Lecture 2: Statistical Decision Theory I

Alex Donov

18 January 2019

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 patients 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?

Two Types of Decision Theory

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

There are two main types of decision theory:

- *Frequentist decision theory* - usually concerned with making decisions that perform best in the "worst possible case" (minimax).
- *Bayesian decision theory* - concerned with making decisions that perform best, based on the information we have about the unknowns.

Today, we will cover *Bayesian decision theory*. Next week, we will consider *Frequentist decision theory*.

Basic Elements of a Decision Problem

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_1), or give the patient medicine (a_2).
- 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).

Basic Elements of a Decision Problem

- $\boldsymbol{\Theta}$ is the *parameter space* which consists of all possible "states of nature" or "states of the world" θ, only one of which will occur. The "true" state of nature θ is unknown.

For example, let θ represent whether a patient has a particular disease.

In this case:

$$\boldsymbol{\Theta} = \{0, 1\} \tag{1}$$

where $\theta = 0$ if a patient is healthy
$\theta = 1$ if a patient has a disease

Basic Elements of a Decision Problem

- $\mathscr{A} = (a_1, a_2, \ldots, a_k)$ is the *action space*, which is the set of all possible actions available, $a \in \mathscr{A}$.

For example, a doctor can choose a particular action a: to treat the patient or send the patient home.

In this case:

$$\mathscr{A} = \{0, 1\} \tag{2}$$

where $a_1 = 0$ send patient home without giving medicine
$a_2 = 1$ prescribe medicine

Basic Elements of a Decision Problem

- Ω contains all possible realisations $y \in \Omega$ of a random variable Y which belongs to the family $\{f(y;\, \theta) | \, \theta \in \Theta\}$

For example, a doctor can perform a blood test and obtain an outcome y, which could be positive or negative:

In this case:

$$\Omega = \{0, 1\} \tag{3}$$

where $y_1 = 0$ the test comes back negative
$y_2 = 1$ the test comes back positive

Basic Elements of a Decision Problem

- $L(\theta, a)$ is a loss function that has the domain $\Theta \times \mathscr{A} = \{(\theta, a) | \, \theta \in \Theta$ and $a \in \mathscr{A}\}$ and codomain $\mathbb{R}$. That is, a loss function maps each combination of states of the world θ and action a onto a numerical loss, $\mathbb{R}$. For technical convenience, $L(\theta, a) \geq -K > -\infty$

For example, a patient is healthy $\theta = 0$, and doctor sends him home without giving medicine $a_1 = 0$, and the loss incurred $L(\theta, a_1) = L(0, 0) = 0$.

In this case:

$$\Theta = \{0, 1\}$$

and

$$\mathscr{A} = \{0, 1\}$$

then

$$\Theta \times \mathscr{A} = \{\underbrace{(0,0)}, (0,1), (1,0), (1,1)\}$$

Loss Function

- The **loss function** $L(\theta, a)$ is a core element of decision making which represents the loss incurred if we choose action a when the true state of the world is θ (usually unknown).

- The losses corresponding to each action and state of world θ can be represented by a *loss matrix*:

	$\theta = 0$	$\theta = 1$
a_1	0	10
a_2	1	0

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

Risk

- In practice, the decision-maker does not know the true θ. The state of the world is uncertain.
- The natural procedure is to consider the "expected" loss of making a decision.

Definition

If $\pi^*(\theta)$ is the believed probability distribution of θ at the time of decision making, the **Bayesian expected loss** of an action a is:

$$\rho(\pi^*, a) = E^{\pi^*} L(\theta, a) = \int_\Theta L(\theta, a) dF^{\pi^*}(\theta)$$

A little note on Notation

Let $F(x)$ be the cumulative distribution function associated with the random variable X. Then the expected value of a random variable X is defined by:

$$E(X) = \int_{-\infty}^{\infty} x \, dF(x) = \begin{cases} \sum_{x \in R(X)} x f(x) & \text{(discrete)} \\ \int_{-\infty}^{\infty} x f(x) dx & \text{(continuous)} \end{cases} \tag{4}$$

where

$$R(X) \equiv \{x : f(x) > 0 \text{ for } x \in \mathbb{R}\}$$

The conditional Bayes Principle

- Once the Bayesian expected loss $\rho(\pi^*, a)$ has been determined for each a, one can easily choose an optimal action.

