Chapter 4

Fitting probability models

This chapter concerns fitting probability models to data $\{\mathbf{x}_i\}_{i=1}^{I}$. This process is referred to as *learning* because we learn about the parameters $\boldsymbol{\theta}$ of the model.[1] It also concerns calculating the probability of a new datum $\mathbf{x}^*$ under the resulting model. This is known as evaluating the *predictive distribution*. We consider three methods: *maximum likelihood*, *maximum a posteriori*, and the *Bayesian approach*.

4.1 Maximum likelihood

As the name suggests, the maximum likelihood (ML) method finds the set of parameters $\hat{\boldsymbol{\theta}}$ under which the data $\{\mathbf{x}_i\}_{i=1}^{I}$ are most likely. To calculate the likelihood function $Pr(\mathbf{x}_i|\boldsymbol{\theta})$ at a single data point $\mathbf{x}_i$, we simply evaluate the probability density function at $\mathbf{x}_i$. Assuming each data point was drawn independently from the distribution, the likelihood function $Pr(\mathbf{x}_{1\ldots I}|\boldsymbol{\theta})$ for a set of points is the product of the individual likelihoods. Hence, the ML estimate of the parameters is

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[Pr(\mathbf{x}_{1\ldots I}|\boldsymbol{\theta})\right]\\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[\prod_{i=1}^{I}Pr(\mathbf{x}_i|\boldsymbol{\theta})\right],\end{aligned}\tag{4.1}$$

where $\operatorname{argmax}_{\boldsymbol{\theta}} f[\boldsymbol{\theta}]$ returns the value of $\boldsymbol{\theta}$ that maximizes the argument $f[\boldsymbol{\theta}]$.

To evaluate the predictive distribution for a new data point $\mathbf{x}^*$ (compute the probability that $\mathbf{x}^*$ belongs to the fitted model), we simply evaluate the probability density function $Pr(\mathbf{x}^*|\hat{\boldsymbol{\theta}})$ using the ML fitted parameters $\hat{\boldsymbol{\theta}}$.

[1] Here we adopt the notation $\boldsymbol{\theta}$ to represent a generic set of parameters when we have not specified the particular probability model.

4.2 Maximum a posteriori

In maximum a posteriori (MAP) fitting, we introduce *prior* information about the parameters $\boldsymbol{\theta}$. From previous experience we may know something about the possible parameter values. For example, in a time-sequence the values of the parameters at time t tell us a lot about the possible values at time $t+1$, and this information would be encoded in the prior distribution.

As the name suggests, maximum a posteriori estimation maximizes the posterior probability $Pr(\boldsymbol{\theta}|\mathbf{x}_{1\ldots I})$ of the parameters

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[Pr(\boldsymbol{\theta}|\mathbf{x}_{1\ldots I})\right] \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[\frac{Pr(\mathbf{x}_{1\ldots I}|\boldsymbol{\theta})Pr(\boldsymbol{\theta})}{Pr(\mathbf{x}_{1\ldots I})}\right] \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[\frac{\prod_{i=1}^{I} Pr(\mathbf{x}_i|\boldsymbol{\theta})Pr(\boldsymbol{\theta})}{Pr(\mathbf{x}_{1\ldots I})}\right], \end{aligned} \tag{4.2}$$

where we have used Bayes' rule between the first two lines and subsequently assumed independence. In fact, we can discard the denominator as it is constant with respect to the parameters and so does not affect the position of the maximum, and we get

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[\prod_{i=1}^{I} Pr(\mathbf{x}_i|\boldsymbol{\theta})Pr(\boldsymbol{\theta})\right]. \tag{4.3}$$

Comparing this to the maximum likelihood criterion (equation 4.1) we see that it is identical except for the additional prior term; maximum likelihood is a special case of maximum a posteriori where the prior is uninformative.

The predictive density (probability of a new datum $\mathbf{x}^*$ under the fitted model) is again calculated by evaluating the pdf $Pr(\mathbf{x}^*|\hat{\boldsymbol{\theta}})$ using the new parameters.

4.3 The Bayesian approach

In the Bayesian approach we stop trying to estimate single fixed values (*point estimates*) of the parameters $\boldsymbol{\theta}$ and admit what is obvious; there may be many values of the parameters that are compatible with the data. We compute a probability distribution $Pr(\boldsymbol{\theta}|\mathbf{x}_{1\ldots I})$ over the parameters $\boldsymbol{\theta}$ based on data $\{\mathbf{x}_i\}_{i=1}^{I}$ using Bayes' rule so that

$$Pr(\boldsymbol{\theta}|\mathbf{x}_{1\ldots I}) = \frac{\prod_{i=1}^{I} Pr(\mathbf{x}_i|\boldsymbol{\theta})Pr(\boldsymbol{\theta})}{Pr(\mathbf{x}_{1\ldots I})}. \tag{4.4}$$

Evaluating the predictive distribution is more difficult for the Bayesian case since we have not estimated a single model but have instead found a probability distribution over possible models. Hence, we calculate

$$Pr(\mathbf{x}^*|\mathbf{x}_{1\ldots I}) = \int Pr(\mathbf{x}^*|\boldsymbol{\theta})Pr(\boldsymbol{\theta}|\mathbf{x}_{1\ldots I})\, d\boldsymbol{\theta}, \tag{4.5}$$

which can be interpreted as follows: the term $Pr(\mathbf{x}^*|\boldsymbol{\theta})$ is the prediction for a given value of $\boldsymbol{\theta}$. So, the integral can be thought of as a weighted sum of the predictions given by different parameters $\boldsymbol{\theta}$, where the weighting is determined by the posterior probability distribution $Pr(\boldsymbol{\theta}|\mathbf{x}_{1\ldots I})$ over the parameters (representing our confidence that different parameters are correct).

The predictive density calculations for the Bayesian, MAP and ML cases can be unified if we consider the ML and MAP estimates to be special probability distributions over the parameters where all of the density is at $\hat{\boldsymbol{\theta}}$. More formally, we can consider them as *delta functions* centered at $\hat{\boldsymbol{\theta}}$. A delta function $\delta[z]$ is a function that integrates to one, and that returns zero everywhere except at $z = 0$. We can now write

$$\begin{aligned} Pr(\mathbf{x}^*|\mathbf{x}_{1\ldots I}) &= \int Pr(\mathbf{x}^*|\boldsymbol{\theta})\delta[\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}]\, d\boldsymbol{\theta} \\ &= Pr(\mathbf{x}^*|\hat{\boldsymbol{\theta}}), \end{aligned} \tag{4.6}$$

which is exactly the calculation we originally prescribed: we simply evaluate the probability of the data under the model with the estimated parameters.

