Decision and Risk
Lecture 2: Statistical Decision Theory

Gordon J. Ross

Overview

In statistical inference, the goal is usually to estimate the unknown parameters θ of a probability distribution $p(y|\theta)$.

However in many real world situations the goal is not simply to learn a model, but to actually make a decision and take action.

Examples:

- should a doctor give medicine to a patient based on their symptoms, or declare them healthy and send them home?
- should an email classification system classify a particular email as spam, or not -spam?
- should an investor buy a company's stock based on its short term expected returns

Statistical decision theory is concerned with the problem of making decisions, in the presence of uncertainty. It translates inference, into action.

Two Types of Decision Theory

There are two main types of decision theory:

- Frequentist decision theory - usually concerned with making decisions that perform best in the ”worst possible case”. I.e. given that there are things about the world we do not know, make a decision that performs best assuming that everything we don’t know has the worst possible value for us (minimax)
- Bayesian decision theory - concerned with making decisions that perform best, based on the information we have about unknowns

The key difference is (again) that frequentists do not put probability distributions on unknown quantities and model parameters – so they will usually assume the parameter has the ”worst possible” value. This makes their decisions more robust, but arguably suboptimal in the case where there is available information.

In this course we will only cover Bayesian decision theory

Basic Elements of a Decision Problem

- θ denotes the true and unknown state of the world (e.g. a model parameter). For example, θ may represent whether a patient has a particular disease.
- $a = (a_1, a_2, \ldots, a_k)$ denotes the k actions which we can take. For example, there may be two actions - give a patient medicine if we think he is sick, or send him home if we think he is healthy
- y denotes the set of observations we have – for example, the results of medical tests carried out on the patient

We start with prior beliefs $p(\theta)$ about the state of the world. After observing the data y, we update these to give $p(\theta|y)$. Based on this, we then choose an action a_i from the set of k actions.

Loss Function

A core element of decision making is the **loss function** $L(\theta, a_i)$ which represents the loss incurred if we choose action a_i when the (usually unknown) true state of the world is θ. This 'loss' is assumed to be a positive real number – it could be either financial loss, or something more subjective (units of happiness/utility, etc).

Quick example: a doctor has to decide whether a patient is healthy ($\theta = 0$) or sick ($\theta = 1$). His possible actions are either to send the patient home empty handed (a_0), or give the patient medicine (a_1). The costs of getting the decision wrong are **not** equal. If the patient is not healthy and gets sent home without medicine, they might become extremely sick. But if the patient is healthy and mistakenly gets prescribed medicine, this is not a huge problem (in this example).

Loss Function

Suppose the first type of mistake (sending the patient home empty-handed when they are sick) is assumed to be 10 times worse than the second type (mistakenly giving drugs to a healthy patient). The losses corresponding to each action and state of world θ are then:

	$\theta = 0$	$\theta = 1$
a_0	0	10
a_1	1	0

This fully specifies the loss function $L(\theta, a)$ for all values of θ and a_i.

Basic Elements of a Decision Problem

- θ denotes the true and unknown state of the world
- $a = (a_1, a_2, \ldots, a_k)$ denotes the k actions which we can take.
- y denotes the set of observations we have
- $L(\theta, a)$ is a loss function that maps each combination of world-states θ and action a_i onto a numerical loss.

The decision maker chooses a decision a_i after observing y. He wants to choose the decision that minimises the loss $L(\theta, a)$. If he knew θ, we would simply choose the action a_i that minimised this:

$$\min_i L(\theta, a_i)$$

where θ is fixed and known.

Risk

In practice, the decision-maker does not know θ. The state of the world is uncertain. Instead, the decision maker has posterior beliefs $p(\theta|y)$ about θ after observing y.

We define the **risk** $R(a_i|y)$ of picking action a_i to be the **expected loss of** a_i **under the posterior distribution for** θ

$$R(a_i|y) = \int L(\theta, a_i)p(\theta|y)d\theta$$

In Bayesian Decision Theory, we pick the action which minimises the risk (=expected loss). An example will make this clearer.

