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):

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

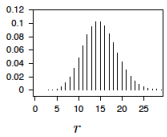

Figure 1.2. The Poisson distribution $P(r \mid \lambda = 15)$.

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

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

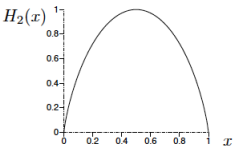

Figure 1.3. The binary entropy function.

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 information 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

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

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 and Maximum Likelihood

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.

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

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

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

2.3 Exercise 22.5

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

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

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.

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

then trivially,

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

$$\frac{\partial^2}{\partial {\mu_k}^2} L = -\sum_n \frac{p_{k|n}}{\sigma^2}$$

The new updated $\boldsymbol{\mu}'$ should maximize the likelihood. Then,

$$\frac{\partial}{\partial \mu'_k} L = \sum_n \frac{p_{k|n}(x_n - \mu'_k)}{\sigma^2} = 0$$

$$\sum_n p_{k|n}\ x_n - \sum_n p_{k|n}\ \mu'_k = \sum_n p_{k|n}\ x_n - \mu'_k \sum_n p_{k|n} = 0$$

Therefore,

$$\mu'_k = \frac{\sum_n p_{k|n}\ x_n}{\sum_n p_{k|n}}$$

Note that this equation is exactly the same as the updated means from the responsibilities and data points in soft K-means clustering. $p_{k|n}$ is the responsibility $r_n^{(k)}$.

$$\frac{\frac{\partial}{\partial \mu_k} L}{\frac{\partial^2}{\partial {\mu_k}^2} L} = \frac{\sum_n p_{k|n}\ x_n - \mu_k \sum_n p_{k|n}}{-\sum_n p_{k|n}} = -\mu'_k + \mu_k$$

Thus,

$$\mu'_k = \mu_k - \frac{\frac{\partial}{\partial \mu_k} L}{\frac{\partial^2}{\partial {\mu_k}^2} L}$$

2.4 Exercise 22.15

$$N = 7$$

$$x_n = (-27.02, 3.57, 8.191, 9.898, 9.603, 9.945, 10.056)$$

$$\frac{\sum_n x_n}{N} = 3.46329$$

It must not be the correct mean.

The likelihood is as followings by the description of the problem.

$$P(\{x_n\}|\sigma_n, \mu) = (\frac{1}{2\pi})^{N/2} \Pi_n \frac{1}{\sigma_n} \exp(-\sum_n \frac{(x_n - \mu)^2}{2\sigma_n^2})$$

To find the maximum likelihood,

$$L \equiv \log P(\{x_n\}|\sigma_n, \mu) = -\sum_n \log\ \sigma_n - \sum_n \frac{(x_n - \mu)^2}{2\sigma_n^2} = 0$$

$$\frac{\partial L}{\partial x_n} = -\sum_n \left(\frac{x_n - \mu}{\sigma_n^2} \right) = -\sum_n \frac{x_n}{\sigma_n^2} + \mu \sum_n \frac{1}{\sigma_n^2}$$

$$\mu = \frac{\sum_n \frac{x_n}{\sigma_n^2}}{\frac{1}{\sigma_n^2}}$$

When $x_n = (-27.02, 3.57)$, the correspondent σ_n are so huge, then it contributes very tiny in μ. Then, the mean should be close to mean of $(8.191, 9.898, 9.603, 9.945, 10.056)$.

2.5 Exercise 24.3

The setting is same as the exercise 22.15.

Bayesian for posterior probability of σ_n is

$$P(\sigma_n|\{x_n\}, \mu) = \frac{P(\{x_n\}|\sigma_n, \mu) P(\sigma_n|\mu)}{P(\{x_n\}|\mu)}$$

Given

$$P(\{x_n\}|\sigma_n, \mu) = (\frac{1}{2\pi})^{N/2} \left(\Pi_n \frac{1}{\sigma_n} \right) \exp\left(-\sum_n \frac{(x_n - \mu)^2}{2\sigma_n^2} \right)$$

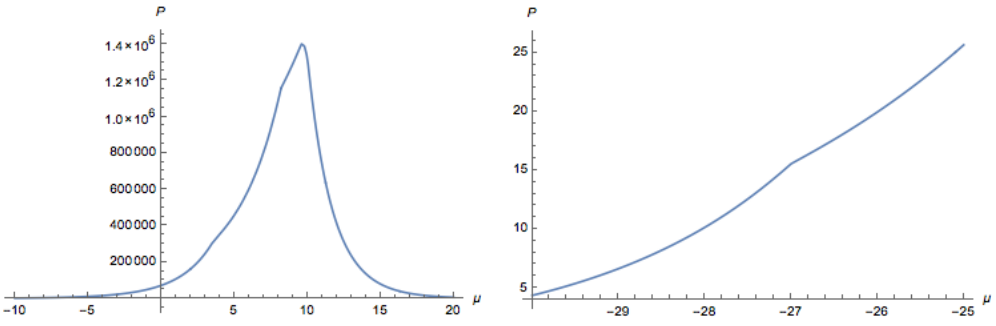

Figure 2: The distribution functions are not normalized yet. However, this plot already shows the maximum likelihood should be around $\mu = 10$. The small bumps around $x_n = (-27,\ 3.6,\ 8)$ are also seen in the graph.