4.4 Worked example 1: univariate normal

To illustrate the above ideas, we will consider fitting a univariate normal model to scalar data $\{x_i\}_{i=1}^{I}$. Recall that the univariate normal model has pdf

$$Pr(x|\mu, \sigma^2) = \text{Norm}_x[\mu, \sigma^2] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-0.5\frac{(x-\mu)^2}{\sigma^2}\right], \tag{4.7}$$

and has two parameters, the mean μ and the variance σ^2. Let us generate I independent data points $x_{1\ldots I}$ from a univariate normal with $\mu = 1$ and $\sigma^2 = 1$. Our goal is to re-estimate these parameters from the data.

4.4.1 Maximum likelihood estimation

The likelihood $Pr(x_{1\ldots I}|\mu, \sigma^2)$ of the parameters $\{\mu, \sigma^2\}$ for observed data $\{x_i\}_{i=1}^{I}$ is computed by evaluating the pdf for each data point separately and taking the product:

Algorithm 4.1

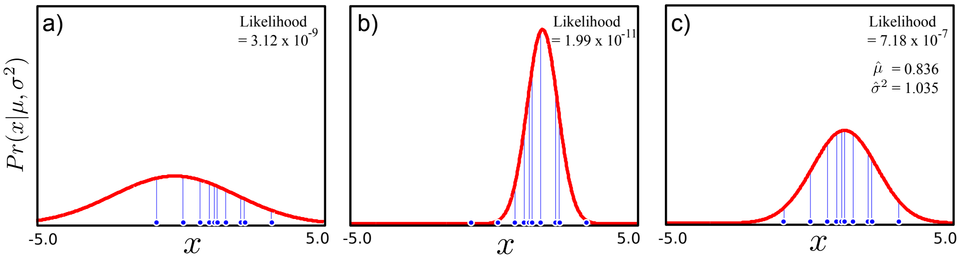

Figure 4.1 Maximum likelihood fitting. The likelihood of the parameters for a single datapoint is the height of the pdf evaluated at that point (blue vertical lines). The likelihood of a set of independently sampled data is the product of the individual likelihoods. a) The likelihood for this normal distribution is low because the large variance means the height of the pdf is low everywhere. b) The likelihood for this normal distribution is even lower as the left-most datum is very unlikely under the model. c) The maximum likelihood solution is the set of parameters for which the likelihood is maximized.

$$\begin{aligned} Pr(x_{1\ldots I}|\mu,\sigma^2) &= \prod_{i=1}^{I} Pr(x_i|\mu,\sigma^2) \\ &= \prod_{i=1}^{I} \text{Norm}_{x_i}[\mu,\sigma^2] \\ &= \frac{1}{(2\pi\sigma^2)^{I/2}} \exp\left[-0.5\sum_{i=1}^{I}\frac{(x_i-\mu)^2}{\sigma^2}\right]. \end{aligned} \tag{4.8}$$

Obviously, the likelihood for some sets of parameters $\{\mu,\sigma^2\}$ will be higher than others (figure 4.1) and it is possible to visualize this as a 2D function of the mean μ and variance σ^2 (figure 4.2). The maximum likelihood solution $\hat{\mu},\hat{\sigma}$ will occur at the peak of this surface so that

$$\hat{\mu},\hat{\sigma}^2 = \underset{\mu,\sigma^2}{\text{argmax}}\left[Pr(x_{1\ldots I}|\mu,\sigma^2)\right]. \tag{4.9}$$

In principle we can maximize this by taking the derivative of equation 4.8 with respect to μ and σ^2, equating the result to zero and solving. In practice, however, the resulting equations are messy. To simplify things, we work instead with the logarithm of this expression (the log likelihood, L). Since the logarithm is a monotonic function (figure 4.3), the position of the maximum in the transformed function remains the same. Algebraically, the logarithm turns the product of the likelihoods of the individual data points into a sum and so decouples the contribution of each. The ML parameters can now be calculated as

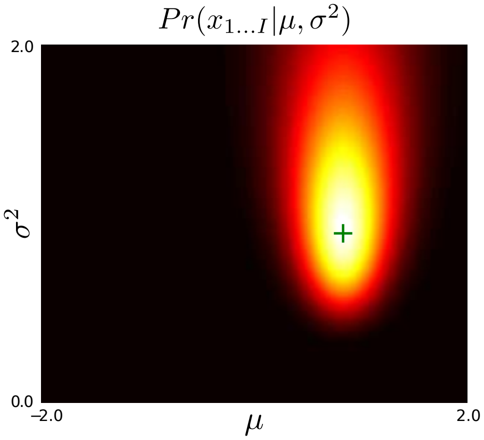

Figure 4.2 The likelihood function for a fixed set of observed data is a function of the mean μ and variance σ^2 parameters. The plot shows that there are many parameter settings which might plausibly be responsible for the ten data points from figure 4.1. A sensible choice for the "best" parameter setting is the maximum likelihood solution (green cross), which corresponds to the maximum of this function.

$$\begin{aligned}\hat{\mu},\hat{\sigma}^2 &= \underset{\mu,\sigma^2}{\operatorname{argmax}}\left[\sum_{i=1}^{I}\log\left[\text{Norm}_{x_i}[\mu,\sigma^2]\right]\right] \qquad (4.10)\\ &= \underset{\mu,\sigma^2}{\operatorname{argmax}}\left[-0.5I\log[2\pi]-0.5I\log\sigma^2-0.5\sum_{i=1}^{I}\frac{(x_i-\mu)^2}{\sigma^2}\right].\end{aligned}$$

To maximize, we differentiate this *log likelihood* $\mathbf{L}$ with respect to μ and equate the result to zero

$$\begin{aligned}\frac{\partial L}{\partial \mu} &= \sum_{i=1}^{I}\frac{(x_i-\mu)}{\sigma^2}\\ &= \frac{\sum_{i=1}^{I}x_i}{\sigma^2}-\frac{I\mu}{\sigma^2}=0 \qquad (4.11)\end{aligned}$$

and re-arranging, we see that

$$\hat{\mu}=\frac{\sum_{i=1}^{I}x_i}{I}. \qquad (4.12)$$

By a similar process, the expression for the variance can be shown to be

Problem 4.1

$$\hat{\sigma}^2=\sum_{i=1}^{I}\frac{(x_i-\hat{\mu})^2}{I}. \qquad (4.13)$$

These expressions are hardly surprising, but the same idea can be used to estimate parameters in other distributions where the results are less familiar.