Medical Example

A person hears on the radio that there has been an outbreak of meningitis in her city. Since she has had a headache for several days, she feels paranoid and goes to the doctor. Let θ denote the true state of nature corresponding to the person's health, which has two possible values

- $\theta = 0$ if the person is healthy (no meningitis) and their headache is not serious
- $\theta = 1$ if the person has meningitis

The doctor has two choices. If he believes the person has meningitis he will prescribe antibiotics, otherwise he will send the patient home. His possible actions are hence:

- a_0: send the patient home with no medication
- a_1: prescribe antibiotics

A Medical Example - Prior

The doctor' knows that most people who have headaches do not have meningitis. His prior is hence that the patient probably does not have meningitis. Before seeing the patient, his prior is:

- $p(\theta = 0) = 0.9$
- $p(\theta = 1) = 0.1$

If the doctor was predicting the person's health using no information other than his prior knowledge, he would send the patient home since he is 90% sure she does not have menangitis.

A Medical Example - Costs

However in this case the costs are **not** equal. If the patient really has meningitis and gets sent home without medicine, they might become extremely sick. On the other hand, if the patient does not have meningitis but gets prescribed antibiotics, this is not a huge problem.

Recall that we assumed the first type of mistake is 10 times worse than the second. The losses corresponding to each action and state of world θ are then:

	$\theta = 0$	$\theta = 1$
a_0	0	10
a_1	1	0

This specifies the loss function $L(\theta, \delta(a))$.

A Medical Example - Prior Risk

Since the doctor has no data, he computes the risk using only his prior knowledge. The prior risk associated with action a_0 (sending the patient home) is:

$$R(a_0) = p(\theta = 0) \times L(\theta = 0, a_0) + p(\theta = 1) \times L(\theta = 1, a_0) =$$
$$= 0.9 \times 0 + 0.1 \times 10 = 1$$

Similarly the prior risk associated with action a_1 (prescribing antibiotics) is

$$R(a_1) = p(\theta = 0) \times L(\theta = 0, a_1) + p(\theta = 1) \times L(\theta = 1, a_1) =$$
$$= 0.9 \times 1 + 0.1 \times 0 = 0.9$$

So the risk is minimised by prescribing antibiotics (action a_1). This is the doctor's optimal decision taking into account both his prior knowledge, and the relative costs

A Medical Example - Data

So far the doctor has not looked at any data. However in practice he may wish to perform a blood test on the patient before making a decision. Let y denote the outcome of the blood test, which has two possible values:

- $y = 0$ if the test comes back negative (no meningitis)
- $y = 1$ if the test comes back positive (meningitis)

Of course, blood tests are not always accurate. Based on previous experience, the doctor knows that the likelihood function $p(y|\theta)$ is:

$$p(y = 0|\theta = 0) = 0.8, \quad p(y = 1|\theta = 0) = 0.2 \text{ (no meningitis case)}$$

$$p(y = 0|\theta = 1) = 0.3, \quad p(y = 1|\theta = 1) = 0.7 \text{ (meningitis case)}$$

A Medical Example - Risk

We now have all the elements of a standard decision problem – the prior $p(\theta)$, the loss function $L(\theta, a)$, the data y, and the likelihood $p(y|\theta)$. Based on all this, the doctor seeks to make the decision minimising the risk.

For a given action a_i the risk is given by the expected value of the loss function under the posterior $p(\theta|y)$ which is obtained by combining the data with the prior:

$$R(a_i) = \int p(\theta|y)L(\theta, a_i)d\theta = \sum_{\theta} p(\theta|y)L(\theta, a_i)$$

This risk is computed for all actions a_i, and then the action which has the smallest risk is chosen

A Medical Example - Computing The Posterior Distributions

The first step is to compute the posterior $p(\theta|y)$ for all possible values of θ and y. This is done directly using Bayes theorem.

$$p(\theta = 0|y = 0) = \frac{p(y = 0|\theta = 0)p(\theta = 0)}{p(y = 0)} = \frac{0.8 \times 0.9}{p(y = 0)}$$

$$p(\theta = 0|y = 1) = \frac{p(y = 1|\theta = 0)p(\theta = 0)}{p(y = 1)} = \frac{0.2 \times 0.9}{p(y = 1)}$$

$$p(\theta = 1|y = 0) = \frac{p(y = 0|\theta = 1)p(\theta = 1)}{p(y = 0)} = \frac{0.3 \times 0.1}{p(y = 0)}$$

$$p(\theta = 1|y = 1) = \frac{p(y = 1|\theta = 1)p(\theta = 1)}{p(y = 1)} = \frac{0.7 \times 0.1}{p(y = 1)}$$

