Introduction to Machine Learning

Bayesian Decision Theory

The 7th lecture, 09.03.2022

Phuc Loi Luu, PhD
p.luu@garvan.org.au
luu.p.loi@googlemail.com

Roadmap for today

- Joint Distributions: Two Discrete Random Variables
- Toy example: Predict gender of a person based on height
- Bayes Decision Theory

Conditional Expectation as a Function of a Random Variable

Remember that the conditional expectation of X given that $Y = y$ is given by

$$E[X|Y = y] = \sum_{x_i \in R_X} x_i P_{X|Y}(x_i|y).$$

Note that $E[X|Y = y]$ depends on the value of y. In other words, by changing y, $E[X|Y = y]$ can also change. Thus, we can say $E[X|Y = y]$ is a function of y, so let's write

$$g(y) = E[X|Y = y].$$

Thus, we can think of $g(y) = E[X|Y = y]$ as a function of the value of random variable Y. We then write

$$g(Y) = E[X|Y].$$

We use this notation to indicate that $E[X|Y]$ is a random variable whose value equals $g(y) = E[X|Y = y]$ when $Y = y$. Thus, if Y is a random variable with range $R_Y = \{y_1, y_2, \cdots\}$, then $E[X|Y]$ is also a random variable with

$$E[X|Y] = \begin{cases} E[X|Y = y_1] & \text{with probability } P(Y = y_1) \\ E[X|Y = y_2] & \text{with probability } P(Y = y_2) \\ . & . \\ . & . \\ . & . \end{cases}$$

Conditional Expectation as a Function of a Random Variable

Example 5.10 Let $X = aY + b$. Then $E[X|Y = y] = E[aY + b|Y = y] = ay + b$. Here, we have $g(y) = ay + b$, and therefore,

$$E[X|Y] = aY + b,$$

which is a function of the random variable Y.

Since $E[X|Y]$ is a random variable, we can find its PMF, CDF, variance, etc. Let's look at an example to better understand $E[X|Y]$.

Example 5.11 Consider two random variables X and Y with joint PMF given in Table 5.2. Let $Z = E[X|Y]$.

a. Find the Marginal PMFs of X and Y.
b. Find the conditional PMF of X given $Y = 0$ and $Y = 1$, i.e., find $P_{X|Y}(x|0)$ and $P_{X|Y}(x|1)$.
c. Find the PMF of Z.
d. Find EZ, and check that $EZ = EX$.
e. Find Var(Z).

Table 5.2: Joint PMF of X and Y in example 5.11

	$Y = 0$	$Y = 1$
$X = 0$	$\frac{1}{5}$	$\frac{2}{5}$
$X = 1$	$\frac{2}{5}$	0

Conditional Expectation as a Function of a Random Variable

a. Using the table we find out

$$P_X(0) = \frac{1}{5} + \frac{2}{5} = \frac{3}{5},$$
$$P_X(1) = \frac{2}{5} + 0 = \frac{2}{5},$$
$$P_Y(0) = \frac{1}{5} + \frac{2}{5} = \frac{3}{5},$$
$$P_Y(1) = \frac{2}{5} + 0 = \frac{2}{5}.$$

Thus, the marginal distributions of X and Y are both $Bernoulli(\frac{2}{5})$. However, note that X and Y are not independent.

b. We have

$$P_{X|Y}(0|0) = \frac{P_{XY}(0,0)}{P_Y(0)}$$
$$= \frac{\frac{1}{5}}{\frac{3}{5}} = \frac{1}{3}.$$

Thus,

$$P_{X|Y}(1|0) = 1 - \frac{1}{3} = \frac{2}{3}.$$

We conclude

$$X|Y=0 \sim Bernoulli\left(\frac{2}{3}\right).$$

Similarly, we find

$$P_{X|Y}(0|1) = 1,$$
$$P_{X|Y}(1|1) = 0.$$

Thus, given $Y = 1$, we have always $X = 0$.

c. We note that the random variable Y can take two values: 0 and 1. Thus, the random variable $Z = E[X|Y]$ can take two values as it is a function of Y. Specifically,

$$Z = E[X|Y] = \begin{cases} E[X|Y=0] & \text{if } Y = 0 \\ E[X|Y=1] & \text{if } Y = 1 \end{cases}$$

