Chapter 5
Bayesian statistics

5.1 Introduction

Using the posterior distribution to summarize everything we know about a set of unknown variables is at the core of Bayesian statistics. In this chapter, we discuss this approach to statistics in more detail.

5.2 Summarizing posterior distributions

The posterior $p(\boldsymbol{\theta}|\mathcal{D})$ summarizes everything we know about the unknown quantities $\boldsymbol{\theta}$. In this section, we discuss some simple quantities that can be derived from a probability distribution, such as a posterior. These summary statistics are often easier to understand and visualize than the full joint.

5.2.1 MAP estimation

We can easily compute a **point estimate** of an unknown quantity by computing the posterior mean, median or mode. In Section 5.7, we discuss how to use decision theory to choose between these methods. Typically the posterior mean or median is the most appropriate choice for a realvalued quantity, and the vector of posterior marginals is the best choice for a discrete quantity. However, the posterior mode, aka the MAP estimate, is the most popular choice because it reduces to an optimization problem, for which efficient algorithms often exist. Futhermore, MAP estimation can be interpreted in non-Bayesian terms, by thinking of the log prior as a regularizer (see Section TODO for more details).

Although this approach is computationally appealing, it is important to point out that there are various drawbacks to MAP estimation, which we briefly discuss below. This will provide motivation for the more thoroughly Bayesian approach which we will study later in this chapter(and elsewhere in this book).

5.2.1.1 No measure of uncertainty

The most obvious drawback of MAP estimation, and indeed of any other *point estimate* such as the posterior mean or median, is that it does not provide any measure of uncertainty. In many applications, it is important to know how much one can trust a given estimate. We can derive such confidence measures from the posterior, as we discuss in Section 5.2.2.

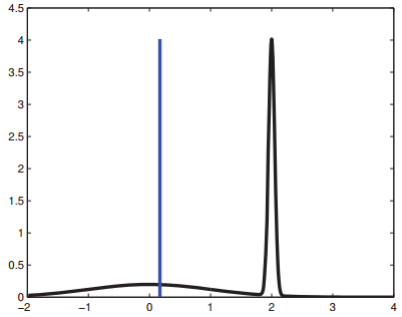
5.2.1.2 Plugging in the MAP estimate can result in overfitting

If we dont model the uncertainty in our parameters, then our predictive distribution will be overconfident. Overconfidence in predictions is particularly problematic in situations where we may be risk averse; see Section 5.7 for details.

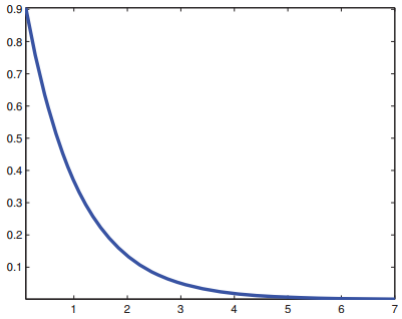
5.2.1.3 The mode is an untypical point

Choosing the mode as a summary of a posterior distribution is often a very poor choice, since the mode is usually quite untypical of the distribution, unlike the mean or median. The basic problem is that the mode is a point of measure zero, whereas the mean and median take the volume of the space into account. See Figure 5.1.

How should we summarize a posterior if the mode is not a good choice? The answer is to use decision theory, which we discuss in Section 5.7. The basic idea is to specify a loss function, where $L(\theta, \hat{\theta})$ is the loss you incur if the truth is θ and your estimate is $\hat{\theta}$. If we use 0-1 loss $L(\theta, \hat{\theta}) = \mathbb{I}(\theta \neq \hat{\theta})$(see section 1.2.2.1), then the optimal estimate is the posterior mode. 0-1 loss means you only get points if you make no errors, otherwise you get nothing: there is no partial credit under this loss function! For continuous-valued quantities, we often prefer to use squared error loss, $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$; the corresponding optimal estimator is then the posterior mean, as we show in Section 5.7. Or we can use a more robust loss function, $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$, which gives rise to the posterior median.



(a)



(b)

Fig. 5.1: (a) A bimodal distribution in which the mode is very untypical of the distribution. The thin blue vertical line is the mean, which is arguably a better summary of the distribution, since it is near the majority of the probability mass. (b) A skewed distribution in which the mode is quite different from the mean.