and

$$P(\sigma_n|\mu) = (\frac{1}{\Gamma(c)\ s^c})^N \Pi_n\ (\sigma_n)^{c-1}\ \ \exp(\frac{\sum_n \sigma_n}{s})^1$$

where $(s,\ c) = (10, 0.1)$.

and the normalizing constant is

$$P(\{x_n\}|\mu) = \int_0^\infty \Pi_n\ d\sigma_n\ P(\{x_n\}|\sigma_n, \mu)P(\sigma_n|\mu)$$

The posterior probability of μ is

$$P(\mu|x_n) = \frac{P(\{x_n\}|\mu)P(\mu)}{P(\{x_n\})}$$

and the prior is determined by some assumption. $P(\mu) = \frac{1}{\sigma_\mu} = const.$, then the normalizing constant is

$$P(\{x_n\}) = \int_{-\infty}^\infty\ d\mu\ P(\{x_n\}|\mu)\frac{1}{\sigma_\mu}$$

3 Monte Carlo

3.1 Exercise 27.1

r is a positive integer, and posterior over lambda

$$P(\lambda|r) = \frac{P(r|\lambda)P(\lambda)}{P(r)}$$

$$P(r|\lambda) = \exp(\lambda)\frac{\lambda^r}{r!}$$

[1] not sure if it is right.

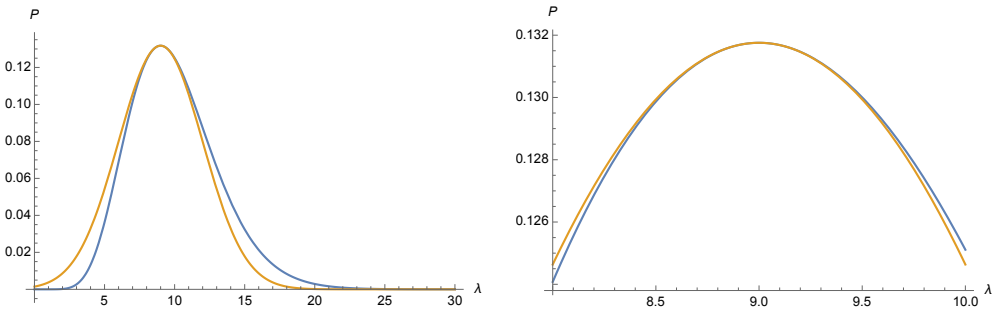

Figure 3: $r = 10$ is set. The blue plot is Poisson distribution and the orange one is the Laplace approximation.

By assumption

$$P(\lambda) = \frac{1}{\lambda}$$

Then, the normalizing constant is

$$P(r) = \int_0^\infty d\lambda \ \exp(\lambda) \frac{\lambda^r}{r!} \frac{1}{\lambda} = \frac{\Gamma(r)}{r!} = \frac{1}{r}$$

We need to find the *lambda* for maximum likelihood. By differentiate the posterior probability distribution function with respect to λ, it has a maximum at $\lambda = \lambda_0 = r - 1$, and

$$P(\lambda = \lambda_0 = r - 1|r) = \frac{(r-1)^{(r-1)} \exp(r-1)}{(r-1)!}$$

$$c = -\frac{\partial^2}{\partial \lambda^2} \log \ P(\lambda|r)|_{\lambda=\lambda_0=r-1} = \frac{1-r}{\lambda^2}|_{\lambda=\lambda_0=r-1} = \frac{1}{r-1}$$

Then, the Laplace approximation is

$$G(\lambda|r) = P(\lambda = \lambda_0 = r - 1|r) \exp\left(-\frac{c}{2}(\lambda - \lambda_0)^2\right)$$

3.2 Exercise 27.3

Bayesian for posterior of ω_0, ω_1. $y(x_n)$ is the mean of t_n data points.

$$P(\{\omega_i\}|t_n, x_n, \sigma_\nu) = \frac{P(t_n|\{\omega_i\}, x_n, \sigma_\nu)P(\omega_0)P(\omega_1)}{P(t_n|x_n\sigma_\nu)}$$ [2]

$P(t_n|\{\omega_i\}, \sigma_\nu)$, $P(\omega_0)$, $P(\omega_1)$ are all Gaussian distributed. Prior $P(\omega_0)$, $P(\omega_1)$ are assumed to be Gaussian.