Now, using the previous part, we have

$$E[X|Y=0] = \frac{2}{3}, \quad E[X|Y=1] = 0,$$

and since $P(y=0) = \frac{3}{5}$, and $P(y=1) = \frac{2}{5}$, we conclude that

$$Z = E[X|Y] = \begin{cases} \frac{2}{3} & \text{with probability } \frac{3}{5} \\ 0 & \text{with probability } \frac{2}{5} \end{cases}$$

So we can write

$$P_Z(z) = \begin{cases} \frac{3}{5} & \text{if } z = \frac{2}{3} \\ \frac{2}{5} & \text{if } z = 0 \\ 0 & \text{otherwise} \end{cases}$$

d. Now that we have found the PMF of Z, we can find its mean and variance. Specifically,

$$E[Z] = \frac{2}{3} \cdot \frac{3}{5} + 0 \cdot \frac{2}{5} = \frac{2}{5}.$$

We also note that $EX = \frac{2}{5}$. Thus, here we have

$$E[X] = E[Z] = E[E[X|Y]].$$

Conditional Expectation as a Function of a Random Variable

e. To find $\text{Var}(Z)$, we write

$$\begin{aligned}\text{Var}(Z) &= E[Z^2] - (EZ)^2 \\ &= E[Z^2] - \frac{4}{25},\end{aligned}$$

where

$$E[Z^2] = \frac{4}{9} \cdot \frac{3}{5} + 0 \cdot \frac{2}{5} = \frac{4}{15}.$$

Thus,

$$\begin{aligned}\text{Var}(Z) &= \frac{4}{15} - \frac{4}{25} \\ &= \frac{8}{75}.\end{aligned}$$

Example 5.12

Let X and Y be two random variables and g and h be two functions. Show that

$$E[g(X)h(Y)|X] = g(X)E[h(Y)|X].$$

Conditional Expectation as a Function of a Random Variable

Example 5.12

Let X and Y be two random variables and g and h be two functions. Show that

$$E[g(X)h(Y)|X] = g(X)E[h(Y)|X].$$

Solution

Note that $E[g(X)h(Y)|X]$ is a random variable that is a function of X. In particular, if $X = x$, then $E[g(X)h(Y)|X] = E[g(X)h(Y)|X = x]$. Now, we can write

$$\begin{aligned} E[g(X)h(Y)|X = x] &= E[g(x)h(Y)|X = x] \\ &= g(x)E[h(Y)|X = x] \qquad (\text{since } g(x) \text{ is a constant}). \end{aligned}$$

Thinking of this as a function of the random variable X, it can be rewritten as $E[g(X)h(Y)|X] = g(X)E[h(Y)|X]$. This rule is sometimes called "taking out what is known." The idea is that, given X, $g(X)$ is a known quantity, so it can be taken out of the conditional expectation.

$$E[g(X)h(Y)|X] = g(X)E[h(Y)|X] \qquad (5.6)$$

Conditional Variance

Similar to the conditional expectation, we can define the conditional variance of X, $\text{Var}(X|Y=y)$, which is the variance of X in the conditional space where we know $Y=y$. If we let $\mu_{X|Y}(y)=E[X|Y=y]$, then

$$\begin{aligned}\text{Var}(X|Y=y) &= E\big[(X-\mu_{X|Y}(y))^2|Y=y\big] \\ &= \sum_{x_i\in R_X} \big(x_i-\mu_{X|Y}(y)\big)^2 P_{X|Y}(x_i) \\ &= E\big[X^2|Y=y\big]-\mu_{X|Y}(y)^2.\end{aligned}$$

Note that $\text{Var}(X|Y=y)$ is a function of y. Similar to our discussion on $E[X|Y=y]$ and $E[X|Y]$, we define $\text{Var}(X|Y)$ as a function of the random variable Y. That is, $\text{Var}(X|Y)$ is a random variable whose value equals $\text{Var}(X|Y=y)$ whenever $Y=y$. Let us look at an example.

Example 5.13

Let X, Y, and $Z=E[X|Y]$ be as in Example 5.11. Let also $V=\text{Var}(X|Y)$.

a. Find the PMF of V.
b. Find EV.
c. Check that $\text{Var}(X)=E(V)+\text{Var}(Z)$.

