Machine Learning - Michaelmas Term 2017
Notes : Information Theory Basics

Lecturers: Christoph Haase & Varun Kanade

1 Information, Entropy, KL Divergence

We'll briefly discuss the connections between some of the concepts introduced in this course to those in information theory. Obviously, give that the goal of machine learning is to extract meaninful patterns out of data, it is no surprise that there are deep connections between machine learning and information theory. Exploring these in detail is beyond the scope of this course, but the interested student may refer to the book by MacKay (2003) or Jaynes (2003).

1.1 Entropy

Let X be a random variable that takes values from a finite set according to distribution p.[1] Then then entropy of X is defined as

$$H(X) = -\sum_{x} p(x) \log p(x) \tag{1}$$

The entropy is a measure of uncertainty of a random variable. If X takes values over a finite set of size n, then X has maximum entropy if it is distributed according to the uniform distribution over these n elements. It has minimum entropy if all the probability mass is concentrated on one of these elements, *i.e.,* in effect it is not a random variable at all, but a constant.

Let us focus on the case of Bernoulli random variables. A Bernoulli random variable is defined by a parameter $\theta \in [0, 1]$ and takes value 1 with probability θ and 0 with probability $1 - \theta$. This can be expressed succintly as

$$p(x \mid \theta) = \theta^x (1 - \theta)^{1-x}$$

In this case, let us write the entropy in terms of the parameter θ and use logarithm base 2 for convenience.

$$H(X) = -\theta \log_2(\theta) - (1 - \theta) \log_2(1 - \theta)$$

Figure 1 plots the entropy as a function of θ. We see that the entropy has a maximum value of 1 for $\theta = 1/2$ and minimum value of 0 at $\theta \in \{0, 1\}$. One way to think of entropy is how much information is obtained when the outcome of an experiment is revealed. For example, if Alice has an unbiased coin, then if she tosses it and reports the outcome we get one *bit* of information. On the other hand if she has a coin that always lands on heads, we get no additional information by being told the outcome of the coin toss, because it was something we could have predicted ourselves with complete certainty!

1.2 Kullback-Leibler Divergence

Let p and q be distributions over some finite set and suppose that the support of p is contained in the suport of q.[1] The Kullback-Leibler (or KL) Divergence between two distributions p and

[1] This can be extended to continuous-valued random variables by using the integral instead of the sum and replacing the probability mass function by the density function.

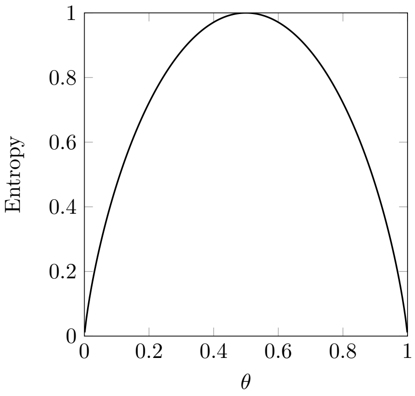

Figure 1: Entropy of the Bernoulli random variable as a function of θ

q is defined as follows

$$\begin{aligned}\mathrm{KL}(p\|q) &= \sum_x p(x)\log\left(\frac{p(x)}{q(x)}\right)\\ &= \sum_x p(x)\log(p(x)) - \sum_x p(x)\log(q(x)) = -H(p) + H(p,q)\end{aligned}$$

Here $H(p) = -\sum_x p(x)\log p(x)$ is the entropy of the distribution p and $H(p,q) = -\sum_x p(x)\log q(x)$ is called the *cross-entropy*. The cross entropy accounts of the expected number of bits required to encode an observation from p if the encoding scheme was based on q. Thus, the KL-divergence $\mathrm{KL}(p\|q)$ gives the expected *excess bits* required to encode an observation from p if the encoding scheme was based on q.

The KL divergence satisfies the following two properties:

1. $\mathrm{KL}(p\|q) \geq 0$

2. $\mathrm{KL}(p\|q) = 0$ if and only if $p = q$

It is worth mentioning that the KL-divergence is not a distance; in particular, it is not symmetric. For example, even when the support of p and q is the same, so that both $\mathrm{KL}(p\|q)$ and $\mathrm{KL}(q\|p)$ are defined, they need not be equal.

Relation to Maximum Likelihood

Let us now see how the maximum likelihood estimate relates to these notions from information theory. Suppose we get data $x_1, \ldots x_N$ from some unknown distribution p (not necessarily of any particular parametric form). However, we wish to fit a distribution that does have some parametric form (say for example Gaussian) that best explains the data. In particular, we will derive the maximum likelihood estimate for the parameters of distributions of a certain parametric form.

Figure 2 shows the actual generating distribution (in thick red). It also shows three possible Gaussian distributions with different means and variances (dotted). Suppose, we want to find maximum likelihood estimate for these parameters.

The mathematical derivation below is more general. It just assumes that the family of distributions we consider are parameterized by some parameters θ. In particular, $q(\cdot \mid \theta)$ is the distribution that we use to model the data and we derive the maximum likelihood estimate for θ.

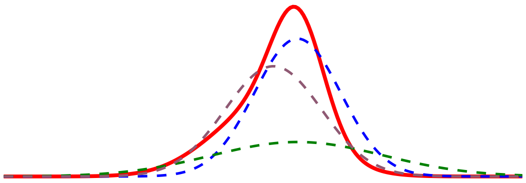

Figure 2: Maximum Likelihood Estimates and KL-divergence: The data is generated according to the distribution shown by the thick red line. The figure also shows three possible Gaussian distributions (dashed). The goal is to find the Gaussian distribution that maximises the likelihood of the observed data.

$$\begin{aligned}
\widehat{\theta}_{\mathrm{ML}} &= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} q(x_i \mid \theta) \\
&= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log(q(x_i \mid \theta)) \\
&= \underset{\theta}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \log(q(x_i \mid \theta)) - \frac{1}{N} \sum_{i=1}^{N} \log(p(x_i)) \qquad (2) \\
&= \underset{\theta}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \log\left(\frac{p(x_i)}{q(x_i \mid \theta)}\right) \qquad (3) \\
&\xrightarrow[N\to\infty]{} \underset{\theta}{\operatorname{argmin}} \int \log\left(\frac{p(x)}{q(x|\theta)}\right) p(x)dx = \mathrm{KL}(p\|q_\theta) \qquad (4)
\end{aligned}$$

Above in Step (2) we replace the sum by the average and added an extra term that does not depend on θ, neither of these operations affects the argmax; in Step (3), we switched the signs and hence changed the argmax to argmin; finally, Step (4) states that in the limit of getting infinite quantities of data, where $x_i \sim p$, the average can be replaced by the expectation under the distribution p. This last term is nothing but the KL-divergence $\mathrm{KL}(p\|q_\theta)$. Thus, the maximum likelihood estimate can be viewed as finding parameters (from some family of distributions) that minimises the KL-divergence between the true distribution generating the data and the modelled distribution from this family. Alternatively, the MLE can be viewed as finding the distribution from a parametric family that has least KL-divergence between the empirical distribution over the data and this particular parametric distribution.

Remark 1. *This section covered somewhat advanced topics and is not examinable. It is introduced to show connections between machine learning methods and information theory.*

References

Edwin T Jaynes. *Probability theory: The logic of science.* Cambridge university press, 2003.

David JC MacKay. *Information theory, inference and learning algorithms.* Cambridge university press, 2003.