5.2.1.4 MAP estimation is not invariant to reparameterization *

A more subtle problem with MAP estimation is that the result we get depends on how we parameterize the probability distribution. Changing from one representation to another equivalent representation changes the result, which is not very desirable, since the units of measurement are arbitrary (e.g., when measuring distance, we can use centimetres or inches).

To understand the problem, suppose we compute the posterior forx. If we define y=f(x), the distribution for yis given by Equation 2.44. The $\frac{dx}{dy}$ term is called the Jacobian, and it measures the change in size of a unit volume passed through f. Let $\hat{x} = \arg\max_x p_x(x)$ be the MAP estimate for x. In general it is not the case that $\hat{x} = \arg\max_x p_x(x)$ is given by $f(\hat{x})$. For example, let $X \sim \mathcal{N}(6,1)$ and $y = f(x)$, where $f(x) = 1/(1+\exp(-x+5))$.

We can derive the distribution of y using Monte Carlo simulation (see Section 2.7). The result is shown in Figure **??**. We see that the original Gaussian has become squashed by the sigmoid nonlinearity. In particular, we see that the mode of the transformed distribution is not equal to the transform of the original mode.

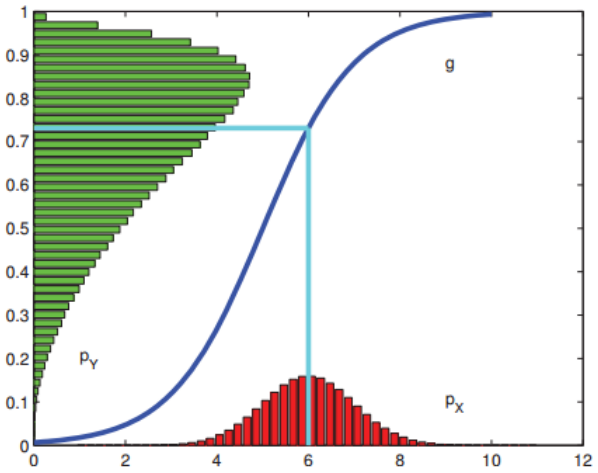


Fig. 5.2: Example of the transformation of a density under a nonlinear transform. Note how the mode of the transformed distribution is not the transform of the original mode. Based on Exercise 1.4 of (Bishop 2006b).

The MLE does not suffer from this since the likelihood is a function, not a probability density. Bayesian inference does not suffer from this problem either, since the change of measure is taken into account when integrating over the parameter space.

5.2.2 Credible intervals

In addition to point estimates, we often want a measure of confidence. A standard measure of confidence in some (scalar) quantity θ is the width of its posterior distribution. This can be measured using a $100(1\alpha)\%$ credible interval, which is a (contiguous) region $C = (\ell, u)$(standing for lower and upper) which contains 1α of the posterior probability mass, i.e.,

$$C_\alpha(\mathcal{D}) \quad \text{where } P(\ell \leq \theta \leq u) = 1 - \alpha \tag{5.1}$$

There may be many such intervals, so we choose one such that there is $(1\alpha)/2$ mass in each tail; this is called a **central interval**.

If the posterior has a known functional form, we can compute the posterior central interval using $\ell = F^{-1}(\alpha/2)$ and $u = F^{-1}(1-\alpha/2)$, where F is the cdf of the posterior.

If we dont know the functional form, but we can draw samples from the posterior, then we can use a Monte Carlo approximation to the posterior quantiles: we simply sort the $\mathcal{S}$ samples, and find the one that occurs at location

$\alpha/\mathcal{S}$ along the sorted list. As $\mathcal{S} \to \infty$, this converges to the true quantile.

People often confuse Bayesian credible intervals with frequentist confidence intervals. However, they are not the same thing, as we discuss in Section TODO. In general, credible intervals are usually what people want to compute, but confidence intervals are usually what they actually compute, because most people are taught frequentist statistics but not Bayesian statistics. Fortunately, the mechanics of computing a credible interval is just as easy as computing a confidence interval.

5.2.3 Inference for a difference in proportions

Sometimes we have multiple parameters, and we are interested in computing the posterior distribution of some function of these parameters. For example, suppose you are about to buy something from Amazon.com, and there are two sellers offering it for the same price. Seller 1 has 90 positive reviews and 10 negative reviews. Seller 2 has 2 positive reviews and 0 negative reviews. Who should you buy from?[18].

On the face of it, you should pick seller 2, but we cannot be very confident that seller 2 is better since it has had so few reviews. In this section, we sketch a Bayesian analysis of this problem. Similar methodology can be used to compare rates or proportions across groups for a variety of other settings.