The conditional Bayes Principle

Choose an action $a \in \mathscr{A}$ which minimizes $\rho(\pi^*, a)$ (assuming the minimum is attained). Such an action will be called a *Bayes action* and will be denoted a^{π^*}.

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_1: send the patient home with no medication
- a_2: 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 meningitis.

A Medical Example - Costs

- However, the costs are **not** equal.
- Recall the *loss matrix*:

	$\theta = 0$	$\theta = 1$
a_1	0	10
a_2	1	0

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

A Medical Example - Prior Risk

- Suppose the doctor has no data, so he computes the *Bayesian expected loss* using only his prior knowledge (denoted by π instead of π^*).
- The *Bayesian expected loss* associated with action a_1 (sending the patient home) is:

$$\rho(\pi, a_1) = p(\theta = 0) \times L(\theta = 0, a_1) \; + \; p(\theta = 1) \times L(\theta = 1, a_1) =$$

$$= 0.9 \times 0 \; + \; 0.1 \times 10 = 1$$

- Similarly, the *Bayesian expected loss* associated with action a_2 (prescribing antibiotics) is:

$$\rho(\pi, a_2) = p(\theta = 0) \times L(\theta = 0, a_2) \; + \; p(\theta = 1) \times L(\theta = 1, a_2) =$$

$$= 0.9 \times 1 \; + \; 0.1 \times 0 = 0.9$$

- What is the doctor's optimal decision?

A Medical Example - Data

- In practice, a doctor 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)

A Medical Example - Data

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

$$p(y=0|\theta=0)=0.8 \quad p(y=1|\theta=0)=0.2 \quad \text{(no meningitis case)}$$
$$p(y=0|\theta=1)=0.3 \quad p(y=1|\theta=1)=0.7 \quad \text{(meningitis case)}$$

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
  - the likelihood $p(y|\theta)$.
- Based on all this, the doctor seeks to make an optimal decision.
- That is, choose an action $a \in \mathscr{A}$ which minimizes the *Bayesian expected loss*, $\rho(\pi^*, a|y)$.

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.

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

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

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

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

A Medical Example - Computing The Posterior Distributions

- We use the Law of Total Probability to compute $p(y = 0)$ and $p(y = 1)$:

$$\begin{aligned} p(y = 0) &= p(y = 0|\theta = 0) \times p(\theta = 0) \ + \ p(y = 0|\theta = 1) \times p(\theta = 1) \\ &= 0.8 \times 0.9 \ + \ 0.3 \times 0.1 = 0.75 \end{aligned}$$

and

$$\begin{aligned} p(y = 1) &= p(y = 1|\theta = 0) \times p(\theta = 0) \ + \ p(y = 1|\theta = 1) \times p(\theta = 1) \\ &= 0.2 \times 0.9 \ + \ 0.7 \times 0.1 = 0.25 \end{aligned}$$

A Medical Example - Computing The Posterior Distributions

- Substituting these back in gives:

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

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

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

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

- This completes the computation of the posterior distribution.