Figure 4.1 shows a set of data points and three possible fits to the data. The mean of the maximum likelihood fit is the mean of the data. The ML fit is neither too narrow (giving very low probabilities to the furthest data points from the mean) nor too wide (resulting in a flat distribution and giving low probability to all points).

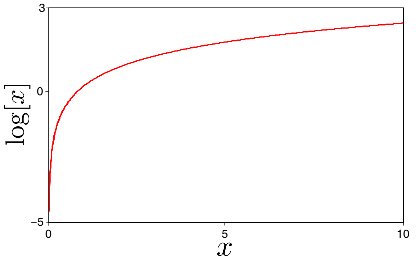

Figure 4.3 The logarithm is a monotonic transformation: if one point is higher than another then it will also be higher after transformation by the logarithmic function. It follows that if we transform the surface in figure 4.2 through the logarithmic function, the maximum will remain in the same position.

Least squares fitting

As an aside, we note that many texts discuss fitting in terms of *least squares*. Consider fitting just the mean parameter μ of the normal distribution using maximum likelihood. Manipulating the cost function so that

$$\begin{aligned}\hat{\mu} &= \underset{\mu}{\operatorname{argmax}}\left[-0.5I\log[2\pi] - 0.5I\log\sigma^2 - 0.5\sum_{i=1}^{I}\frac{(x_i-\mu)^2}{\sigma^2}\right]\\ &= \underset{\mu}{\operatorname{argmax}}\left[-\sum_{i=1}^{I}(x_i-\mu)^2\right]\\ &= \underset{\mu}{\operatorname{argmin}}\left[\sum_{i=1}^{I}(x_i-\mu)^2\right], \end{aligned} \tag{4.14}$$

leads to a formulation where we minimize the sum of squares. In other words, least squares fitting is equivalent to fitting the mean parameter of a normal distribution using the maximum likelihood method.

4.4.2 Maximum a posteriori estimation

Algorithm 4.2

Returning to the main thread, we will now demonstrate maximum a posteriori fitting of the parameters of the normal distribution. The cost function becomes

$$\begin{aligned}\hat{\mu},\hat{\sigma}^2 &= \underset{\mu,\sigma^2}{\operatorname{argmax}}\left[\prod_{i=1}^{I}Pr(x_i|\mu,\sigma^2)Pr(\mu,\sigma^2)\right]\\ &= \underset{\mu,\sigma^2}{\operatorname{argmax}}\left[\prod_{i=1}^{I}\text{Norm}_{x_i}[\mu,\sigma^2]\text{NormInvGam}_{\mu,\sigma^2}[\alpha,\beta,\gamma,\delta]\right], \end{aligned} \tag{4.15}$$

where we have chosen normal inverse gamma prior with parameters $\alpha,\beta,\gamma,\delta$ (figure 4.4) as this is conjugate to the normal distribution. The expression for the prior is

$$Pr(\mu,\sigma^2) = \frac{\sqrt{\gamma}}{\sigma\sqrt{2\pi}}\frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left[-\frac{2\beta+\gamma(\delta-\mu)^2}{2\sigma^2}\right]. \tag{4.16}$$

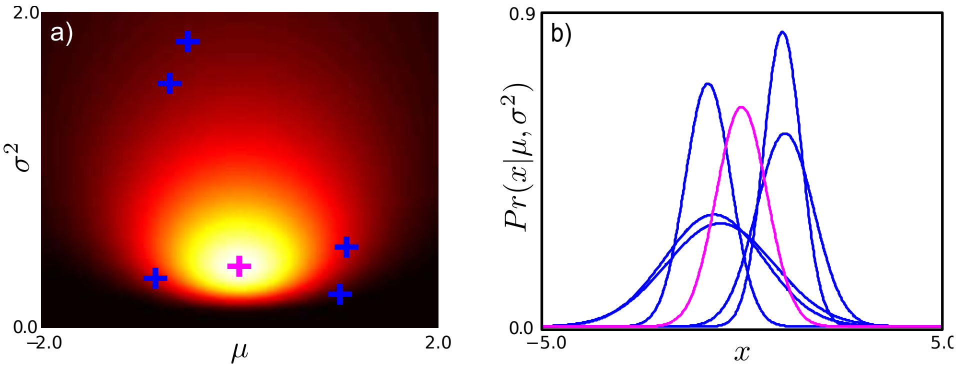

Figure 4.4 Prior over normal parameters. a) A normal inverse gamma with $\alpha, \beta, \gamma = 1$ and $\delta = 0$ gives a broad prior distribution over univariate normal parameters. The magenta cross indicates the peak of this prior distribution. The blue crosses are five samples randomly drawn from the distribution. b) The peak and the samples can be visualized by plotting the normal distributions that they represent.

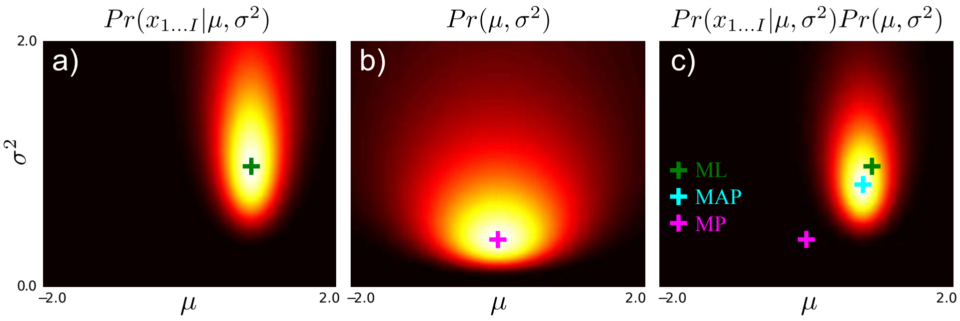

Figure 4.5 MAP inference for normal parameters. a) The likelihood function is multiplied by b) the prior probability to give a new function c) that is proportional to the posterior distribution. The maximum a posteriori (MAP) solution (cyan cross) is found at the peak of the posterior distribution. It lies between the maximum likelihood (ML) solution (green cross) and the maximum of the prior distribution (MP, magenta cross).

The posterior distribution is proportional to the product of the likelihood and the prior (figure 4.5), and has the highest density in regions that both agree with the data *and* were a priori plausible.