Let θ_1 and θ_2 be the unknown reliabilities of the two sellers. Since we dont know much about them, well endow them both with uniform priors, $\theta_i \sim \text{Beta}(1,1)$. The posteriors are $p(\theta_1|\mathcal{D}_1) = \text{Beta}(91,11)$ and $p(\theta_2|\mathcal{D}_2) = \text{Beta}(3,1)$.

We want to compute $p(\theta_1 > \theta_2|\mathcal{D})$. For convenience, let us define $\delta = \theta_1 - \theta_2$ as the difference in the rates. (Alternatively we might want to work in terms of the log-odds ratio.) We can compute the desired quantity using numerical integration

$$p(\delta > 0|\mathcal{D}) = \int_0^1 \int_0^1 \mathbb{I}(\theta_1 > \theta_2)\text{Beta}(\theta_1|91,11) \text{Beta}(\theta_2|3,1)\text{d}\theta_1\text{d}\theta_2 \tag{5.2}$$

We find $p(\delta > 0|\mathcal{D}) = 0.710$, which means you are better off buying from seller 1!

[18] This example is from http://www.johndcook.com/blog/2011/09/27/bayesian-amazon/

5.3 Bayesian model selection

In general, when faced with a set of models (i.e., families of parametric distributions) of different complexity, how should we choose the best one? This is called the **model selection** problem.

One approach is to use cross-validation to estimate the generalization error of all the candidate models, and then to pick the model that seems the best. However, this requires fitting each model K times, where K is the number of CV folds. A more efficient approach is to compute the posterior over models,

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{\sum_{m'} p(\mathcal{D}|m')p(m')} \tag{5.3}$$

From this, we can easily compute the MAP model, $\hat{m} = \arg\max_m p(m|\mathcal{D})$. This is called **Bayesian model selection**.

If we use a uniform prior over models, this amounts to picking the model which maximizes

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|m)\text{d}\boldsymbol{\theta} \tag{5.4}$$

This quantity is called the **marginal likelihood**, the **integrated likelihood**, or the **evidence** for model m. The details on how to perform this integral will be discussed in Section 5.3.2. But first we give an intuitive interpretation of what this quantity means.

5.3.1 Bayesian Occam's razor

One might think that using $p(\mathcal{D}|m)$ to select models would always favour the model with the most parameters. This is true if we use $p(\mathcal{D}|\hat{\boldsymbol{\theta}}_m)$ to select models, where $\hat{\boldsymbol{\theta}}_m)$ is the MLE or MAP estimate of the parameters for model m, because models with more parameters will fit the data better, and hence achieve higher likelihood. However, if we integrate out the parameters, rather than maximizing them, we are automatically protected from overfitting: models with more parameters do not necessarily have higher *marginal likelihood*. This is called the **Bayesian Occams razor** effect (MacKay 1995b; Murray and Ghahramani 2005), named after the principle known as **Occams razor**, which says one should pick the simplest model that adequately explains the data.

One way to understand the Bayesian Occams razor is to notice that the marginal likelihood can be rewritten as follows, based on the chain rule of probability (Equation 2.3):

$$p(D) = p((\boldsymbol{x}_1, y_1))p((\boldsymbol{x}_2, y_2)|(\boldsymbol{x}_1, y_1)) \\ p((\boldsymbol{x}_3, y_3)|(\boldsymbol{x}_1, y_1) : (\boldsymbol{x}_2, y_2)) \cdots \\ p((\boldsymbol{x}_N, y_N)|(\boldsymbol{x}_1, y_1) : (\boldsymbol{x}_{N-1}, y_{N-1})) \tag{5.5}$$

This is similar to a leave-one-out cross-validation estimate (Section 1.3.4) of the likelihood, since we predict each future point given all the previous ones. (Of course, the order of the data does not matter in the above expression.) If a model is too complex, it will overfit the early examples and will then predict the remaining ones poorly.

Another way to understand the Bayesian Occams razor effect is to note that probabilities must sum to one. Hence $\sum_{p(\mathcal{D}')} p(m|\mathcal{D}') = 1$, where the sum is over all possible data sets. Complex models, which can predict many things, must spread their probability mass thinly, and hence will not obtain as large a probability for any given data set as simpler models. This is sometimes called the **conservation of probability mass** principle, and is illustrated in Figure 5.3.

