Notes on prml

Hao Su

July 17, 2025

Contents

Chapter 1

Introduction

1.6 Information Theory

We begin by considering a discrete random variable x and we ask how much information is received when we observe a specific value for this variable, which can be viewed as the 'degree of surprise' on learning the value of x. Our measure of information content will therefore depend on the probability distribution $p(x)$, and we therefore look for a quantity $h(x)$ that is a monotonic function of the probability $p(x)$ and that expresses the information content. The form of h can be found by noting that if we have two events x and y that are unrelated, then the information gain from observing both of them should be the sum of the information gained from each of them separately, so that $h(x, y) = h(x)+h(y)$. Two unrelated events will be statistically independent and so $p(x, y) = p(x)p(y)$. From these two relationships, it is easily shown that $h(x)$ must be given by the logarithm of $p(x)$ and so we have

$$h(x) = -\log_2 p(x)$$

where the choice of basis is arbitrary. Now suppose that a sender wishes to transmit the value of a random variable to a receiver. The average amount of information that they transmit in the process is obtained by taking expectation with respect to the distribution $p(x)$, given by

$$\mathrm{H}[x] = -\sum_x p(x) \log_2 p(x),$$

which is called the *entropy* of x. When using 2 as the basis, $\mathrm{H}[x]$ is indeed the lower bound on the number of *bits* needed to *encode* the state of x.

To gain an alternative view of entropy, consider a set of N distinguishable objects that are to be divided amongst a set of bins, such that there are n_i objects in the i^{th} bin. The number of different ways of allocation is $W = \dfrac{N!}{\prod_i n_i!}$ and is called *multiplicity*. The entropy is then defined

$$\mathrm{H} = \frac{1}{N} \ln W = \frac{1}{N} \ln N! - \frac{1}{N} \sum_i \ln n_i!.$$

Consider the limit $N \to \infty$, in which the fractions n_i/N are held fixed, and apply Stirling's approximation

$$\ln N! \simeq N \ln N - N,$$

then we have

$$\begin{aligned}
\mathrm{H} &= \frac{1}{N} \left\{ (N \ln N - N) - \left(\sum_i (n_i \ln n_i - n_i) \right) \right\} \\
&= \frac{1}{N} \left\{ \left(\sum_i n_i \right) \ln N - \left(\sum_i (n_i \ln n_i) \right) \right\} \\
&= \frac{1}{N} \left(\sum_i n_i (\ln N - \ln n_i) \right) \\
&= -\sum_i \left(\frac{n_i}{N} \right) \ln \left(\frac{n_i}{N} \right) = -\sum_i p_i \ln p_i
\end{aligned}$$

where $p_i = \lim_{N \to \infty} (n_i/N)$ is the probability of an object being assigned to the i^{th} bin. A specific arrangement is called a *microstate* and the overall distribution of arrangements, expressed through p_i, is called a *macrostate*. Think of the objects as particles and bins as states to see that entropy can be viewed as a measure of disorder of a macrostate.

We extend the definition of entropy to include distributions $p(x)$ over continuous variables x as follows. First divide x into bins of width Δ. Then, assuming $p(x)$ is continuous, the *mean value theorem* (Weisstein, 1999) tells us that, for each such bin, there must exist a value x_i such that

$$\int_{i\Delta}^{(i+1)\Delta} p(x)\, dx = p(x_i)\Delta.$$

This gives a discrete distribution for which the entropy takes the form

$$\begin{aligned} H_\Delta &= -\sum_i p(x_i)\Delta \ln\left(p(x_i)\Delta\right) \\ &= -\sum_i p(x_i)\Delta \ln p(x_i) - \ln \Delta. \end{aligned}$$

Omit the second term $-\ln\Delta$ to optain

$$\lim_{\Delta\to 0}\left\{\sum_i p(x_i)\Delta \ln p(x_i)\right\} = -\int p(x)\ln p(x)\,dx.$$

where the quantity on the right-hand side is called the *differential entropy*. We see that the discrete and continuous forms of the entropy differ by a quantity $\ln\Delta$. This reflects the fact that to specify a continuous variable very precisely requires a large number of bits. For a density defined over multiple continuous variables, denoted collectively by the vector $\mathbf{x}$, the differential entropy is given by

$$H[\mathbf{x}] = -\int p(\mathbf{x})\ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

Given standard deviation σ, the distribution that maximizes the differential entropy is the Gaussian, given by

$$\mathrm{H}[x] = \frac{1}{2}\left\{1 + \ln(2\pi\sigma^2)\right\}.$$

This maximization can be performed using Lagrange multipliers. Thus a broader Gaussian has a higher entropy. And the differencial entropy can be negative (which implies nothing specifically.)