A Medical Example - Computing The Posterior Distributions

To commute the denominators $p(y = 0)$ and $p(y = 1)$ we use the usual theorem of total probability (see previous lecture):

$$p(y = 0) = p(y = 0|\theta = 0)p(\theta = 0) + p(y = 0|\theta = 1)p(\theta = 1) =$$
$$= 0.8 \times 0.9 + 0.3 \times 0.1 = 0.75$$

and

$$p(y = 1) = p(y = 1|\theta = 0)p(\theta = 0) + p(y = 1|\theta = 1)p(\theta = 1) =$$
$$= 0.2 \times 0.9 + 0.7 \times 0.1 = 0.25$$

A Medical Example - Computing The Posterior Distributions

Substituting these back in gives:

$$p(\theta = 0|y = 0) = \frac{p(y = 0|\theta = 0)p(\theta = 0)}{p(y = 0)} = \frac{0.8 \times 0.9}{0.75} = 0.96$$

$$p(\theta = 0|y = 1) = \frac{p(y = 1|\theta = 0)p(\theta = 0)}{p(y = 1)} = \frac{0.2 \times 0.9}{0.25} = 0.72$$

$$p(\theta = 1|y = 0) = \frac{p(y = 0|\theta = 1)p(\theta = 1)}{p(y = 0)} = \frac{0.3 \times 0.1}{0.75} = 0.04$$

$$p(\theta = 1|y = 1) = \frac{p(y = 1|\theta = 1)p(\theta = 1)}{p(y = 1)} = \frac{0.7 \times 0.1}{0.25} = 0.28$$

This completes the computation of the posterior distribution.

A Medical Example - Computing The Risk

We can now compute the risk $R(a_i|y)$ for each action a_i and value of y. First, let's do a_0

$$R(a_0|y) = \sum_{\theta} p(\theta|y)L(\theta, a_0) = p(\theta = 0|y)L(\theta = 0, a_0) + p(\theta = 1|y)L(\theta = 1, a_0) =$$

$$= p(\theta = 0|y) \times 0 + p(\theta = 1|y) \times 10 =$$

$$= 10p(\theta = 1|y)$$

So:

$$R(a_0|y = 0) = 10 \times 0.04 = 0.4$$

$$R(a_0|y = 1) = 10 \times 0.28 = 2.8$$

A Medical Example - Computing The Risk

Similarly for a_1:

$$R(a_1|y) = \sum_{\theta} p(\theta|y)L(\theta, a_1) = p(\theta = 0|y)L(\theta = 0, a_1) + p(\theta = 1|y)L(\theta = 1, a_1) =$$

$$= p(\theta = 0|y) \times 1 + p(\theta = 1|y) \times 0 =$$

$$= p(\theta = 0|y)$$

So:

$$R(a_1|y = 0) = 0.96$$

$$R(a_1|y = 1) = 0.72$$

A Medical Example - Final Decision

In summary:

- $R(a_0|y=0) = 0.4$
- $R(a_1|y=0) = 0.96$
- $R(a_0|y=1) = 2.8$
- $R(a_1|y=1) = 0.72$

So if the blood test comes back negative ($y = 0$), then a_0 has a lower risk than a_1, i.e. the patient should be sent home

And if the blood test comes back positive ($y = 1$), then a_1 has a lower risk than a_0, i.e. the patient should be given antibiotics

Classification

A particular type of decision problem which often occurs is trying to classify an object into one of two categories, based on some associated data (side note: in the machine learning literature, this is called either classification, or supervised learning).

Some examples:

- Classifying an email as being either **spam** or **not spam**
- Classifying a patient as being either **sick** or **healthy**
- Classifying a particular earthquake scenario as **worth evacuating the village** or **not worth evacuating the village**
- Classifying a stock as being **worth buying** or **not worth buying**

The previous example was a specific instance of this. We will now treat it more generally

Classification

Assume that the object can have one of two classes $\theta \in \{0, 1\}$. There are two actions a_0 and a_1, corresponding to the decision to allocate the object to class 0 and 1 respectively.

The data y has a (known) likelihood function $p(y|\theta)$, and loss function is $L(\theta, a_i)$.

As before, we take the action which minimises the risk. We allocate the object to class 0 if $R(a_0|y) < R(a_1|y)$, i.e if:

$$\int L(\theta, a_0)p(\theta|y)d\theta < \int L(\theta, a_1)p(\theta|y)d\theta$$

and to class 1 otherwise.

Classification - Example

A company produces widgets on an assembly line. Due to inherent defects in the manufacturing process, each widget has a probability 0.01 of being defective. The company does not want to send too many defective widgets to the market.

