pr(X^n = x^n) = 2^{-n[H(P_{x^n})+D(P_{x^n}\|Q_X)]} \tag{11}$$

With this notation we can ask for a particular sequence x_n what iid source Q_x maximizes the probability of that sequence:

$$\max_{Q_X} \quad Q_X(x^n) = 2^{-n[H(P_{x^n})+D(P_{x^n}\|P_{x^n})]} = 2^{-nH(P_{x^n})} \tag{12}$$

where the minimum is achieved by $Q_X = P_{x^n}$. Similarly for $(X_i, Y_i, Z_i) \sim \text{iid } Q_{X,Y,Z}$

$$Q_{X,Y,Z}(x^n, y^n, z^n) = Pr((X^n, Y^n, Z^n) = (x^n, y^n, z^n)) = 2^{-n[H(P_{x^n,y^n,z^n})+D(P_{x^n,y^n,z^n}||Q_{X,Y,Z})]} \tag{13}$$

where again P_{x^n,y^n,z^n} is the empirical joint distribution, i.e., $P_{x^n,y^n,z^n}(x,y,z)$ is the fraction of i's where $(x_i, y_i, z_i) = (x, y, z)$.

3 Tree Distributions

Definition 5. *A **tree** is an undirected graph with no cycles (loops). See Figure 1.*

A tree with nodes corresponding to random variables defines a conditional independence structure on the variables. Conditioned on any node, the subtrees on its edges are independent. For example, the tree in Figure 1 corresponds to $p(x)p(y|x)p(z|x)p(u|y)p(v|y)p(w|z)p(s|z)$. Note that a tree structure has much smaller number of parameters (linear in the number of nodes) as compared to exponentially many parameters for a general distribution.

A **Markov n-tuplet** $X - Y - Z - W$ is a special case of a tree (see Fig. 2)

Recall that a Markov Triplet $X - Y - Z$ has the property that X and Z are conditionally independent given Y. This means that $p(x,y|z) = p(x|y)p(z|y)$, $p(x|y,z) = p(x|y)$ and $p(z|y,x) = p(z|y)$. Intuitively, it means that X and Y are more closely related than X and Z. Furthermore, using data processing inequality and conditional independence:

1. $H(X|Y) = H(X|Y,Z)$
2. $H(Z|Y) = H(Z|X,Y)$
3. $H(X|Y) \leq H(X|Z)$
4. $I(X;Y) \geq I(X;Z)$
5. $I(Y;Z) \geq I(X;Z)$

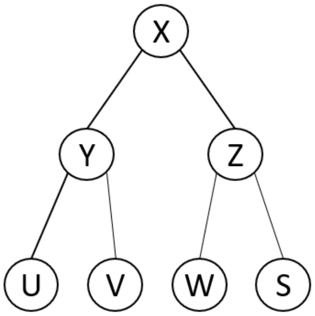


Figure 1: This tree corresponds to $P(x,y,z,u,v,w,s) = p(x)p(y|x)p(z|x)p(u|y)p(v|y)p(w|z)p(s|z)$

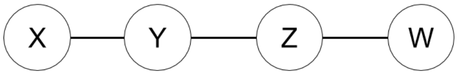


Figure 2: This tree corresponds to Markov 4-tuple $X - Y - Z - W$

3.1 Learning tree distributions

Suppose $(X_i, Y_i, Z_i) \sim$ iid $Q_{X,Y,Z}$. Recall from the previous lectures on the Method of Types and empirical distributions that $Q_{X,Y,Z}(x^n, y^n, z^n)$ (where x^n, y^n, z^n is a 3-tuple of n-tuples) is equal to

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

Question: In practice, Q is unknown. Among all $Q_{X,Y,Z}$ corresponding to the fixed tree $Y - X - Z$, which one maximizes $Q_{X,Y,Z}(z^n, y^n, z^n)$?

Answer: One might expect the answer to be $Q_{X,Y,Z} = P_{x^n,y^n,z^n}$, but it is not. That distribution does not necessarily respect the Markov graph. Instead, the best $Q_{X,Y,Z}$ that respects the tree has the same *marginals* as $P_{X,Y,Z}$.