Like the maximum likelihood case, it is easier to maximize the logarithm of equation 4.15:

$$\hat{\mu}, \hat{\sigma}^2 = \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \log[\text{Norm}_{x_i}[\mu,\sigma^2]] + \log\left[\text{NormInvGam}_{\mu,\sigma^2}[\alpha,\beta,\gamma,\delta]\right]\right]. \tag{4.17}$$

Problem 4.2

To find the MAP parameters, we substitute in the expressions, differentiate with respect to μ and σ, equate to zero, and rearrange to give

$$\hat{\mu} = \frac{\sum_{i=1}^{I} x_i + \gamma\delta}{I+\gamma} \qquad \text{and} \qquad \hat{\sigma^2} = \frac{\sum_{i=1}^{I}(x_i-\hat{\mu})^2 + 2\beta + \gamma(\delta-\hat{\mu})^2}{I+3+2\alpha}. \tag{4.18}$$

The formula for the mean can be more easily understood if we write it as

$$\hat{\mu} = \frac{I\overline{x} + \gamma\delta}{I+\gamma}. \tag{4.19}$$

This is a weighted sum of two terms. The first term is the data mean $\overline{x}$ and is weighted by the number of training examples I. The second term is δ, the value of μ favored by the prior, and is weighted by γ.

This gives some insight into the behavior of the MAP estimate (figure 4.6). With a large amount of data, the first term dominates, and the MAP estimate $\hat{\mu}$ is very close to the data mean (and the ML estimate). With intermediate amounts of data, $\hat{\mu}$ is a weighted sum of the prediction from the data and the prediction from the prior. With no data at all, the estimate is completely governed by the prior. The hyperparameter (parameter of the prior) γ controls the concentration of the prior with respect to μ and determines the extent of its influence. Similar conclusions can be drawn about the MAP estimate of the variance.

Where there is a single data point (figure 4.6e-f), the data tells us nothing about the variance and the maximum likelihood estimate $\hat{\sigma}^2$ is actually zero; the best fit is an infinitely thin and infinitely tall normal distribution centered on the one data point. This is unrealistic, not least because it accords the datum an infinite likelihood. However, MAP estimation is still valid as the prior ensures that sensible parameter values are chosen.

4.4.3 The Bayesian approach

Algorithm 4.3

In the Bayesian approach, we calculate a posterior distribution $Pr(\mu,\sigma^2|x_{1\ldots I})$ over possible parameter values using Bayes' rule,

$$\begin{aligned} Pr(\mu,\sigma^2|x_{1\ldots I}) &= \frac{\prod_{i=1}^{I} Pr(x_i|\mu,\sigma^2)Pr(\mu,\sigma^2)}{Pr(x_{1\ldots I})} \\ &= \frac{\prod_{i=1}^{I}\text{Norm}_{x_i}[\mu,\sigma^2]\text{NormInvGam}_{\mu,\sigma^2}[\alpha,\beta,\gamma,\delta]}{Pr(x_{1\ldots I})} \\ &= \frac{\kappa\text{NormInvGam}_{\mu,\sigma^2}[\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}]}{Pr(x_{1\ldots I})}, \end{aligned} \tag{4.20}$$

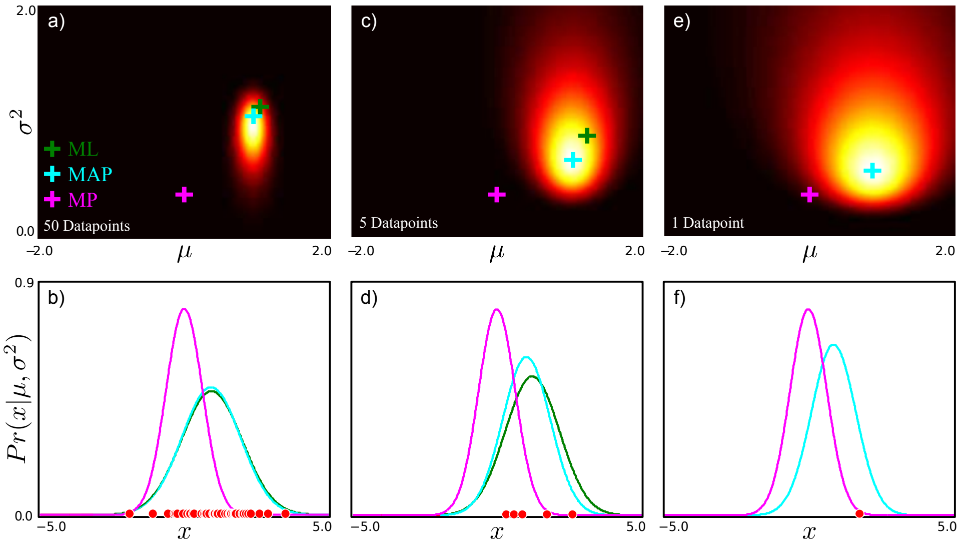

Figure 4.6 Maximum a posteriori estimation. a) MAP solution (cyan cross) lies between ML (green cross) and the peak of the prior (purple cross). b) Normal distributions corresponding to MAP solution, ML solution and peak of prior. c-d) With fewer data points, the prior has a greater effect on the final solution. e-f) With only one data point, the maximum likelihood solution cannot be computed (you cannot calculate the variance of a single point). However, the MAP solution can still be calculated.

where we have used the conjugate relationship between likelihood and prior (section 3.9) and κ is the associated constant. The product of the normal likelihood and normal inverse gamma prior creates a posterior over μ and σ^2, which is a new normal inverse gamma distribution and can be shown to have parameters

$$\tilde{\alpha} = \alpha + I/2, \qquad \tilde{\gamma} = \gamma + I \qquad \tilde{\delta} = \frac{(\gamma\delta + \sum_i x_i)}{\gamma + I}$$
$$\tilde{\beta} = \frac{\sum_i x_i^2}{2} + \beta + \frac{\gamma\delta^2}{2} - \frac{(\gamma\delta + \sum_i x_i)^2}{2(\gamma + I)}. \tag{4.21}$$

Note that the posterior (left-hand side of equation 4.20) must be a valid probability distribution and sum to one, so the constant κ from the conjugate product and the denominator from the right-hand side must exactly cancel to give

$$Pr(\mu, \sigma^2|x_{1\ldots I}) \quad = \quad \text{NormInvGam}_{\mu,\sigma^2}[\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\delta}]. \tag{4.22}$$

Now we see the major advantage of using a conjugate prior: we are guaranteed a closed form expression for the posterior distribution over the parameters.