A Medical Example - Computing The Risk

- We can now compute the Bayesian expected loss, $\rho(\pi^*, a|y)$ for each action a and value of y. First, let's consider a_1

$$\begin{aligned}
\rho(\pi^*, a_1|y) &= \sum_{\Theta} p(\theta|y)L(\theta, a_1) \\
&= p(\theta = 0|y) \times L(\theta = 0, a_1) \; + \; p(\theta = 1|y) \times L(\theta = 1, a_1) \\
&= p(\theta = 0|y) \times 0 \; + \; p(\theta = 1|y) \times 10 \\
&= 10 \times p(\theta = 1|y)
\end{aligned}$$

So:

$$\rho(\pi^*, a_1|y = 0) = 10 \times 0.04 = 0.4 \tag{5}$$

$$\rho(\pi^*, a_1|y = 1) = 10 \times 0.28 = 2.8 \tag{6}$$

A Medical Example - Computing The Risk

- Similarly for a_2:

$$\begin{aligned}\rho(\pi^*, a_2|y) &= \sum_{\Theta} p(\theta|y)L(\theta, a_2) \\ &= p(\theta = 0|y)L(\theta = 0, a_2) \; + \; p(\theta = 1|y)L(\theta = 1, a_2) \\ &= p(\theta = 0|y) \times 1 \; + \; p(\theta = 1|y) \times 0 \\ &= p(\theta = 0|y)\end{aligned}$$

So:

$$\rho(\pi^*, a_2|y = 0) = 0.96 \tag{7}$$

$$\rho(\pi^*, a_2|y = 1) = 0.72 \tag{8}$$

A Medical Example - Final Decision

In summary:

- $\rho(\pi^*, a_1|y=0) = 0.4$
- $\rho(\pi^*, a_2|y=0) = 0.96$
- $\rho(\pi^*, a_1|y=1) = 2.8$
- $\rho(\pi^*, a_2|y=1) = 0.72$

- So if the blood test comes back negative ($y=0$), then a_1 has a lower *Bayesian expected loss* than a_2, i.e. the patient should be sent home

- And if the blood test comes back positive ($y=1$), then a_2 has a lower *Bayesian expected loss* than a_1, i.e. the patient should be given antibiotics

Example

- Consider a drug company deciding whether or not to market a new pain reliever.
- One of the many factors affecting its decision is the proportion of the market θ the drug will capture.
- Hence, the company desires to estimate θ.
- Since θ is a proportion, it is clear that:

$$\Theta = \{\theta : 0 \leq \theta \leq 1\} = [0, 1]$$

- Since the goal is to estimate θ, the action taken is simply the choice of a number as an estimate for θ. Hence $\mathscr{A} = [0, 1]$.

Example

- If a company underestimates demand, i.e. $\theta - a \geq 0$, the loss incurred is $\theta - a$.
- However, if a company overestimates demand, i.e. $\theta - a \leq 0$, then it is twice as costly as underestimating the demand with the cost being $2(a - \theta)$.
- Assume no data is obtained, however there is a prior information about θ arising from previous introductions of new similar drugs into the market.
- In the past drugs tended to capture between $\frac{1}{10}$ and $\frac{1}{5}$ of the market, with all values between $\frac{1}{10}$ and $\frac{1}{5}$ being equally likely.

Example

- The prior density $p(\theta)$ that will reflect this prior information is as follows:

$$p(\theta) = 10\mathbb{I}_{\{0.1,0.2\}}(\theta)$$

- The *Bayesian expected loss* $\rho(\pi, a)$ of an action a is:

$$\begin{aligned}
\rho(\pi, a) &= \int_0^1 L(\theta, a)p(\theta)d\theta \\
&= \int_0^a 2(a-\theta)10\mathbb{I}_{\{0.1,0.2\}}(\theta)d\theta + \int_a^1 (\theta-a)10\mathbb{I}_{\{0.1,0.2\}}(\theta)d\theta \\
&= \begin{cases} 0.15 - a & \text{if} \quad a \leq 0.1 \\ 15a^2 - 4a + 0.3 & \text{if} \quad 0.1 \leq a \leq 0.2 \\ 2a - 0.3 & \text{if} \quad a \geq 0.2 \end{cases}
\end{aligned}$$

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.