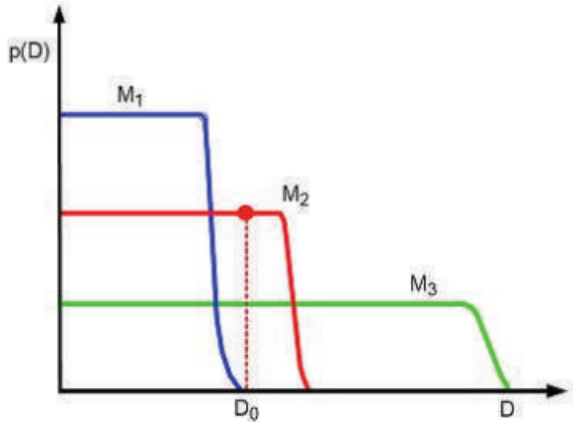


Fig. 5.3: A schematic illustration of the Bayesian Occams razor. The broad (green) curve corresponds to a complex model, the narrow (blue) curve to a simple model, and the middle (red) curve is just right. Based on Figure 3.13 of (Bishop 2006a).

When using the Bayesian approach, we are not restricted to evaluating the evidence at a finite grid of values. Instead, we can use numerical optimization to find $\lambda^* = \arg\max_\lambda p(\mathcal{D}|\lambda)$. This technique is called **empirical Bayes** or **type II maximum likelihood** (see Section 5.6 for details). An example is shown in Figure TODO(b): we see that the curve has a similar shape to the CV estimate, but it can be computed more efficiently.

5.3.2 Computing the marginal likelihood (evidence)

When discussing parameter inference for a fixed model, we often wrote

$$p(\boldsymbol{\theta}|\mathcal{D}, m) \propto p(\boldsymbol{\theta}|m)p(\mathcal{D}|\boldsymbol{\theta}, m) \tag{5.6}$$

thus ignoring the normalization constant $p(\mathcal{D}|m)$. This is valid since $p(\mathcal{D}|m)$is constant wrt $\boldsymbol{\theta}$. However, when comparing models, we need to know how to compute the marginal likelihood, $p(\mathcal{D}|m)$. In general, this can be quite hard, since we have to integrate over all possible parameter values, but when we have a conjugate prior, it is easy to compute, as we now show.

Let $p(\boldsymbol{\theta}) = q(\boldsymbol{\theta})/Z_0$ be our prior, where $q(\boldsymbol{\theta})$ is an unnormalized distribution, and Z_0 is the normalization constant of the prior. Let $p(\mathcal{D}|\boldsymbol{\theta}) = q(\mathcal{D}|\boldsymbol{\theta})/Z_\ell$ be the likelihood, where Z_ℓ contains any constant factors in the likelihood. Finally let $p(\boldsymbol{\theta}|\mathcal{D}) = q(\boldsymbol{\theta}|\mathcal{D})/Z_N$ be our posterior , where $q(\boldsymbol{\theta}|\mathcal{D}) = q(\mathcal{D}|\boldsymbol{\theta})q(\boldsymbol{\theta})$ is the unnormalized posterior, and Z_N is the normalization constant of the posterior. We have

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} \tag{5.7}$$

$$\frac{q(\boldsymbol{\theta}|\mathcal{D})}{Z_N} = \frac{q(\mathcal{D}|\boldsymbol{\theta})q(\boldsymbol{\theta})}{Z_\ell Z_0 p(\mathcal{D})} \tag{5.8}$$

$$p(\mathcal{D}) = \frac{Z_N}{Z_0 Z_\ell} \tag{5.9}$$

So assuming the relevant normalization constants are tractable, we have an easy way to compute the marginal likelihood. We give some examples below.

5.3.2.1 Beta-binomial model

Let us apply the above result to the Beta-binomial model. Since we know $p(\boldsymbol{\theta}|\mathcal{D}) = \text{Beta}(\boldsymbol{\theta}|a', b')$, where $a' = a + N_1$, $b' = b + N_0$, we know the normalization constant of the posterior is $B(a', b')$. Hence

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \tag{5.10}$$

$$= \frac{1}{p(\mathcal{D})}\left[\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\right] \left[\binom{N}{N_1}\theta^{N_1}(1-\theta)^{N_0}\right] \tag{5.11}$$

$$= \binom{N}{N_1}\frac{1}{p(\mathcal{D})}\frac{1}{B(a,b)}\left[\theta^{a+N_1-1}(1-\theta)^{b+N_0-1}\right] \tag{5.12}$$

So

$$\frac{1}{B(a+N_1,b+N_0)} = \binom{N}{N_1}\frac{1}{p(\mathcal{D})}\frac{1}{B(a,b)} \tag{5.13}$$

$$p(\mathcal{D}) = \binom{N}{N_1}\frac{B(a+N_1,b+N_0)}{B(a,b)} \tag{5.14}$$

The marginal likelihood for the Beta-Bernoulli model is the same as above, except it is missingthe $\binom{N}{N_1}$ term.

