?

1 intro

1.1 some formula

Let's derive Stirling's approximation by an unconventional route. We start from the Poisson distribution with mean λ,

$$P(r \mid \lambda) = e^{-\lambda} \frac{\lambda^r}{r!} \quad r \in \{0, 1, 2, \ldots\}. \tag{1.8}$$

For large λ, this distribution is well approximated – at least in the vicinity of $r \simeq \lambda$ – by a Gaussian distribution with mean λ and variance λ:

$$e^{-\lambda} \frac{\lambda^r}{r!} \simeq \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{(r-\lambda)^2}{2\lambda}}. \tag{1.9}$$

Let's plug $r = \lambda$ into this formula, then rearrange it.

$$e^{-\lambda} \frac{\lambda^\lambda}{\lambda!} \simeq \frac{1}{\sqrt{2\pi\lambda}} \tag{1.10}$$

$$\Rightarrow \quad \lambda! \simeq \lambda^\lambda e^{-\lambda} \sqrt{2\pi\lambda}. \tag{1.11}$$

This is Stirling's approximation for the factorial function.

$$x! \simeq x^x e^{-x} \sqrt{2\pi x} \quad \Leftrightarrow \quad \ln x! \simeq x \ln x - x + \tfrac{1}{2} \ln 2\pi x. \tag{1.12}$$

We have derived not only the leading order behaviour, $x! \simeq x^x e^{-x}$, but also, at no cost, the next-order correction term $\sqrt{2\pi x}$. We now apply Stirling's approximation to $\ln \binom{N}{r}$:

$$\ln \binom{N}{r} \equiv \ln \frac{N!}{(N-r)!\, r!} \simeq (N-r) \ln \frac{N}{N-r} + r \ln \frac{N}{r}. \tag{1.13}$$

Since all the terms in this equation are logarithms, this result can be rewritten in any base. We will denote natural logarithms ($\log_e$) by 'ln', and logarithms to base 2 ($\log_2$) by 'log'.

If we introduce the *binary entropy function*,

$$H_2(x) \equiv x \log \frac{1}{x} + (1-x) \log \frac{1}{(1-x)}, \tag{1.14}$$

then we can rewrite the approximation (1.13) as

$$\log \binom{N}{r} \simeq N H_2(r/N), \tag{1.15}$$

or, equivalently,

$$\binom{N}{r} \simeq 2^{N H_2(r/N)}. \tag{1.16}$$

If we need a more accurate approximation, we can include terms of the next order from Stirling's approximation (1.12):

$$\log \binom{N}{r} \simeq N H_2(r/N) - \tfrac{1}{2} \log \left[2\pi N \frac{N-r}{N} \frac{r}{N} \right]. \tag{1.17}$$

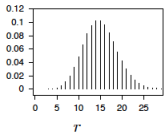

Figure 1.2. The Poisson distribution $P(r \mid \lambda = 15)$.

Recall that $\log_2 x = \frac{\log_e x}{\log_e 2}$.

Note that $\frac{\partial \log_2 x}{\partial x} = \frac{1}{\log_e 2} \frac{1}{x}$.

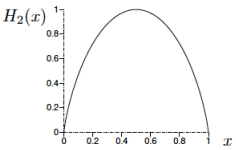

Figure 1.3. The binary entropy function.

Figure 1: Some formula

1.2 assumptions in inference

First, once assumptions are made, the inferences are objective and unique, reproducible with complete agreement by anyone who has the same informa-

tion and makes the same assumptions. For example, given the assumptions listed above, $\mathcal{H}$, and the data D, everyone will agree about the posterior probability of the decay length

$$P(\lambda|D,\mathcal{H}) = \frac{P(D|\lambda,\mathcal{H})P(\lambda|\mathcal{H})}{P(D|\mathcal{H})}$$

Second, when the assumptions are explicit, they are easier to criticize, and easier to modify indeed, we can quantify the sensitivity of our inferences to the details of the assumptions.

Third, when we are not sure which of various alternative assumptions is the most appropriate for a problem, we can treat this question as another inference task. Thus, given data D, we can compare alternative assumptions $\mathcal{H}$ using Bayes' theorem

$$P(\mathcal{H}|D,I) = \frac{P(D|\mathcal{H},I)P(\mathcal{H}|I)}{P(D|I)},$$

where I denotes the highest assumptions, which we are not questioning.

Fourth, we can take into account our uncertainty regarding such assumptions when we make subsequent predictions. Rather than choosing one particular assumption $\mathcal{H}^*$, and working out our predictions about some quantity t, $P(t|D,\mathcal{H},I)$, we obtain predictions that take into account our uncertainty about H by using the sum rule

$$P(\mathbf{t}|D,I) = \Sigma_{\mathcal{H}} P(\mathbf{t}|D,\mathcal{H},I)P(\mathcal{H}|D,I)$$

This is another contrast with orthodox statistics, in which it is conventional to 'test' a default model, and then, if the test 'accepts the model' at some 'significance level', to use exclusively that model to make predictions. *probability theory reaches parts that ad hoc methods cannot reach.*