The widgets are produced in batches of 10,000. For each batch, there is a chance that the manufacturing process can go drastically wrong, in which case each of the widgets in the batch has probability 0.05 of being defective (which is five times the usual probability)

Ideally the company would test each widget individually to find whether it is defective. However testing widgets is expensive. So instead the company randomly selects 100 widgets from each batch, and tests only these. The goal is to determine whether each particular batch is bad (i.e.has a defective rate of 0.05 rather than 0.01). If a batch is bad, it is thrown out, otherwise it is sent to the market to be sold

Classification - Example

For a particular batch, the company selects 100 widgets at random and tests them. Of these, $y = 3$ are found to be defective.

Question: Does observing $y = 3$ justify concluding that the batch is bad, and throwing out the batch?

Classification - Example

For a particular batch, the company selects 100 widgets at random and tests them. Of these, $y = 3$ are found to be defective.

Question: Does observing $y = 3$ justify concluding that the batch is bad, and throwing out the batch?

Answer: like all decision making, this depends on the relative costs, i.e. on the loss function $L(\theta, a_i)$. Without specifying this, the question cannot be answered

Classification - Example

There are two classes of batch, good and bad. These correspond to $\theta = 0$ and $\theta = 1$ respectively. Actions a_0 and a_1 correspond to classifying the batch as good (and keeping it) and classifying the batch as bad (and throwing it out).

The company estimates the cost of sending a bad batch to market as being equal to 20 times the cost of throwing out a good batch (due to the cost of potential lawsuits, replacing defective products, etc).

The loss function is hence:

	$\theta = 0$	$\theta = 1$
a_0	0	20
a_1	1	0

Classification - Example

Finally, based on previous experience, the company knows that only 0.3% of batches are bad. The prior is hence $p(\theta = 0) = 0.997$.

This is all the information needed to classify a batch as good/bad based on observing y defectives out of the 100 items sampled.

As before, we first compute the posterior distribution $p(\theta|y)$ for both values of θ.

Classification - Example - Posterior

If $\theta = 0$ then $p(y|\theta = 0)$ is a Binomial$(100, 0.01)$ distribution. Similarly if $\theta = 1$ then it is Binomial$(100, 0.05)$. The posteriors are hence:

$$p(\theta = 0|y) = \frac{p(y|\theta = 0)p(\theta = 0)}{p(y)} = \frac{0.997\binom{100}{3}0.01^3 0.99^{97}}{p(y)}$$

$$p(\theta = 1|y) = \frac{p(y|\theta = 1)p(\theta = 1)}{p(y)} = \frac{0.003\binom{100}{3}0.05^3 0.95^{97}}{p(y)}$$

and:

$$p(y) = p(y|\theta = 0)p(\theta = 0) + p(y|\theta = 1)p(\theta = 1) =$$

$$= 0.997 \times \binom{100}{3}0.01^3 0.99^{97} + 0.003 \times \binom{100}{3}0.05^3 0.95^{97}$$

$$= 0.061$$

(remember you can use the dnorm() R function to compute the $p(y|\theta)$ parts quickly)

Classification - Example - Posterior

Substituting in $p(y)$ gives the posterior:

$$p(\theta = 0|y) = 0.993162$$

$$p(\theta = 1|y) = 0.006838$$

(note of course that these must sum to 1, since there are only two possible values for θ – this will help you check your algebra!)

Classification - Example - Risk

We can now compute the risk associated with each action:

$$R(a_0|y) = p(\theta = 0|y)L(\theta = 0, a_0) + p(\theta = 1|y)L(\theta = 1, a_0) =$$
$$= 0 + 0.006838 \times 20 = 0.1367$$

and

$$R(a_1|y) = p(\theta = 0|y)L(\theta = 0, a_1) + p(\theta = 1|y)L(\theta = 1, a_1) =$$
$$= 0.993 \times 1 = 0.993$$

So we pick action a_0, i.e. classify the batch as good and send it to market

Classification - Multi-class

In some cases there may be more than one class that the object can be allocated to. Lets say the number of possible classes is k. Examples:

- When designing an earthquake prediction system, we may wish to classify future earthquakes as **minor**, **medium**, or **major** based on the evidence ($k = 3$)
- When designing handwriting recognition systems, numerical digits can be classified as '0', '1', '2',...'9' ($k = 10$)

In principle this is the same as in the two class case: we number the available classes from 1 to k, associate an action a_i with each class, and choose the action which minimises the risk. E.g. if $k = 3$, we allocate to class 1 if $R(a_1) < R(a_2)$ and $R(a_1) < R(a_3)$.