5.3.2.2 Dirichlet-multinoulli model

By the same reasoning as the Beta-Bernoulli case, one can show that the marginal likelihood for the Dirichlet-multinoulli model is given by

$$p(\mathcal{D}) = \frac{B(\boldsymbol{N}+\boldsymbol{\alpha})}{B(\boldsymbol{\alpha})} \tag{5.15}$$

$$= \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(N+\sum_k \alpha_k)}\prod_k \frac{\Gamma(N_k+\alpha_k)}{\Gamma(\alpha_k)} \tag{5.16}$$

5.3.2.3 Gaussian-Gaussian-Wishart model

Consider the case of an MVN with a conjugate NIW prior. Let Z_0 be the normalizer for the prior, Z_N be normalizer for the posterior, and let $Z_\ell(2\pi)^{ND/2} =$ be the normalizer for the likelihood. Then it is easy to see that

$$p(\mathcal{D}) = \frac{Z_N}{Z_0 Z_\ell} \tag{5.17}$$

$$= \frac{1}{(2\pi)^{ND/2}}\frac{\left(\frac{2\pi}{\kappa_N}\right)^{D/2}|\boldsymbol{S}_N|^{-\nu_N/2}2^{(\nu_0+N)D/2}\Gamma_D(\nu_N/2)}{\left(\frac{2\pi}{\kappa_0}\right)^{D/2}|\boldsymbol{S}_0|^{-\nu_0/2}2^{\nu_0 D/2}\Gamma_D(\nu_0/2)} \tag{5.18}$$

$$= \frac{1}{\pi^{ND/2}}\left(\frac{\kappa_0}{\kappa_N}\right)^{D/2}\frac{|\boldsymbol{S}_0|^{\nu_0/2}}{|\boldsymbol{S}_N|^{\nu_N/2}}\frac{\Gamma_D(\nu_N/2)}{\Gamma_D(\nu_0/2)} \tag{5.19}$$

5.3.2.4 BIC approximation to log marginal likelihood

In general, computing the integral in Equation 5.4 can be quite difficult. One simple but popular approximation is known as the **Bayesian information criterion** or **BIC**, which has the following form (Schwarz 1978):

$$\text{BIC} \triangleq \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}) - \frac{\text{dof}(\hat{\boldsymbol{\theta}})}{2}\log N \tag{5.20}$$

where $\text{dof}(\hat{\boldsymbol{\theta}})$ is the number of **degrees of freedom** in the model, and $\hat{\boldsymbol{\theta}}$ is the MLE for the model. We see that this has the form of a **penalized log likelihood**, where the penalty term depends on the models complexity. See Section TODO for the derivation of the BIC score.

As an example, consider linear regression. As we show in Section TODO, the MLE is given by $\hat{\boldsymbol{w}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ and $\sigma^2 = \frac{1}{N}\sum_{i=1}^N(y_i - \hat{\boldsymbol{w}}^T\boldsymbol{x}_i)$. The corresponding log likelihood is given by

$$\log p(\mathcal{D}|\hat{\boldsymbol{\theta}}) = -\frac{N}{2}\log(2\pi\hat{\sigma}^2) - \frac{N}{2} \tag{5.21}$$

Hence the BIC score is as follows (dropping constant terms)

$$\text{BIC} = -\frac{N}{2}\log(\hat{\sigma}^2) - \frac{D}{2}\log N \tag{5.22}$$

where D is the number of variables in the model. In the statistics literature, it is common to use an alternative definition of BIC, which we call the BIC *cost*(since we want to minimize it):

$$\text{BIC-cost} \triangleq -2\log p(\mathcal{D}|\hat{\boldsymbol{\theta}}) - \text{dof}(\hat{\boldsymbol{\theta}})\log N \approx -2\log p(\mathcal{D}) \tag{5.23}$$

In the context of linear regression, this becomes

$$\text{BIC-cost} = N\log(\hat{\sigma}^2) + D\log N \tag{5.24}$$

The BIC method is very closely related to the **minimum description length** or **MDL** principle, which characterizes the score for a model in terms of how well it fits the data, minus how complex the model is to define. See (Hansen and Yu 2001) for details.