Conditional Variance, example

In Example 5.11, we found out that $X, Y \sim Bernoulli(\frac{2}{5})$. We also obtained

$$\begin{aligned} X|Y=0 &\sim Bernoulli\left(\frac{2}{3}\right), \\ P(X=0|Y=1) &= 1, \\ \mathrm{Var}(Z) &= \frac{8}{75}. \end{aligned}$$

a. To find the PMF of V, we note that V is a function of Y. Specifically,

$$V = \mathrm{Var}(X|Y) = \begin{cases} \mathrm{Var}(X|Y=0) & \text{if } Y=0 \\ \mathrm{Var}(X|Y=1) & \text{if } Y=1 \end{cases}$$

Therefore,

$$V = \mathrm{Var}(X|Y) = \begin{cases} \mathrm{Var}(X|Y=0) & \text{with probability } \frac{3}{5} \\ \mathrm{Var}(X|Y=1) & \text{with probability } \frac{2}{5} \end{cases}$$

Now, since $X|Y=0 \sim Bernoulli\left(\frac{2}{3}\right)$, we have

$$\mathrm{Var}(X|Y=0) = \frac{2}{3}\cdot\frac{1}{3} = \frac{2}{9},$$

and since given $Y=1, X=0$, we have

$$\mathrm{Var}(X|Y=1) = 0.$$

Thus,

$$V = \mathrm{Var}(X|Y) = \begin{cases} \frac{2}{9} & \text{with probability } \frac{3}{5} \\ 0 & \text{with probability } \frac{2}{5} \end{cases}$$

So we can write

$$P_V(v) = \begin{cases} \frac{3}{5} & \text{if } v = \frac{2}{9} \\ \frac{2}{5} & \text{if } v = 0 \\ 0 & \text{otherwise} \end{cases}$$

b. To find EV, we write

$$EV = \frac{2}{9}\cdot\frac{3}{5} + 0\cdot\frac{2}{5} = \frac{2}{15}.$$

c. To check that $\mathrm{Var}(X) = E(V) + \mathrm{Var}(Z)$, we just note that

$$\begin{aligned} \mathrm{Var}(X) &= \frac{2}{5}\cdot\frac{3}{5} = \frac{6}{25}, \\ EV &= \frac{2}{15}, \\ \mathrm{Var}(Z) &= \frac{8}{75}. \end{aligned}$$

How to make decision in the prsence of uncertainty?

Decision theory is the science of making decisions. Contributions have come from a number of areas: philosophy, psychology, statistics, the social sciences, political science, economics, ... It has been mostly developed during the 20th century:

> Bernoulli (1738), de Finetti (1937), Ramsay (1931), Wald (1950), Lindley (1953), Savage (1954), Ferguson (1967), DeGroot (1970), Berger (1980,1985), Bernardo and Smith (1994), Robert (1994), ...

We make decisions all the time.

- The umbrella question—do I take umbrella today or not?
- Pregnancy and the risk of Down's syndrome.
- Should I fix or float my mortgage?
- Civil defence alerts.

How to make decision in the prsence of uncertainty?

Hypothetically, we could apply decision theory to any of these, but in many situations the difficulties may out-weigh the benefits.

In statistics, decision theory is concerned with quantifying the decision making process. We must choose a particular course of action from a set of possible decisions. The consequences of each possible decision depends on the "true" **state of nature** which we don't know. We collect data which provides information about the true state of nature and based on this we make a decision.

A **decision matrix** is often used to organise information. Consequences of alternative courses of action are tabulated against the possible "states of nature." For example,

	State of Nature: Size of Tsunami				
Action	0–0.5m	0.5–1m	1–2m	...	>5m
Issue Advisory					
Issue Warning					
Close Beaches					
Evacuation					

How to make decision in the prsence of uncertainty?

Example 7.1 (Game Show)**.** As part of a TV game show, suppose you need to choose between one of the following two options:

- Option A
  - win \$100,000 — probability $= 1$
- Option B
  - win \$500,000 — probability $= 0.10$
  - win \$100,000 — probability $= 0.89$
  - win \$0 — probability $= 0.01$

What would you do?

Now suppose you need to choose between one of the following two options:

- Option C
  - win \$100,000 — probability $= 0.11$
  - win \$0 — probability $= 0.89$
- Option D
  - win \$500,000 — probability $= 0.10$
  - win \$0 — probability $= 0.90$

What would you do? ◊

Elements of Decision Theory

Rice describes decision theory as "a mathematical approach for making decisions in face of uncertainty." Key elements are:

(a) An **action** a is chosen from a set of possible actions $\mathcal{A}$.

(b) This choice is based on observations from a **random variable** X which has a probability distribution that depends on the state of nature Θ.