Classification - Multi-class - Loss function

Specifying the loss function in the multiclass case can be difficult. Remember: with k classes we need to specify the loss associated with each wrong decision – what is the loss of allocation a class 1 object to class 2? And the cost of allocating it to class 3?

To keep things simple, a $0-1$ loss function is often assumed:

$$L(\theta, a_i) = \begin{cases} 0 & \text{if } \theta = i \\ 1 & \text{if } \theta \neq i \end{cases}$$

i.e. the loss is 1 if the object is allocated to the wrong class, otherwise it is 0

Classification - Multi-class - Decision Rule

In this case it is easy to show that the risk is minimised if we allocate the object to the class for which the posterior is highest

I.e we allocate to the class i for which $p(\theta = i|y)$ is largest.

So compute $p(\theta = 1|y), p(\theta = 2|y), \ldots, p(\theta = k|y)$, and go with the class for which this is largest

Parameter Estimation Using Decision Theory

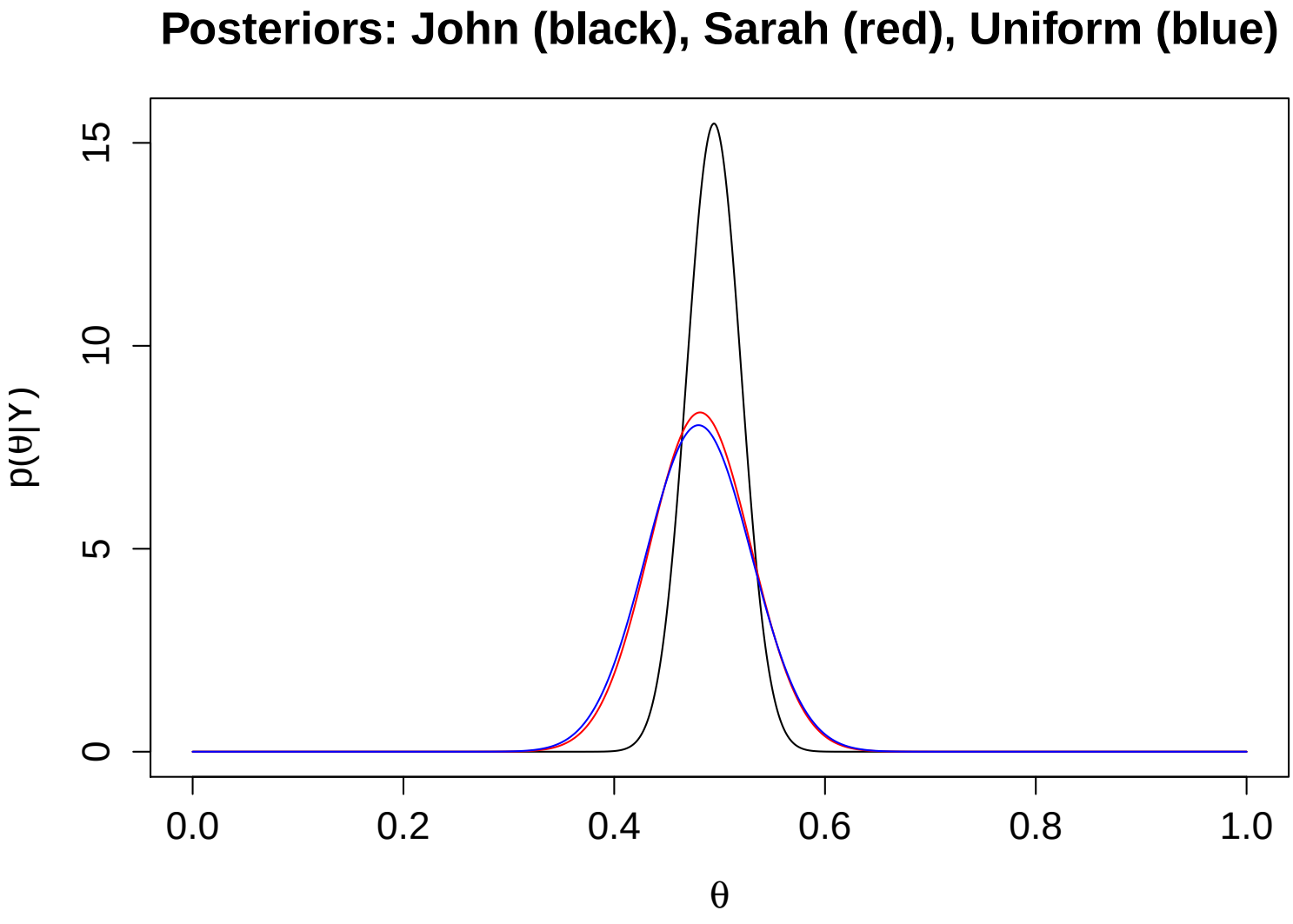
Parameter Estimation

We saw in the previous lecture how Bayesian inference was used for estimating the parameters of a probability distribution.