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**

Classification

- Assume that the object can have one of two classes $\theta \in \{0, 1\}$. There are two actions a_1 and a_2, 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 $L(\theta, a)$.
- As before, we take the action which minimises the *Bayesian expected loss*. We allocate the object to class 0 if $\rho(\pi^*, a_1|y) < \rho(\pi^*, a_2|y)$, i.e if:

$$\int_\Theta L(\theta, a_1) dF^{\pi^*}(\theta) < \int_\Theta L(\theta, a_2) dF^{\pi^*}(\theta)$$

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.

Classification - Example

- 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?
- **Answer**: like all decision making, this depends on the relative costs, i.e. on the loss function $L(\theta, a)$. 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_1 and a_2 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 matrix* is hence:

	$\theta = 0$	$\theta = 1$
a_1	0	20
a_2	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 or 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:

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

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

and:

$$\begin{aligned} p(y) &= p(y|\theta = 0) \times p(\theta = 0) \quad + \quad p(y|\theta = 1) \times p(\theta = 1) \\ &= 0.997 \times \binom{100}{3} 0.01^3 0.99^{97} \quad + \quad 0.003 \times \binom{100}{3} 0.05^3 0.95^{97} \\ &= 0.061 \end{aligned}$$

Classification - Example - Posterior

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

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

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

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

Classification - Example

- We can now compute the *Bayesian expected loss* associated with each action:

$$\begin{aligned}\rho(\pi^*, a_1|y) &= p(\theta = 0|y)L(\theta = 0, a_1) \quad + \quad p(\theta = 1|y)L(\theta = 1, a_1) \\ &= 0 + 0.006838 \times 20 \\ &= 0.1367\end{aligned}$$

and

$$\begin{aligned}\rho(\pi^*, a_2|y) &= p(\theta = 0|y)L(\theta = 0, a_2) \quad + \quad p(\theta = 1|y)L(\theta = 1, a_2) \\ &= 0.993 \times 1 \\ &= 0.993\end{aligned}$$

- So we pick action a_1, i.e. classify the batch as good and send it to market.

Classification - Multiclass

- 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' ($k = 10$)

Classification - Multiclass

- 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 *Bayesian expected loss*.
- For example, if $k = 3$, we allocate to class 1 if $\rho(\pi^*, a_1) < \rho(\pi^*, a_2)$ and $\rho(\pi^*, a_1) < \rho(\pi^*, a_3)$.

Classification - Multiclass - 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:

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

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

Classification - Multiclass - Decision Rule

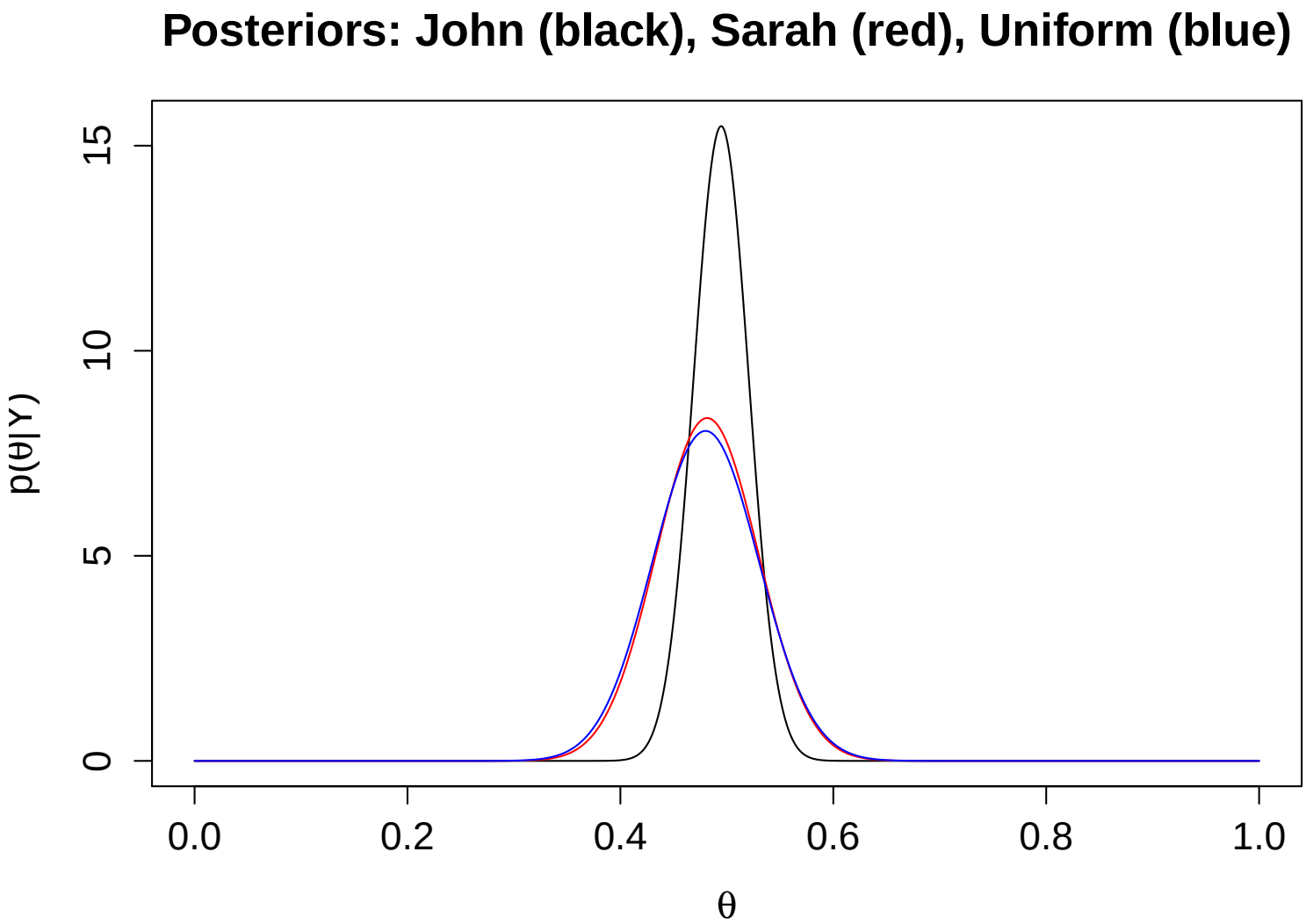
- In this case it is easy to show that the risk is minimised if we allocate the object to the class for which the posterior is highest. (Why?)
- That is, we allocate to the class i for which $p(\theta = i|y)$ is largest.
- So compute $p(\theta = 1|y)$, $p(\theta = 2|y)$, $\ldots$, $p(\theta = k|y)$, 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.
- In practice we may need to pick our single "best guess" for θ. That is, 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 can we do this?

John and Sarah's Posteiror



Parameter Estimation

- You have probably been taught how to do this from the frequentist perspective.
- For example, estimating θ by choosing the value of $\widehat{\theta}$ that maximises the *likelihood function*.
- 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 a **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 results can be proved;

- 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**, $\hat{\theta} = \max_\theta p(\theta|y)$

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!

Proof

- The proof for the squared loss function is as follows:

$$\begin{aligned}\rho(\pi^*, \hat{\theta}|y) &= \int (\theta - \hat{\theta})^2 p(\theta|y) d\theta \\ &= \hat{\theta}^2 \int p(\theta|y) d\theta - 2\hat{\theta} \int \theta p(\theta|y) d\theta + \int \theta^2 p(\theta|y) d\theta \\ &= \hat{\theta}^2 - 2\hat{\theta} E[\theta] + E[\theta^2]\end{aligned}$$

- Using the result $Var(\theta) = E[\theta^2] - E[\theta]^2$, and rearranging:

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

It can be seen that $\rho(\pi^*, \hat{\theta}|y)$ is minimised when $\hat{\theta} = E[\theta]$