Model comparison as inference. Assume we have two hypotheses. In order to perform model comparison, We wish to know how probable $P(\mathcal{H}_1)$ is given the data. By Bayes' theorem,

$$P(\mathcal{H}_1|\mathbf{s},F) = \frac{P(s|F,\mathcal{H}_1)P(\mathcal{H}_1)}{P(\mathbf{s}|F)},$$

and

$$P(\mathcal{H}_0|\mathbf{s},F) = \frac{P(\mathbf{s}|F,\mathcal{H}_0)P(\mathcal{H}_0)}{P(\mathbf{s}|F)}$$

The normalizing constant in both cases is the total probability of geting the observed data. and

$$P(s|F) = P(\mathbf{s}|F,\mathcal{H}_1)P(\mathcal{H}_1) + P(\mathbf{s}|F,\mathcal{H}_0)P(\mathcal{H}_0)$$

To evaluate the posterior probabilities of the hypotheses we need to assign values to the prior probabilities $P(\mathcal{H}_1)$ and $P(\mathcal{H}_0)$; in this case, we

might set these to 1/2 each. And we need to evaluate the data-dependent terms $P(s|F,\mathcal{H}_1)$ and $P(s|F,\mathcal{H}_0)$. We can give names to these quantities. The quantity $P(s|F,\mathcal{H}_1)$ is a measure of how much the data favour $\mathcal{H}_1$), and we call it the evidence for model $\mathcal{H}_1$). *How model comparison works : The evidence for a model is usually the normalizing constant of an earlier Bayesian inference.*

2 Clustering

2.1 Motivation of clustering

First, a good clustering has predictive power. Second, clusters can be a useful aid to communication because they allow lossy compression. A third reason for making a cluster model is that failures of the cluster model may highlight interesting objects that deserve special attention. A fourth reason for liking clustering algorithms is that they may serve as models of learning processes in neural systems.

2.2 K-means

The K-means algorithm is an algorithm for putting N data points in an M-dimensional space into K clusters. Each cluster is parameterized by a vector $\mathbf{m}^{(k)}$ called its mean.

First of all, set K means $\mathbf{m}^{(k)}$ to random values. In the assignment step, each data point n is assigned to the nearest mean.

$$\hat{k}^{(n)} = argmin_k d(\mathbf{m}^{(k)}, \mathbf{x}^{(n)})$$

An alternative, equivalent representation of this assignment of points to clusters is given by 'responsibilities', which are indicator variables $r_k^{(n)}$. In the assignment step, we set $r_k^{(n)}$ to one if mean k is the closest mean to datapoint $\mathbf{x}^{(n)}$; otherwise, $r_k^{(n)}$ is zero.

$$r_k^{(n)} = \{ \begin{array}{ll} 1 & if\ \hat{k}^{(n)} = k \\ 0 & if\ \hat{k}^{(n)} \neq k \end{array}$$

In the update step, the means are adjusted to match the sample means of the data points that they are responsible for. The update step is very similar to how to find the center of the mass in physics.

$$\mathbf{m}^{(k)} = \frac{\Sigma_n\ r_k^{(n)} \mathbf{x}^{(n)}}{R^{(k)}}$$

where $R^{(k)}$ is the total responsibility of mean k

$$R^{(k)} = \Sigma_n\ r_k^{(n)}$$

2.3 Exercise 22.5

$$P(k_n = 1|x_n, \boldsymbol{\theta}) = \frac{P(x_n|k_n = 1, \boldsymbol{\theta})P(k_n = 1, \boldsymbol{\theta})}{P(x_n, \boldsymbol{\theta})}$$

$$= \frac{P(x_n|k_n = 1, \boldsymbol{\theta})P(k_n = 1, \boldsymbol{\theta})}{\Sigma_{k_n} P(x_n|k_n = 1, \boldsymbol{\theta})P(k_n = 1, \boldsymbol{\theta}) + P(x_n|k_n = 2, \boldsymbol{\theta})P(k_n = 2, \boldsymbol{\theta})}$$

where $\boldsymbol{\theta} = \mu_k,\ \sigma_k$.

$$P(k_n = 1, \boldsymbol{\theta}) \equiv p_1, \quad P(k_n = 2, \boldsymbol{\theta}) \equiv p_2$$

Then,

$$P(k_n = 1|x_n, \boldsymbol{\theta}) = \frac{p_1}{p_1 + p_2 \exp[-(w_1 x_n + w0)]}$$

$$P(k_n = 2|x_n, \boldsymbol{\theta}) = \frac{p_2}{p_2 + p_1 \exp[-(w_1 x_n + w0)]}$$

where $w_1 = 2(\mu_1 - \mu_2)$, $w_0 = -(\mu_1 - \mu_2)(\mu_1 + \mu_2)$

$$P(k_n = k|x_n, \boldsymbol{\theta}) \equiv p_{k|n}$$

By assumption, the prior probability $p_1 = p_2 = 1/2$ then, (22.17) of the book is satisfied.

$$L \equiv \log \Pi_n P(x_n|\{\mu_k\},\ \sigma)$$

$$\frac{\partial}{\partial \mu_k} L = \Sigma_n \frac{p_{k|n}(x_n - \mu_k)}{\sigma^2}$$

3 Monte Carlo