There is a very similar expression to BIC/ MDL called the **Akaike information criterion** or **AIC**, defined as

$$\text{AIC}(m,\mathcal{D}) = \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}_{MLE}) - \text{dof}(m) \tag{5.25}$$

This is derived from a frequentist framework, and cannot be interpreted as an approximation to the marginal likelihood. Nevertheless, the form of this expression is very similar to BIC. We see that the penalty for AIC is less than for BIC. This causes AIC to pick more complex models. However, this can result in better predictive accuracy. See e.g., (Clarke et al. 2009, sec 10.2) for further discussion on such information criteria.

5.3.2.5 Effect of the prior

Sometimes it is not clear how to set the prior. When we are performing posterior inference, the details of the prior may not matter too much, since the likelihood often overwhelms the prior anyway. But when computing the marginal likelihood, the prior plays a much more important role, since we are averaging the likelihood over all possible parameter settings, as weighted by the prior.

If the prior is unknown, the correct Bayesian procedure is to put a prior on the prior. If the prior is unknown, the correct Bayesian procedure is to put a prior on the prior.

5.3.3 Bayes factors

Suppose our prior on models is uniform, $p(m) \propto 1$. Then model selection is equivalent to picking the model with the highest marginal likelihood. Now suppose we just have two models we are considering, call them the **null hypothesis**, M_0, and the **alternative hypothesis**, M_1. Define the **Bayes factor** as the ratio of marginal likelihoods:

$$\mathrm{BF}_{1,0} \triangleq \frac{p(\mathcal{D}|M_1)}{p(\mathcal{D}|M_0)} = \frac{p(M_1|\mathcal{D})}{p(M_2|\mathcal{D})} \Big/ \frac{p(M_1)}{p(M_0)} \tag{5.26}$$

5.4 Priors

The most controversial aspect of Bayesian statistics is its reliance on priors. Bayesians argue this is unavoidable, since nobody is a **tabula rasa** or **blank slate**: all inference must be done conditional on certain assumptions about the world. Nevertheless, one might be interested in minimizing the impact of ones prior assumptions. We briefly discuss some ways to do this below.

5.4.1 Uninformative priors

If we dont have strong beliefs about what θ should be, it is common to use an **uninformative** or **non-informative** prior, and to let the data speak for itself.

5.4.2 Robust priors

In many cases, we are not very confident in our prior, so we want to make sure it does not have an undue influence on the result. This can be done by using **robust priors**(Insua and Ruggeri 2000), which typically have heavy tails, which avoids forcing things to be too close to the prior mean.

5.4.3 Mixtures of conjugate priors

Robust priors are useful, but can be computationally expensive to use. Conjugate priors simplify the computation, but are often not robust, and not flexible enough to encode our prior knowledge. However, it turns out that a **mixture of conjugate priors** is also conjugate, and can approximate any kind of prior (Dallal and Hall 1983; Diaconis and Ylvisaker 1985). Thus such priors provide a good compromise between computational convenience and flexibility.

5.5 Hierarchical Bayes

A key requirement for computing the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ is the specification of a prior $p(\boldsymbol{\theta}|\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ are the hyperparameters. What if we dont know how to set $\boldsymbol{\eta}$? In some cases, we can use uninformative priors, we we discussed above. A more Bayesian approach is to put a prior on our priors! In terms of graphical models (Chapter TODO), we can represent the situation as follows:

$$\boldsymbol{\eta} \rightarrow \boldsymbol{\theta} \rightarrow \mathcal{D} \tag{5.27}$$

This is an example of a **hierarchical Bayesian model**, also called a **multi-level** model, since there are multiple levels of unknown quantities.