The intuition of 'information' works as expected for joint distributions. That is,

$$\begin{aligned} \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}] &= -\iint p(\mathbf{y},\mathbf{x})\ln p(\mathbf{y}|\mathbf{x})\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} - \iint p(\mathbf{y},\mathbf{x})\ln p(\mathbf{x})\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} \\ &= -\iint p(\mathbf{y},\mathbf{x})\ln p(\mathbf{y},\mathbf{x})\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} = \mathrm{H}[\mathbf{y},\mathbf{x}] \end{aligned}$$

where $\mathrm{H}[\mathbf{x},\mathbf{y}]$ is the differential entropy of $p(\mathbf{x},\mathbf{y})$ and $\mathrm{H}[\mathbf{x}]$ is the differential entropy of the marginal distribution $p(\mathbf{x}) = \int p(\mathbf{y},\mathbf{x})\mathrm{d}\mathbf{y}$.

Consider some unknown distribution $p(\mathbf{x})$, and suppose that we have modelled this using an approximating distribution $q(\mathbf{x})$. If we use $q(\mathbf{x})$ to construct a coding scheme for the purpose of transmitting values of $\mathbf{x}$ to a receiver, then the average *additional* amount of information (in nats) required to specify the value of $\mathbf{x}$ (assuming we choose an efficient coding scheme) as a result of using $q(\mathbf{x})$ instead of the true distribution $p(\mathbf{x})$ is given by

$$\begin{aligned}\mathrm{KL}(p \parallel q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x}) \, \mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \right) \\ &= -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} \mathrm{d}\mathbf{x}.\end{aligned}$$

This is known as the *relative entropy* or *Kullback–Leibler divergence*, or *KL divergence* (Kullback and Leibler, 1951), between the distributions $p(\mathbf{x})$ and $q(\mathbf{x})$. Note that it is not a symmetrical quantity, that is, $\mathrm{KL}(p \parallel q) \not\equiv \mathrm{KL}(q \parallel p)$.

Consider *Jensen's inequility* in the following form:

$$f\left(\mathbb{E}[x]\right) \leq \mathbb{E}[f(x)]$$

where $f(x)$ is a convex function. Taking $f(x) = -\ln x$ gives us

$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} \mathrm{d}\mathbf{x} \\ &= \mathbb{E}\left[f\left(\frac{q(\mathbf{x})}{p(\mathbf{x})} \right) \right] \geq f\left(\mathbb{E}\left[\frac{q(\mathbf{x})}{p(\mathbf{x})} \right] \right) \\ &= -\ln \int q(\mathbf{x}) \, \mathrm{d}\mathbf{x} = 0\end{aligned}$$

where the equality holds iff $q(\mathbf{x}) = p(\mathbf{x})$ for all $\mathbf{x}$ for that $-\ln x$ is strictly convex. Thus KL divergence can be thought of as a measure of the dissimilarity of two distributions.

Suppose that data is being generated from an unknown distribution $p(\mathbf{x})$ that we wish to approximate using some parametric distribution $q(\mathbf{x}|\theta)$. One way to determine θ is to minimize the KL divergence between $p(\mathbf{x})$ and $q(\mathbf{x}|\theta)$ with respect to θ. Suppose that we have observed a finite set of training points $\mathbf{x}_n$, for $n = 1, \ldots, N$, drawn from $p(\mathbf{x})$. Then the expectation with respect to $p(\mathbf{x})$ can be approximated by

$$\mathrm{KL}(p\|q) \simeq \frac{1}{N} \sum_{n=1}^{N} \left\{ -\ln q(\mathbf{x}_n|\theta) + \ln p(\mathbf{x}_n) \right\}.$$

The second term on the right-hand side is independent of θ, and the first term is the negative log likelihood function for θ under the distribution $q(\mathbf{x}|\theta)$ evaluated using the training set. Thus we see that minimizing this KL divergence is equivalent to maximizing the likelihood function.

Now consider the joint distribution $p(\mathbf{x}, \mathbf{y})$. Intuitively we can gain some idea of whether they are 'close' to being independent by considering the KL divergence between the joint distribution and the product of the marginals, given by

$$\begin{aligned} \mathrm{I}[\mathbf{x}, \mathbf{y}] &\equiv \mathrm{KL}\left(p(\mathbf{x}, \mathbf{y}) \,\|\, p(\mathbf{x})p(\mathbf{y})\right) \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln\left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})}\right) \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}. \end{aligned}$$

This is called the *mutual information* between the variables $\mathbf{x}$ and $\mathbf{y}$. From the properties of the KL divergence, we see that $I(\mathbf{x}, \mathbf{y}) \geq 0$ with equality iff $\mathbf{x}$ and $\mathbf{y}$ are independent.

The mutual information is related to the conditional entropy through

$$\mathrm{I}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}].$$

From a Bayesian perspective, the mutual information therefore represents the reduction in uncertainty about $\mathbf{x}$ as a consequence of the new observation $\mathbf{y}$.

Chapter 3

Linear Models for Regression

A helpful insight is that when

$$Ax = b$$

is irresolvable (A being not invertible, etc), we turn to $A^{\dagger}$ instead to achieve an optimal result in the sense that $\|Ax - b\|$ is minimized. This is a motivation for *Moore-Penrose pseudo-inverse* matrix.