208891 Probabilistic Graphical Models **Spring 2019**

Lecture 13: February 14

Lecturer: Donlapark Pornnopparath

13.1 Basic information theory

Information theory dated back in 1948 in Shannon's seminal *A Mathemaical Theory of Communication*. Here, *information* means an insight, or an additional knowledge that a receiver gains after receiving a signal from a sender. For example, we can think of signals as a sequence of words from an article, and the information as the main context of that article. We know that not all words give the same amount of information: the word "eat" is not as meaningful as the word "barbecue". From this observation, we see that the information the receiver gains from each word is not the content of word itself but *the probability of obtaining that word*. Therefore, the function h that describes information should satisfy

1. $h(x)$ is a decreasing function of $p(x)$.

Moreover, the information that we gain should be something additive, that is

2. If X and Y are independent, then $h(x, y) = h(x) + h(y)$.

Since $p(x, y) = p(x)p(y)$, we can show that h must be in the form of $-C \log p(x)$ for some constant $C > 0$. Since the choice of C specifies the base of the logarithm, we can simply write

$$h(x) = -\log p(x).$$

We can think of this as a "surprise" function. The word "barbecue" contains more surprise elements than "eat", and so its value of h should be higher.

After sending a random sequence of words to the receiver, the average amount of information can be computed by taking the expected value.

Definition 13.1 (Entropy)**.** The entropy of a random variable X is given by

$$H(X) = -\sum_x p(x) \log p(x) = \mathbb{E}[-\log p(X)].$$

There are two seemingly contradictory notions of entropy: average information and *uncertainly*. These interpretation can coexist if we treat them at different point in time of knowing X; The entropy measures the average uncertainty *before receiving* X was sent out, which is equivalent to the average information gained *after receiving* X. To illustrate this point, we consider a simple coin toss. If the probability of head was 0.99, then before tossing we would be certain that it would turn up head. Alternatively, as we partially knew that the coin would most likely turn up head, each coin result would give us less elements of surprise i.e. less information.

With this intuition, it is easy to see that any deterministic variable has the lowest entropy and a uniform distribution has the highest entropy. Formally, we have

Proposition 13.2. Let X be a discrete random variable with n possible states. The entropy of a random variable X satisfies

$$0 \leq H(X) \leq \log n$$

The maximum is attained when X is uniformly distributed.

We also have an analogous definition of entropy for continuous distributions.

Definition 13.3 (Entropy (continuous version))**.** The entropy of a continuous variable X is

$$-\int p(x) \log p(x)\ dx.$$

Specifically, when we consider the joint probability of two random variables X and Y,

$$H(X,Y) = -\int\int p(x,y) \log p(x,y)\ dxdy.$$

In particular, H also exhibits logarithmic behavior; if X and Y are independent, then

$$H(X,Y) = H(X) + H(Y)$$

Suppose that the value of X is already known, then what we want to compute is the additional information that we can gain from Y after observing X, denoted by $H(Y|X)$. We derive a formula for $H(Y|X)$ from our intuition that

$$\begin{aligned} H(Y|X) &= H(X,Y) - H(X) \\ &= -\int\int p(x,y) \left[\log p(x,y) - \log p(x)\right]\ dxdy \\ &= -\int\int p(x,y) \log p(y|x)\ dxdy \\ &= \mathbb{E}_{X,Y}[-\log p(Y|X)]. \end{aligned}$$

The *Bayes' rule for entropy* can be easily derived from the decomposition of joint entropy.

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

In general, we have the *chain rule of entropies*:

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_n|X_{n-1}, \ldots, X_1).$$

Consider the information diagram of X and Y in Figure 13.1. The only region that have not been discussed yet is the intersection, which corresponds to the part of the information of one variable that can be described by other variable. We call this part the *mutual information*.

Definition 13.4 (Mutual information)**.** The mutual information between two random variables X and Y is given by

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

By the symmetry, we have that $I(X,Y) = I(Y,X)$. In terms of the original distributions, we have

$$\begin{aligned} I(X,Y) &= -\int\int p(x,y)\left[\log p(x) - \log p(x)\right]\, dxdy \\ &= -\int\int p(x,y)\log\frac{p(x)p(y)}{p(x,y)}\, dxdy. \end{aligned}$$

13.2 Kullback-Leibler divergence

Suppose that we sample from an unknown distribution $p(x)$ and obtain an empirical distribution $q(x)$ then send codes to the receiver based on $q(x)$. The Kullback-Leiber divergence (Kullback and Leibler, 1951) measures an additional amount of information required to send X in order to match the information obtained by just using p.

$$\begin{aligned} D_{KL}(p\|q) &= -\int p(x)\log q(x)dx - \left(-\int p(x)\log p(x)\right) \\ &= -\int p(x)\log\frac{q(x)}{p(x)}\, dx \end{aligned}$$

Since $-\log(\cdot)$ is convex, Jensen's inequality yields

$$D_{KL}(p\|q) = -\int p(x)\log\frac{q(x)}{p(x)}\, dx \geq -\log\int q(x)\, dx = 0.$$

Thus the KL-divergence can be used as a distance function. However, one must be careful as it is not symmetric: $D_{KL}(p\|q) \neq D_{KL}(q\|p)$.

One can use KL-divergence to find an appropriate set of parameters $\boldsymbol{\theta}$ from a given set of data. Suppose that we are trying to approximate an unknown distribution $p(x)$ with some parametric distribution $q(x|\boldsymbol{\theta})$. This can be done by minimizing the KL-divergence. Suppose that we have a sample of N data points, then we can use the empirical approximation.

$$\begin{aligned} D_{KL}(p\|q) &= \int p(x)\log\frac{q(x|\boldsymbol{\theta})}{p(x)}\, dx \\ &= \mathbb{E}_X\left[-\log\frac{q(X|\boldsymbol{\theta})}{p(X)}\right] \\ &\approx \frac{1}{N}\sum_{i=1}^{N} -\log q(x_i|\boldsymbol{\theta}) + \log p(x_i). \end{aligned}$$

Note that the second term on the right-hand side does not depend on $\boldsymbol{\theta}$, and so it remains to minimize the first term. Thus the minimizing the KL-divergence is equivalent to the likelihood maximization problem.