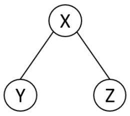
Proof: By the chain rule,

$$\min_{Q_{X,Y,Z:Y-X-Z}} D(P_{x^n,y^n,z^n}||Q_{X,Y,Z}) = \min D(P_X||Q_X) + D(P_{Y|X}||Q_{Y|X}|P_X) + D(P_{Z|Y,Z}||Q_{Z|Y,X}|P_{Y,X})$$

The first two terms can be made 0 by setting $Q_X = P_X$ and $Q_{Y|X} = P_{Y|X}$. For the third term, note that $Q_{Z|Y,X} = Q_{Z|X}$ by the Markov relation, and the minimum is obtained using a version of exercise 3 [1]:

$$\min_{Q_{Z|X}} D(P_{Z|X,Y} \parallel Q_{Z|X} \mid P_{X,Y}) = D(P_{Z|X,Y} \parallel P_{Z|X} \mid P_{X,Y}) = I(Z;Y \mid X) \quad (14)$$

We find that the best approximation of $P_{X,Y,Z}$ among $Q_{X,Y,Z}$ corresponding to



[1] Recall that $P_{Z|X} \neq P_{Z|X,Y}$

is the one with the same marginals as $P_{X,Y,Z}$, and in this case:

$$\max_{Q_{X,Y,Z}:Y-X-Z} \qquad Q_{X,Y,Z}(x^n, y^n, z^n) = 2^{-n[H(X,Y,Z)+I(Z;Y|X)]} \tag{15}$$

where $H(X,Y,Z)$ and $I(Z;Y \mid X)$ are with respect to

$$(X,Y,Z) \sim P(x^n, y^n, z^n) \quad \text{i.e. the empirical distribution of the data} \tag{16}$$

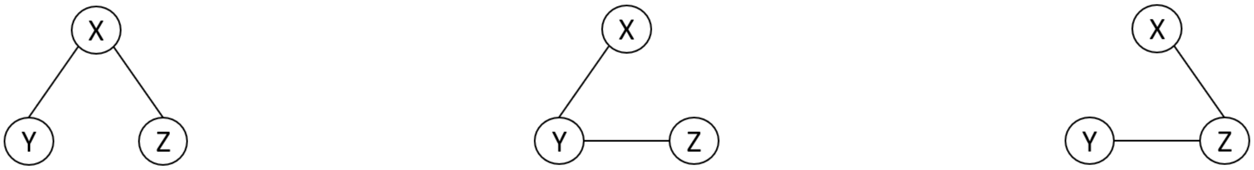
Using the results from exercise 4, specifically that

$$H(X,Y,Z) + I(Y;Z \mid X) = H(X) + H(Y) + H(Z) - I(X;Y) - I(X;Z) \tag{17}$$

We can reformulate this result in the following way:

$$2^{-n[H(X,Y,Z)+I(Z;Y|X)]} = 2^{-n[H(X)+H(Y)+H(Z)]}2^{n[I(X;Y)+I(X;Z)]} \tag{18}$$

In this reformulation the two mutual information terms $I(X;Y)$ and $I(X;Z)$ correspond to the edges of our three node tree. This allows us to answer the additional question where we constrain our $Q_{X,Y,Z}$ to come from some 3 node tree, but are unsure of exactly which one, i.e. each of the following trees are possible



We can now answer this problem as:

$$\begin{array}{ll} \max & Q_{X,Y,Z}(x^n, y^n, z^n) = 2^{-n[H(X)+H(Y)+H(Z)]}2^{n[\max\left\{\begin{matrix} I(X,Y)+I(X,Z) \\ I(X,Z)+I(Y,Z) \\ I(X,Y)+I(Y,Z) \end{matrix}\right\}]} \\ Q_{X,Y,Z}: & Y-X-Z \\ & \text{or} \\ & X-Y-Z \\ & \text{or} \\ & X-Z-Y \end{array}$$

From this we can conclude that among all tree models containing 3 nodes on (X,Y,Z) the likelihood of the data is maximized by the model which maximizes the sum of the weights on the edges (aka the maximal weight spanning tree) when the weight for a edge is given by the empirical mutual information between the variables corresponding to the two nodes the edge connects. For this maximal model, the probabilities and conditional probabilities between these random variables are given by the empirical distribution of the data. This method of selecting models from the data works not only for tree models with 3 variables, but also for tree models with any number of variables. One might ask why we could be interested in constraining our $Q_{X,Y,Z}$ to be a tree model. Constraining our Q in this way is akin to regularization in machine learning, and through limiting our model in this way we actually improve the how accurately the model coincides with new data. If no constraints are applied to $Q_{X,Y,Z}$, and allow for our $Q_{X,Y,Z}$ to simply be $P(x^n, y^n, z^n)$, this model would be too 'greedy', over-fitting the data and wouldn't generalize well to new data.