(c) A **loss function** evaluates the consequences of the possible actions given the state of nature.

(d) A **decision rule** that maps the sample space of X into $\mathcal{A}$. Given a set of observations the decision rule identifies which action will be taken.

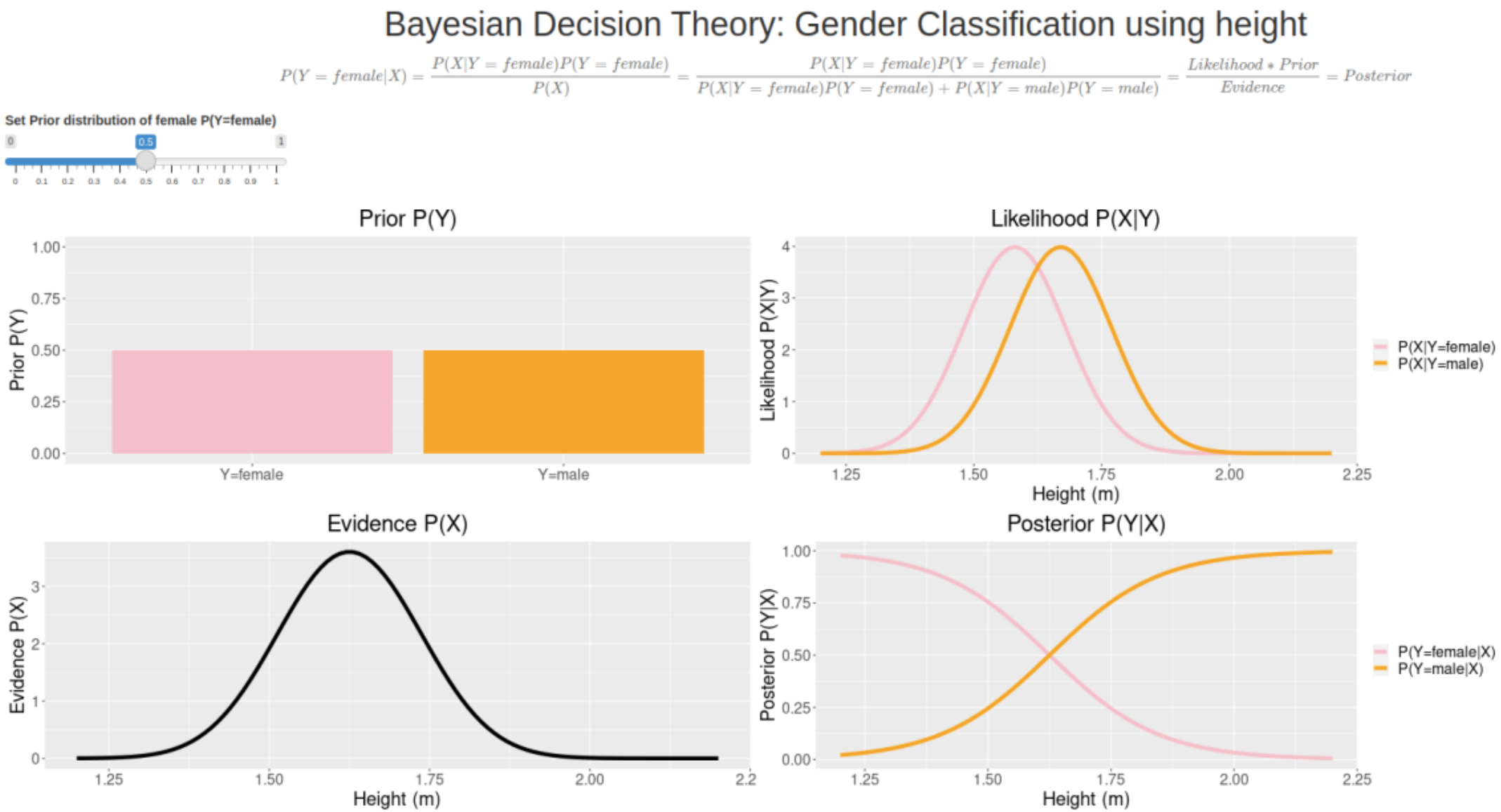
(e) The idea is to identify the **optimal decision rule**.