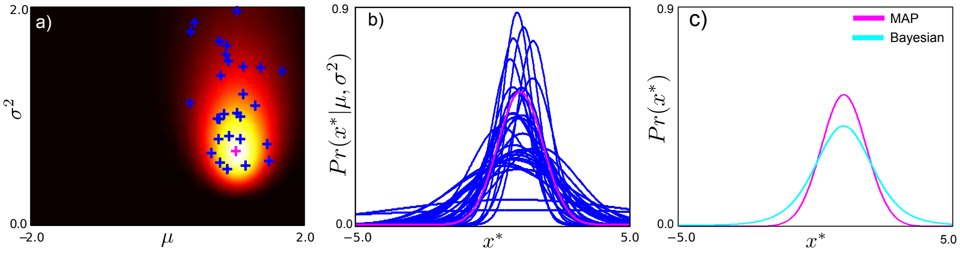

Figure 4.7 Bayesian predictions. a) Posterior probability distribution over parameters. b) Samples from posterior probability distribution correspond to normal distributions. c) The predictive distribution for the Bayesian case is the average of an infinite set of samples. Alternately, we can think of choosing the parameters from a uniform distribution and computing a weighted average where the weights correspond to the posterior distribution.

This posterior distribution represents the relative plausibility of various parameter settings μ and σ^2 having created the data. At the peak of the distribution is the MAP estimate, but there are many other plausible configurations (figure 4.6).

When data are plentiful (figure 4.6a), the parameters are well specified, and the probability distribution is concentrated. In this case, placing all of the probability mass at the MAP estimate is a good approximation to the posterior. However, when data are scarce (figure 4.6c), many possible parameters might have explained the data and the posterior is broad. In this case approximation with a point mass is inadequate.

Predictive density

For the maximum likelihood and MAP estimates, we evaluate the predictive density (probability that a new data point x^* belongs to the same model) by simply evaluating the normal pdf with the estimated parameters. For the Bayesian case, we compute a weighted average of the predictions for each possible parameter set, where the weighting is given by the posterior distribution over parameters (figures 4.6a-c and 4.7),

$$\begin{aligned} Pr(x^*|x_{1\ldots I}) &= \iint Pr(x^*|\mu,\sigma^2)Pr(\mu,\sigma^2|x_{1\ldots I})\, d\mu d\sigma \\ &= \iint \text{Norm}_{x^*}[\mu,\sigma^2]\text{NormInvGam}_{\mu,\sigma^2}[\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}]\, d\mu d\sigma \\ &= \iint \kappa(x^*,\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta})\text{NormInvGam}_{\mu,\sigma^2}[\breve{\alpha},\breve{\beta},\breve{\gamma},\breve{\delta}]\, d\mu d\sigma. \end{aligned} \tag{4.23}$$

Here we have used the conjugate relation for a second time. The integral contains a constant with respect to μ and σ^2 multiplied by a probability distribution. Taking the constant outside the integral, we get

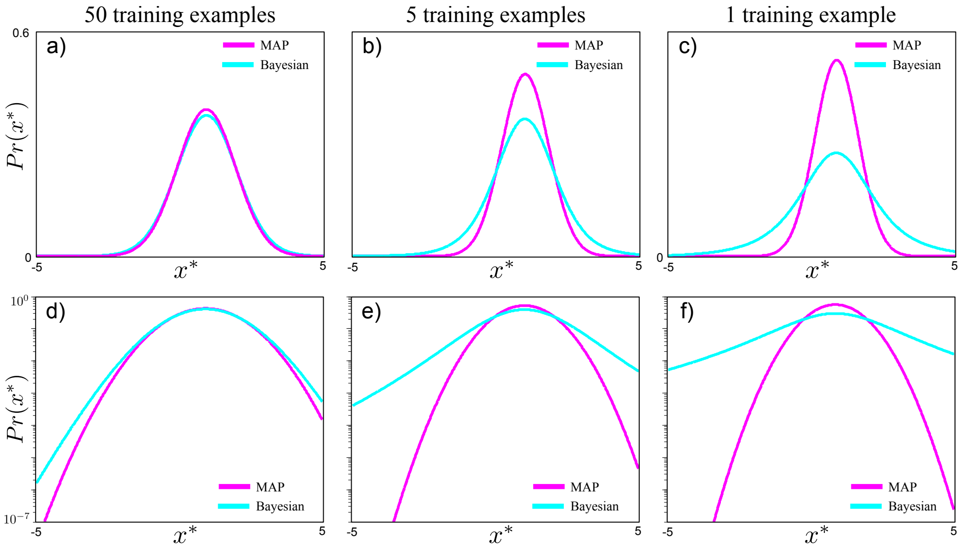

Figure 4.8 a-c) Predictive densities for MAP and Bayesian approaches with 50, 5, and 1 training examples, respectively. As the training data decreases, the Bayesian prediction becomes less certain but the MAP prediction is erroneously overconfident. d-f) This effect is even more clear on a log scale.

$$\begin{aligned} Pr(x^*|x_{1\ldots I}) &= \kappa(x^*,\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}) \iint \text{NormInvGam}_{\mu,\sigma^2}[\breve{\alpha},\breve{\beta},\breve{\gamma},\breve{\delta}]\, d\mu d\sigma \\ &= \kappa(x^*,\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}), \end{aligned} \tag{4.24}$$

which follows because the integral of a pdf is one. It can be shown that the constant is given by

$$\kappa(x^*,\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}) = \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\tilde{\gamma}}\tilde{\beta}^{\tilde{\alpha}}}{\sqrt{\breve{\gamma}}\breve{\beta}^{\breve{\alpha}}} \frac{\Gamma[\breve{\alpha}]}{\Gamma[\tilde{\alpha}]}, \tag{4.25}$$

where

$$\begin{aligned} \breve{\alpha} = \tilde{\alpha} + 1/2, \qquad\qquad & \breve{\gamma} = \tilde{\gamma} + 1 \\ \breve{\beta} = \frac{x^{*2}}{2} + \tilde{\beta} + \frac{\tilde{\gamma}\tilde{\delta}^2}{2} - & \frac{(\tilde{\gamma}\tilde{\delta} + x^*)^2}{2(\tilde{\gamma}+1)}. \end{aligned} \tag{4.26}$$

Here, we see the second advantage of using the conjugate prior; it means that the integral can be computed and so we get a nice closed form expression for the predictive density.