A statistician might ponder whether this method of selecting trees is effective by supposing that our data

$P(x^n, y^n, z^n)$ was truly generated by a distribution which is a tree model and asking if this method would choose the true tree which actually generated the data as n becomes large, i.e., is this method *consistent*. This answer of this question turns out to be yes. Suppose our three variables did come from a 3 node tree $X - Y - Z$, by the data processing inequality:

$$I(Y,Z) + I(X,Z) \leq I(X,Y) + I(Y,Z) \geq I(X,Y) + I(X,Z) \tag{19}$$

These are the true mutual informations. The quantity $I(X,Y) + I(Y,Z)$ corresponds to the true mutual information of the edges of the true tree. The quantities $I(X,Y) + I(X,Z)$ and $I(Y,Z) + I(X,Z)$ would correspond to false trees. From this we see that if X,Y,Z are truly governed by a tree distribution the maximal weight spanning tree, when the weights are the true mutual informations, is the true tree. This serves to confirm that our method of choosing the true tree by the empirical mutual information will with high probability (as the empirical mutual information gets closer and closer to the true mutual information as n increases) select the true tree.

3.2 Some further points on tree learning

Here we discuss a few points about learning tree distribution (also referred to as Chow-Liu [2] tree).

1. Since the number of possible trees on d nodes is d^{d-2} (Cayley's formula), it is not feasible to search for the maximum weight tree by a brute force search. However, as taught in any algorithms course, the maximum weight spanning tree can be found in time $\mathcal{O}(d \log d)$ using a cleverer algorithm such as Kruskal's algorithm or Prim's algorithm.

2. As we saw above using data processing inequality, if the data is truly governed by a tree distribution, Chow-Liu algorithm will recover the true tree with high probability as $n \to \infty$. But even if the true distribution is not a tree, it might be useful to use this algorithm to find the best tree approximation to the data, which provides regularization. In practice, the class of tree distributions is usually rich enough to capture structure and dependence among variables while not being too rich so as to overfit the data. The best approximation of the empirical distribution is the empirical distribution itself, but that is unlikely to generalize well especially in high dimension settings, where the amount of data might not be much larger than the number of variables.

3. We talked about how the empirical mutual information can be thought of as estimating the true mutual information. It turns out that there are better mutual information estimators than the empirical mutual information, and hence Chow-Liu tree performance can be boosted by replacing the empirical mutual information by these estimators (see this link for an example). For more details on how this can improve the performance in practice, we refer you to this paper [3].

4. Tree learning can be used for machine learning tasks such as classification. Suppose we wish to predict a label Y given a feature vector $\mathbf{X}$. Then, we can learn a tree distribution $\mathbf{X}|Y = y$ for each label y. For prediction, we can use Bayes rule to estimate $P(Y = y|\mathbf{X})$ and maximize this to classify new data points.

5. Tree learning can also be used for data compression [4]. Suppose we have tabular data where each row is an independent sample while the columns within each row might be correlated. Then we can learn a tree over the columns and then use something like arithmetic coding based on the conditional distributions on the tree edges.

[2] Chow, C., and Cong Liu. "Approximating discrete probability distributions with dependence trees." *IEEE transactions on Information Theory* 14.3 (1968): 462-467.

[3] Jiao, Jiantao, Yanjun Han, and Tsachy Weissman. "Beyond maximum likelihood: Boosting the Chow-Liu algorithm for large alphabets." *Signals, Systems and Computers, 2016 50th Asilomar Conference on.* IEEE, 2016.

[4] Pavlichin, Dmitri S., Amir Ingber, and Tsachy Weissman. "Compressing Tabular Data via Pairwise Dependencies." *Data Compression Conference (DCC), 2017.* IEEE, 2017.