If y has (e.g) a Binomial(n, θ) or Exponenital(θ) distribution then we find the posterior distribution for θ by combining our prior knowledge about θ with the likelihood.

E.g. we considered the example where a (possibly biased) coin was tossed 100 times and produced 48 heads. John and Sarah both had different priors for the bias θ of the coin (i.e. the probability of landing heads)

John and Sarah's Posteiror



Parameter Estimation

In practice we may need to pick our single "best guess" for θ. I.e. rather than using the full posterior distribution $p(\theta|y)$, we may want a single point estimate $\hat{\theta}$.

We must hence summarise the posterior distribution by a single number. How to do this?

You will previously have been taught to do this by (e.g) estimating θ by the maximum of the likelihood function. However this does not take costs into account!

In Bayesian decision theory, the way in which we summarise the posterior depends on our particular choice of loss function

Parameter Estimation - Loss Function

Here we make e **decision** to estimate θ using the estimate $\hat{\theta}$. In terms of actions, we now have a (possibly infinite) set where action a_i corresponds to estimating θ by $\hat{\theta} = i$.

The loss function $L(\theta, \hat{\theta})$ defines the loss incurred if we estimate the true value of θ by $\hat{\theta}$

As before, we want to choose the estimate $\hat{\theta}$ to minimise the expected loss.

Parameter Estimation - Loss Function

There are three popular loss functions:

- Squared loss: $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$
- Absolute loss: $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$
- Binary loss: $L(\theta, \hat{\theta}) = \begin{cases} 1 & \text{if } \theta \neq \hat{\theta} \\ 0 & \text{if } \theta = \hat{\theta} \end{cases}$

The loss function we choose depends on the problem. For example, binary loss means we only care about getting θ exactly right, and any deviation is equally bad (usually more sensible when θ is discrete). The difference between absolute and squared loss is that squared loss punishes big mistakes more, which should lead to a more conservative estimate.

Parameter Estimation

The following (very elegant!) results can be proved (see supplementary material on moodle for a proof):

- The squared loss is minimised if $\hat{\theta}$ is chosen to be the posterior **mean** $\hat{\theta} = \int \theta p(\theta|y) d\theta$
- The absolute loss is minimised if $\hat{\theta}$ is chosen to be the posterior **median**
- The binary 0-1 loss is minimised if $\hat{\theta}$ is chosen to be the posterior **mode**, $\hat{\theta} = \max_\theta p(\theta|y)$

So all three of the intuitively sensible ways of summarising the posterior we might have tried (mean, median, mode) can be shown to correspond to optimal estimators of θ under different choices of the loss function.

This is a very nice result because it means our estimates are **principled**. We aren't just (e.g.) using the posterior mean because it feels like a sensible estimate, it is actually the best possible estimate under squared error loss! Other procedures like maximum likelihood typically do not have this property.

Proof

We prove the squared loss case only, the other two are similar.

$$R(\hat{\theta}) = \int (\theta - \hat{\theta})^2 p(\theta|y) d\theta =$$

$$= \hat{\theta}^2 \int p(\theta|y) d\theta - 2\hat{\theta} \int \theta p(\theta|y) d\theta + \int \theta^2 p(\theta|y) d\theta =$$

$$= \hat{\theta}^2 - 2\hat{\theta} E[\theta] + E[\theta^2]$$

The first term integrated to 1 since it is a probability distributions. The expectations are with respect to $p(\theta|y)$. Recall that $Var(\theta) = E[\theta^2] - E[\theta]^2$, so the above becomes:

$$= \hat{\theta}^2 - 2\hat{\theta} E[\theta] + E[\theta]^2 + E[(\theta - E[\theta])^2] =$$

$$= (\hat{\theta} - E[\theta])^2 + E[(\theta - E[\theta])^2]$$

which is minimised when $\hat{\theta} = E[\theta]$