Figure 4.8 shows the predictive distribution for the Bayesian and MAP cases, for varying amounts of training data. With plenty of training data, they are quite

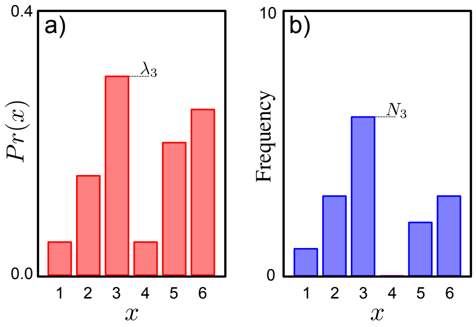

Figure 4.9 a) Categorical probability distribution over six discrete values with parameters $\{\lambda_k\}_{k=1}^6$ where $\sum_{k=1}^6 \lambda_k = 1$. This could be the relative probability of a biased die landing on its six sides. b) Fifteen observations $\{x_i\}_{i=1}^I$ randomly sampled from this distribution. We denote the number of times category k was observed by N_k so that here the total observations $\sum_{k=1}^6 N_k = 15$.

similar but as the amount of data decreases, the Bayesian predictive distribution has a significantly longer tail. This is typical of Bayesian solutions: they are more moderate (less certain) in their predictions. In the MAP case, erroneously committing to a single estimate of μ and σ^2 causes overconfidence in our future predictions.

4.5 Worked example 2: categorical distribution

As a second example, we consider discrete data $\{x_i\}_{i=1}^I$ where $x_i \in \{1, 2, \ldots, 6\}$ (figure 4.9). This could represent observed rolls of a die with unknown bias. We will describe the data using a categorical distribution (normalized histogram) where

$$Pr(x=k|\lambda_{1\ldots K}) = \lambda_k. \tag{4.27}$$

For the ML and MAP techniques, we estimate the six parameters $\{\lambda_k\}_{k=1}^6$. For the Bayesian approach, we compute a probability distribution over the parameters.

4.5.1 Maximum Likelihood

Algorithm 4.4

To find the maximum likelihood solution, we maximize the product of the likelihoods for each individual data point with respect to the parameters $\lambda_{1\ldots6}$.

$$\begin{aligned}
\hat{\lambda}_{1\ldots6} &= \underset{\lambda_{1\ldots6}}{\operatorname{argmax}} \left[\prod_{i=1}^I Pr(x_i|\lambda_{1\ldots6})\right] && \text{s.t.} \sum_k \lambda_k = 1\\
&= \underset{\lambda_{1\ldots6}}{\operatorname{argmax}} \left[\prod_{i=1}^I \text{Cat}_{x_i}[\lambda_{1\ldots6}]\right] && \text{s.t.} \sum_k \lambda_k = 1\\
&= \underset{\lambda_{1\ldots6}}{\operatorname{argmax}} \left[\prod_{k=1}^6 \lambda_k^{N_k}\right] && \text{s.t.} \sum_k \lambda_k = 1, \qquad (4.28)
\end{aligned}$$

where N_k is the total number of times we observed bin k in the training data. As before, it is easier to maximize the log probability, and we use the criterion

$$L = \sum_{k=1}^{6} N_k \log[\lambda_k] + \nu \left(\sum_{k=1}^{6} \lambda_k - 1 \right), \tag{4.29}$$

where the second term uses the Lagrange multiplier ν to enforce the constraint on the parameters $\sum_{k=1}^{6} \lambda_k = 1$. We differentiate L with respect to λ_k and ν, set the derivatives equal to zero, and solve for λ_k to obtain

$$\hat{\lambda}_k = \frac{N_k}{\sum_{m=1}^{6} N_m}. \tag{4.30}$$

In other words, λ_k is the proportion of times that we observed bin k.

4.5.2 Maximum a posteriori

To find the maximum a posteriori solution we need to define a prior. We choose the Dirichlet distribution as it is conjugate to the categorical likelihood. This prior over the six categorical parameters is hard to visualize but samples can be drawn and examined (figure 4.10a-e). The MAP solution is given by

Algorithm 4.5

$$\begin{aligned}
\hat{\lambda}_{1\ldots 6} &= \underset{\lambda_{1\ldots 6}}{\operatorname{argmax}} \left[\prod_{i=1}^{I} Pr(x_i|\lambda_{1\ldots 6}) Pr(\lambda_{1\ldots 6}) \right] \\
&= \underset{\lambda_{1\ldots 6}}{\operatorname{argmax}} \left[\prod_{i=1}^{I} \text{Cat}_{x_i}[\lambda_{1\ldots 6}] \text{Dir}_{\lambda_{1\ldots 6}}[\alpha_{1\ldots 6}] \right] \\
&= \underset{\lambda_{1\ldots 6}}{\operatorname{argmax}} \left[\prod_{k=1}^{6} \lambda_k^{N_k} \prod_{k=1}^{6} \lambda_k^{\alpha_k - 1} \right] \\
&= \underset{\lambda_{1\ldots 6}}{\operatorname{argmax}} \left[\prod_{k=1}^{6} \lambda_k^{N_k + \alpha_k - 1} \right].
\end{aligned} \tag{4.31}$$

which is again subject to the constraint that $\sum_{k=1}^{6} \lambda_k = 1$. As in the maximum likelihood case, this constraint is enforced using a Lagrange multiplier. The MAP estimate of the parameters can be shown to be

Problem 4.4

$$\hat{\lambda}_k = \frac{N_k + \alpha_k - 1}{\sum_{m=1}^{6} (N_m + \alpha_m - 1)}, \tag{4.32}$$

where N_k is the number of times that observation k occurred in the training data. Note that if all the values α_k are set to one, the prior is uniform and this expression reverts to the maximum likelihood solution (equation 4.30).