[2] not sure if it is correct yet. I am still confused at Bayesian theory.

$$P(t_n|\sigma_\nu) = \int_{-\infty}^{\infty} d\omega_0 \int_{-\infty}^{\infty} d\omega_1 \; P(t_n|\{\omega_i\}, \sigma_\nu) P(\omega_0) P(\omega_1)$$

$$P(\{t_n\}|\{\omega_i\}, x_n, \sigma) = (\frac{1}{2\pi\sigma_\nu^2})^{N/2} \exp\left(- \sum_n \frac{(t_n - y(x_n))^2}{2\sigma^2} \right)$$

$$y(x_n) = \omega_0 + \omega_1 x_n$$

$$P(\omega_i) = (\frac{1}{2\pi\sigma^2})^{1/2} \exp\left(-\frac{\omega_i^2}{2\sigma_i^2} \right)$$

Because the mean of ω_i should be zero obviously, and the variances of ω_0 and ω_1, i.e. σ_1, σ_2 could be same without loss of generality. The variance of t_n would have a very broad range of values dependent on x_n. I assumed $\sigma_i^2 = \sigma/3$ at Figure 4.

To use Laplace approximation, find $\{\omega_i\}$ satisfy $\frac{\partial}{\partial \omega_i} \log P(\{\omega_i\}|t_n, x_n, \sigma)$. [3]

$$\log P(\{\omega_i\}|t_n, x_n, \sigma) = - \left(\sum_n \frac{(t_n - \omega_0 - \omega_1 x_n)^2}{2\sigma^2} \right) - \frac{\omega_0^2}{2\sigma_0^2} - \frac{\omega_1^2}{2\sigma_1^2} \qquad (3.1)$$

Some reader might notice that $-\sum_n \frac{(t_n - \omega_0 - \omega_1 x_n)^2}{2\sigma^2}$ term is minimized for the linear regression. We have two extra terms from the priors. It is *Bayesian linear regression*.

By the way, ω_0^* and ω_1^* satisfy Eq. 3.1 is quite complicated. I tried to find the simplified form, but can't. I would just keep symbolic expressions of ω_0^*, ω_1^*.

Using the Eq. (27.6) of the book,

$$\mathbf{A} = \frac{1}{\sigma^2} \begin{pmatrix} N+1 & -\sum_n x_n \\ -\sum_n x_n & (\sum_n x_n)^2 + 1 \end{pmatrix}$$

Then, the Laplace approximation is

$$G(\{\omega_i\}|t_n) = P(\{\omega_i^*\}|t_n, x_n, \sigma) exp\left(-\frac{1}{2} (\boldsymbol{\omega} - \boldsymbol{\omega}_0)^T \mathbf{A} (\boldsymbol{\omega} - \boldsymbol{\omega}_0) \right)$$

where $\boldsymbol{\omega} = (\omega_0, \omega_1)$ and $\boldsymbol{\omega}^* = (\omega_0^*, \omega_1^*)$

[3] Do you have any better suggestion?

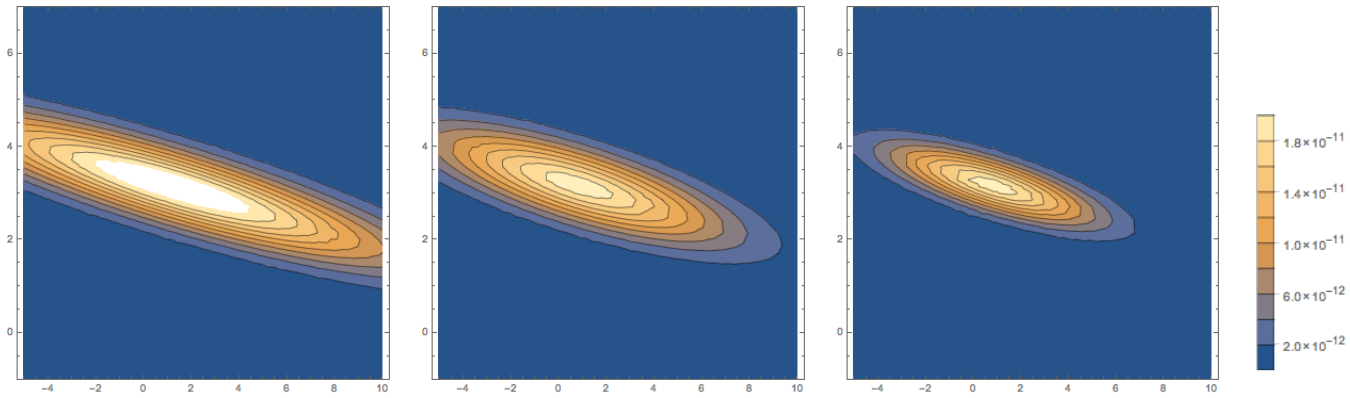

Figure 4: The likelihood and posterior probability distribution plot and its Laplace approximation. I have put 10 random data points x_n and t_n $3x_n + 2 \pm 2$. The left plot has the maximum likelihood around $(\omega_0, \omega_2) = (1.79794, 3.02217)$, and the central posterior is about $(\omega_0, \omega_2) = (0.812961, 3.16977)$. The discrepancy occurs by the assumed priors. Because the mean of ω_i should be zero obviously, and the variances of ω_0 and ω_1, i.e. σ_1, σ_2 could be same without loss of generality. The variance of t_n would have a very broad range of values dependent on x_n. I assumed $\sigma_i^2 = \sigma/3$. It is chosen from base of the mean value of x_n. The last one is obtained by Laplace approximation. I did not normalize them.