5.6 Empirical Bayes

Method	**Definition**
Maximum likelihood	$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\mathcal{D}\|\boldsymbol{\theta})$
MAP estimation	$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\mathcal{D}\|\boldsymbol{\theta}) p(\boldsymbol{\theta}\|\boldsymbol{\eta})$
ML-II (**Empirical Bayes**)	$\hat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} \int p(\mathcal{D}\|\boldsymbol{\theta}) p(\boldsymbol{\theta}\|\boldsymbol{\eta}) \mathrm{d}\boldsymbol{\theta}$ $= \arg\max_{\boldsymbol{\eta}} p(\mathcal{D}\|\boldsymbol{\eta})$
MAP-II	$\hat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} \int p(\mathcal{D}\|\boldsymbol{\theta}) p(\boldsymbol{\theta}\|\boldsymbol{\eta}) p(\boldsymbol{\eta}) \mathrm{d}\boldsymbol{\theta}$ $= \arg\max_{\boldsymbol{\eta}} p(\mathcal{D}\|\boldsymbol{\eta}) p(\boldsymbol{\eta})$
Full Bayes	$p(\boldsymbol{\theta}, \boldsymbol{\eta}\|\mathcal{D}) \propto p(\mathcal{D}\|\boldsymbol{\theta}) p(\boldsymbol{\theta}\|\boldsymbol{\eta})$

5.7 Bayesian decision theory

We have seen how probability theory can be used to represent and updates our beliefs about the state of the world. However, ultimately our goal is to convert our beliefs into

actions. In this section, we discuss the optimal way to do this.

Our goal is to devise a **decision procedure** or **policy**, $f(\boldsymbol{x}) : \mathcal{X} \to \mathcal{Y}$, which minimizes the **expected loss** $R_{\text{exp}}(f)$(see Equation 1.1).

In the Bayesian approach to decision theory, the optimal output, having observed $\boldsymbol{x}$, is defined as the output a that minimizes the **posterior expected loss**:

$$\rho(f) = \mathbb{E}_{p(y|\boldsymbol{x})}[L(y, f(\boldsymbol{x}))] = \begin{cases} \sum_y L[y, f(\boldsymbol{x})] p(y|\boldsymbol{x}) \\ \int_y L[y, f(\boldsymbol{x})] p(y|\boldsymbol{x}) \mathrm{d}y \end{cases} \tag{5.28}$$

Hence the **Bayes estimator**, also called the **Bayes decision rule**, is given by

$$\delta(\boldsymbol{x}) = \arg\min_{f \in \mathcal{H}} \rho(f) \tag{5.29}$$

5.7.1 Bayes estimators for common loss functions

5.7.1.1 MAP estimate minimizes 0-1 loss

When $L(y, f(x))$ is **0-1 loss**(Section 1.2.2.1), we can proof that MAP estimate minimizes 0-1 loss,

$$\begin{aligned} \arg\min_{f \in \mathcal{H}} \rho(f) &= \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{K} L[C_k, f(\boldsymbol{x})] p(C_k|\boldsymbol{x}) \\ &= \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{K} \mathbb{I}(f(\boldsymbol{x}) \neq C_k) p(C_k|\boldsymbol{x}) \\ &= \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{K} p(f(\boldsymbol{x}) \neq C_k|\boldsymbol{x}) \\ &= \arg\min_{f \in \mathcal{H}} [1 - p(f(\boldsymbol{x}) = C_k|\boldsymbol{x})] \\ &= \arg\max_{f \in \mathcal{H}} p(f(\boldsymbol{x}) = C_k|\boldsymbol{x}) \end{aligned}$$

5.7.1.2 Posterior mean minimizes ℓ_2(quadratic) loss

For continuous parameters, a more appropriate loss function is **squared error**, ℓ_2 **loss**, or **quadratic loss**, defined as $L(y, f(\boldsymbol{x})) = [y - f(\boldsymbol{x})]^2$.

The posterior expected loss is given by

$$\begin{aligned} \rho(f) &= \int_y L[y, f(\boldsymbol{x})] p(y|\boldsymbol{x}) \mathrm{d}y \\ &= \int_y [y - f(\boldsymbol{x})]^2 p(y|\boldsymbol{x}) \mathrm{d}y \\ &= \int_y \left[y^2 - 2y f(\boldsymbol{x}) + f(\boldsymbol{x})^2\right] p(y|\boldsymbol{x}) \mathrm{d}y \end{aligned} \tag{5.30}$$

Hence the optimal estimate is the posterior mean:

$$\begin{aligned} &\frac{\partial \rho}{\partial f} = \int_y [-2y + 2f(\boldsymbol{x})] p(y|\boldsymbol{x}) \mathrm{d}y = 0 \Rightarrow \\ &\int_y f(\boldsymbol{x}) p(y|\boldsymbol{x}) \mathrm{d}y = \int_y y p(y|\boldsymbol{x}) \mathrm{d}y \\ &f(\boldsymbol{x}) \int_y p(y|\boldsymbol{x}) \mathrm{d}y = \mathbb{E}_{p(y|\boldsymbol{x})}[y] \\ &f(\boldsymbol{x}) = \mathbb{E}_{p(y|\boldsymbol{x})}[y] \end{aligned} \tag{5.31}$$

This is often called the **minimum mean squared error** estimate or **MMSE** estimate.