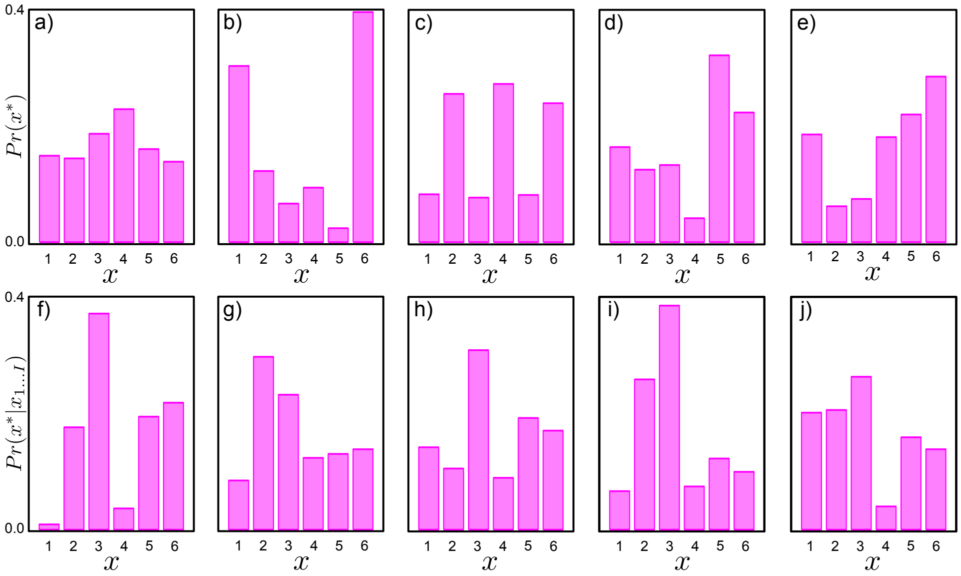

Figure 4.10 a-e) Five samples drawn from Dirichlet prior with hyperparameters $\alpha_{1\ldots 6} = 1$. This defines a uniform prior, so each sample looks like a random unstructured probability distribution. f-j) Five samples from Dirichlet posterior. The distribution favors histograms where bin three is larger and bin four is small as suggested by the data.

4.5.3 Bayesian Approach

Algorithm 4.6

In the Bayesian approach we calculate a posterior over the parameters

$$\begin{aligned}
Pr(\lambda_1 \ldots \lambda_6 | x_{1\ldots I}) &= \frac{\prod_{i=1}^{I} Pr(x_i|\lambda_{1\ldots 6})Pr(\lambda_{1\ldots 6})}{Pr(x_{1\ldots I})} \\
&= \frac{\prod_{i=1}^{I} \text{Cat}_{x_i}[\lambda_{1\ldots 6}]\text{Dir}_{\lambda_{1\ldots 6}}[\alpha_{1\ldots 6}]}{Pr(x_{1\ldots I})} \\
&= \frac{\kappa(\alpha_{1\ldots 6}, x_{1\ldots I})\text{Dir}_{\lambda_{1\ldots 6}}[\tilde{\alpha}_{1\ldots 6}]}{Pr(x_{1\ldots I})} \\
&= \text{Dir}_{\lambda_{1\ldots 6}}[\tilde{\alpha}_{1\ldots 6}], \qquad (4.33)
\end{aligned}$$

where $\tilde{\alpha}_k = N_k + \alpha_k$. We have again exploited the conjugate relationship to yield a posterior distribution with the same form as the prior. The constant κ must again cancel with the denominator to ensure a valid probability distribution on the left hand side. Samples from this distribution are shown in figure 4.10f-j.

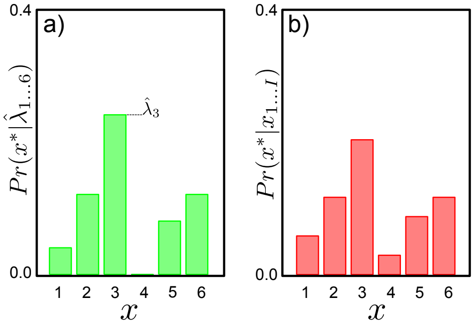

Figure 4.11 Predictive distributions with $\alpha_{1...6} = 1$ for a) maximum likelihood / maximum a posteriori approaches and b) Bayesian approach. The ML/MAP approaches predict the same distribution that exactly follows the data frequencies. The Bayesian approach predicts a more moderate distribution and allots some probability to the case $x = 4$ despite having seen no training examples in this category.

Predictive Density

For the ML and MAP estimates we evaluate the predictive density (probability that a new data point x^* belongs to the same model) by simply evaluating the categorical pdf with the estimated parameters. With the uniform prior ($\alpha_{1...6} = 1$) the MAP and ML predictions are identical (figure 4.11a) and both are exactly proportional to the frequencies of the observed data.

For the Bayesian case, we compute a weighted average of the predictions for each possible parameter set, where the weighting is given by the posterior distribution over parameters so that

$$\begin{aligned} Pr(x^*|x_{1...I}) &= \int Pr(x^*|\lambda_{1...6})Pr(\lambda_{1...6}|x_{1...I})\, d\lambda_{1...6} \\ &= \int \text{Cat}_{x^*}[\lambda_{1...6}]\text{Dir}_{\lambda_{1...6}}[\tilde{\alpha}_{1...6}]\, d\lambda_{1...6} \\ &= \int \kappa(x^*, \tilde{\alpha}_{1...6})\text{Dir}_{\lambda_{1...6}}[\breve{\alpha}_{1...6}]\, d\lambda_{1...6} \\ &= \kappa(x^*, \tilde{\alpha}_{1...6}). \end{aligned} \tag{4.34}$$

Here, we have again exploited the conjugate relationship to yield a constant multiplied by a probability distribution and the integral is simply the constant as the integral of the pdf is one. For this case, it can be shown that

$$Pr(x^* = k|x_{1...I}) = \kappa(x^*, \tilde{\alpha}_{1...6}) = \frac{N_k + \alpha_k}{\sum_{j=1}^{6}(N_j + \alpha_j)}. \tag{4.35}$$

This is illustrated in figure 4.11b. It is notable that once more the Bayesian predictive density is less confident than the ML/MAP solutions. In particular, it does not allot zero probability to observing $x^* = 4$ despite the fact that this value was never observed in the training data. This is sensible; just because we have not drawn a 4 in 15 observations does not imply that it is inconceivable that we will ever see one. We may have just been unlucky. The Bayesian approach takes this into account and allots this category a small amount of probability.

Summary

We presented three ways to fit a probability distribution to data and to predict the probability of new points. Of the three methods discussed, the Bayesian approach is the most desirable. Here it is not necessary to find a point estimate of the (uncertain) parameters, and so errors are not introduced because this point estimate is inaccurate.

However, the Bayesian approach is only tractable when we have a conjugate prior, which makes it easy to calculate the posterior distribution over the parameters $Pr(\boldsymbol{\theta}|\mathbf{x}_{1\ldots I})$ and also to evaluate the integral in the predictive density. When this is not the case, we will usually have to rely on maximum a posteriori estimates. Maximum likelihood estimates can be thought of as a special case of maximum a posteriori estimates in which the prior is uninformative.

Notes

For more information about the Bayesian approach to fitting distributions consult chapter 3 of Gelman *et al.* (2004). More information about Bayesian model selection (problem 4.6), including an impassioned argument for its superiority as a method of hypothesis testing can be found in Mackay (2003).

Problems

Problem 4.1 Show that the maximum likelihood solution for the variance σ^2 of the normal distribution is given by

$$\sigma^2 = \sum_{i=1}^{I} \frac{(x_i - \hat{\mu})^2}{I}.$$

Problem 4.2 Show that the MAP solution for the mean μ and variance σ^2 of the normal distribution are given by

$$\hat{\mu} = \frac{\sum_{i=1}^{I} x_i + \gamma\delta}{I + \gamma} \qquad \text{and} \qquad \hat{\sigma^2} = \frac{\sum_{i=1}^{I}(x_i - \hat{\mu})^2 + 2\beta + \gamma(\delta - \hat{\mu})^2}{I + 3 + 2\alpha},$$

when we use the conjugate normal-scaled inverse gamma prior

$$Pr(\mu, \sigma^2) = \frac{\sqrt{\gamma}}{\sigma\sqrt{2\pi}} \frac{\beta^\alpha}{\Gamma[\alpha]} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left[-\frac{2\beta + \gamma(\delta - \mu)^2}{2\sigma^2}\right].$$

Problem 4.3 Taking equation 4.29 as a starting point, show that the maximum likelihood parameters for the categorical distribution are given by

$$\hat{\lambda}_k = \frac{N_k}{\sum_{m=1}^{6} N_m},$$

where N_k is the number of times that category K was observed in the training data.

Problem 4.4 Show that the MAP estimate for the parameters $\{\lambda\}_{k=1}^{K}$ of the categorical distribution is given by

$$\hat{\lambda}_k = \frac{N_k + \alpha_k - 1}{\sum_{m=1}^{6}(N_m + \alpha_m - 1)},$$

under the assumption of a Dirichlet prior with hyperparameters $\{\alpha_k\}_{k=1}^{K}$. The terms N_k again indicate the number of times that category k was observed in the training data.

Problem 4.5 The denominator of Bayes' rule

$$Pr(x_{1\ldots I}) = \int \prod_{i=1}^{I} Pr(x_i|\theta) Pr(\theta)\, d\theta$$

is known as the *evidence*. It is a measure of how well the distribution fits *regardless* of the particular values of the parameters. Find an expression for the evidence term for (i) the normal distribution and (ii) the categorical distribution assuming conjugate priors in each case.

Problem 4.6 The evidence term can be used to compare models. Consider two sets of data $\mathcal{S}_1 = \{0.1, -0.5, 0.2, 0.7\}$ and $\mathcal{S}_2 = \{1.1, 2.0, 1.4, 2.3\}$. Let us pose the question of whether these two data sets came from the same normal distribution or from two different normal distributions.

Let model M_1 denote the case where all of the data comes from the one normal distribution. The evidence for this model is

$$Pr(\mathcal{S}_1 \cup \mathcal{S}_2|M_1) = \int \prod_{i \in \mathcal{S}_1 \cup \mathcal{S}_2} Pr(x_i|\boldsymbol{\theta})Pr(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

where $\boldsymbol{\theta} = \{\mu, \sigma^2\}$ contains the parameters of this normal distribution. Similarly, we will let M_2 denote the case where the two sets of data belong to different normal distributions

$$Pr(\mathcal{S}_1 \cup \mathcal{S}_2|M_2) = \int \prod_{i \in \mathcal{S}_1} Pr(x_i|\boldsymbol{\theta}_1)Pr(\boldsymbol{\theta}_1)\, d\boldsymbol{\theta}_1 \int \prod_{i \in \mathcal{S}_2} Pr(x_i|\boldsymbol{\theta}_2)Pr(\boldsymbol{\theta}_2)\, d\boldsymbol{\theta}_2,$$

where $\boldsymbol{\theta}_1 = \{\mu_1, \sigma_1^2\}$ and $\boldsymbol{\theta}_2 = \{\mu_2, \sigma_2^2\}$.

Now it is possible to compare the probability of the data under each of these two models using Bayes' rule

$$Pr(M_1|\mathcal{S}_1 \cup \mathcal{S}_2) = \frac{Pr(\mathcal{S}_1 \cup \mathcal{S}_2|M_1)Pr(M_1)}{\sum_{n=1}^{2} Pr(\mathcal{S}_1 \cup \mathcal{S}_2|M_n)Pr(M_n)}$$

Use this expression to compute the posterior probability that the two datasets came from the same underlying normal distribution. You may assume normal-scaled inverse gamma priors over $\boldsymbol{\theta}$, $\boldsymbol{\theta}_1$, and $\boldsymbol{\theta}_2$ with parameters $\alpha = 1, \beta = 1, \gamma = 1, \delta = 0$.

Note that this is (roughly) a Bayesian version of the two-sample t-test, but it is much neater - we get a posterior probability distribution over the two hypotheses rather than the potentially misleading p value of the t-test. The process of comparing evidence terms in this way is known as *Bayesian model selection* or *the evidence framework*. It is rather clever in that two normal distributions fitted with maximum likelihood will *always* explain the data better than one; the additional parameters simply make the model more flexible. However because we have marginalized these parameters away here, it is valid to compare these models in the Bayesian case.

Problem 4.7 In the Bernoulli distribution, the likelihood $Pr(x_{1\ldots I}|\lambda)$ of the data $\{x_i\}_{i=1}^{I}$ where $x_i \in \{0, 1\}$ given the parameter λ is

$$Pr(x_{1\ldots I}|\lambda) = \prod_{i=1}^{I} \lambda^{x_i}(1-\lambda)^{1-x_i}.$$

Find an expression for the maximum likelihood estimate of the parameter λ.

Problem 4.8 Find an expression for the MAP estimate of the Bernoulli parameter λ (see problem 4.7) assuming a beta distributed prior

$$Pr(\lambda) = \text{Beta}_{\lambda}[\alpha, \beta].$$

Problem 4.9 Now consider the Bayesian approach to fitting Bernoulli data, using a beta distributed prior. Find expressions for (i) the posterior probability distribution over the Bernoulli parameters given observed data $\{x_i\}_{i=1}^{I}$ and (ii) the predictive distribution for new data $\mathbf{x}^*$.

Problem 4.10 Staying with the Bernoulli distribution, consider observing data $0, 0, 0, 0$ from four trials. Assuming a uniform beta prior ($\alpha = 1, \beta = 1$), compute the predictive distribution using the (i) maximum likelihood, (ii) maximum a posteriori and (iii) Bayesian approaches. Comment on the results.