5.7.1.3 Posterior median minimizes ℓ_1(absolute) loss

The ℓ_2 loss penalizes deviations from the truth quadratically, and thus is sensitive to outliers. A more robust alternative is the absolute or ℓ_1 loss. The optimal estimate is the posterior median, i.e., a value a such that $P(y < a|\boldsymbol{x}) = P(y \geq a|\boldsymbol{x}) = 0.5$.

Proof.

$$\begin{aligned} \rho(f) &= \int_y L[y, f(\boldsymbol{x})] p(y|\boldsymbol{x}) \mathrm{d}y = \int_y |y - f(\boldsymbol{x})| p(y|\boldsymbol{x}) \mathrm{d}y \\ &= \int_y [f(\boldsymbol{x}) - y] p(y < f(\boldsymbol{x})|\boldsymbol{x}) + \\ &\quad [y - f(\boldsymbol{x})] p(y \geq f(\boldsymbol{x})|\boldsymbol{x}) \mathrm{d}y \\ &\frac{\partial \rho}{\partial f} = \int_y [p(y < f(\boldsymbol{x})|\boldsymbol{x}) - p(y \geq f(\boldsymbol{x})|\boldsymbol{x})] \, \mathrm{d}y = 0 \Rightarrow \\ &p(y < f(\boldsymbol{x})|\boldsymbol{x}) = p(y \geq f(\boldsymbol{x})|\boldsymbol{x}) = 0.5 \\ &\therefore f(\boldsymbol{x}) = \text{median} \end{aligned}$$

5.7.1.4 Reject option

In classification problems where $p(y|\boldsymbol{x})$ is very uncertain, we may prefer to choose a reject action, in which we refuse to classify the example as any of the specified classes, and instead say dont know. Such ambiguous cases

can be handled by e.g., a human expert. This is useful in **risk averse** domains such as medicine and finance.

We can formalize the reject option as follows. Let choosing $f(\boldsymbol{x}) = c_{K+1}$ correspond to picking the reject action, and choosing $f(\boldsymbol{x}) \in \{C_1, ..., C_k\}$ correspond to picking one of the classes. Suppose we define the loss function as

$$L(f(\boldsymbol{x}), y) = \begin{cases} 0 & \text{if } f(\boldsymbol{x}) = y \text{ and } f(\boldsymbol{x}), y \in \{C_1, ..., C_k\} \\ \lambda_s & \text{if } f(\boldsymbol{x}) \neq y \text{ and } f(\boldsymbol{x}), y \in \{C_1, ..., C_k\} \\ \lambda_r & \text{if } f(\boldsymbol{x}) = C_{K+1} \end{cases} \tag{5.32}$$

where λ_s is the cost of a substitution error, and λ_r is the cost of the reject action.

5.7.1.5 Supervised learning

We can define the loss incurred by $f(\boldsymbol{x})$ (i.e., using this predictor) when the unknown state of nature is $\boldsymbol{\theta}$(the parameters of the data generating mechanism) as follows:

$$L(\boldsymbol{\theta}, f) \triangleq \mathbb{E}_{p(\boldsymbol{x}, y|\boldsymbol{\theta})}[\ell(y - f(\boldsymbol{x}))] \tag{5.33}$$

This is known as the **generalization error**. Our goal is to minimize the posterior expected loss, given by

$$\rho(f|\mathcal{D}) = \int p(\boldsymbol{\theta}|\mathcal{D}) L(\boldsymbol{\theta}, f) \mathrm{d}\boldsymbol{\theta} \tag{5.34}$$

This should be contrasted with the frequentist risk which is defined in Equation TODO.

5.7.2 The false positive vs false negative tradeoff

In this section, we focus on binary decision problems, such as hypothesis testing, two-class classification, object/event detection, etc. There are two types of error we can make: a **false positive**(aka **false alarm**), or a **false negative**(aka **missed detection**). The 0-1 loss treats these two kinds of errors equivalently. However, we can consider the following more general loss matrix:

TODO