Toy example 1: Predict gender of a person based on height

- Setting: binary classification, that is Y = {male, female} = {M, F} = {0, 1} and X is height of a person
- The joint density p(x, y) of the probability measure P on X × Y can be decomposed as follows
  - The **class-conditional density** or **likelihood** p(x|y). It models the occurrence of the features x of class y.
  - The **conditional probability** p(y|x) or **Posterior**. The probability that we observe y given that the input is x. The most probable class y for the features x is then used for prediction.
  - The **marginal distribution or evidence** p(x). It models the cumulated occurrence of features x over all classes.
  - $$P(Y = female|X) = \frac{P(X|Y = female)P(Y = female)}{P(X)} = \frac{Likelihood * Prior}{Evidence} = Posterior$$ **or.**

$$P(X) = P(X|Y = female)P(Y = female) + P(X|Y = male)P(Y = male)$$

The effect of Prior on Posterior/Decision Rule



How to decide male of female: 1) Using Prior

- Derive Bayes decision rule $\alpha(x)$

$$\hat{y} = \max_{y} P(y)$$

$\rightarrow \alpha(x)$: P(Y=female) > P(Y=male) -> y = female

or $\alpha(x)$ can be write as a classifier $f_n(X)$:

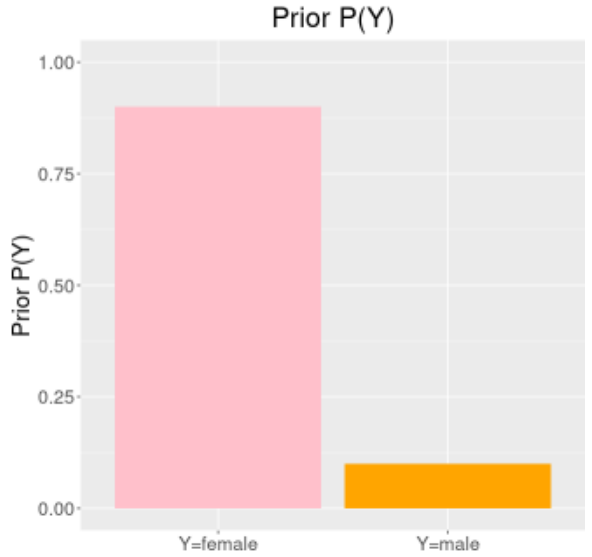
$$f_n(X) = \begin{cases} f & if\ P(Y=f) > P(Y=m) \\ m & otherwise \end{cases}$$

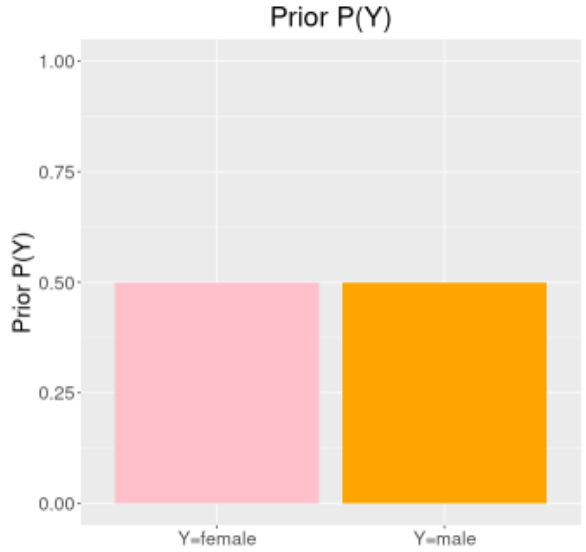
$$f_n(X) = \begin{cases} 1 & if\ P(Y=1) > P(Y=0) \\ 0 & otherwise \end{cases}$$

Example:

$\alpha(x)$: P(Y=female) = 0.9 > P(Y=male) = 0.1 -> y = female

- BUT, we always make the same decision even though we know that both female and male will appear.





How to decide male of female: 2) Using Maximum Likelihood

- Using Likelihood or Class conditional Probability: predict the label with the higher likelihood

$$\hat{y} = \max_y P(x|y)$$

-> $P(x|\hat{y}) >= P(x|y)$

-> $\alpha(x)$: $P(x|y=male) > P(x|y=female)$ -> y = male

$P(x|y=male) < P(x|y=female)$ -> y = female

-> or $\alpha(x)$ can be write as a classifier $f_n(X)$:

$$f_n(X) = \begin{cases} f & if\ P(x|y=f) > P(x|y=m) \\ m & otherwise \end{cases}$$

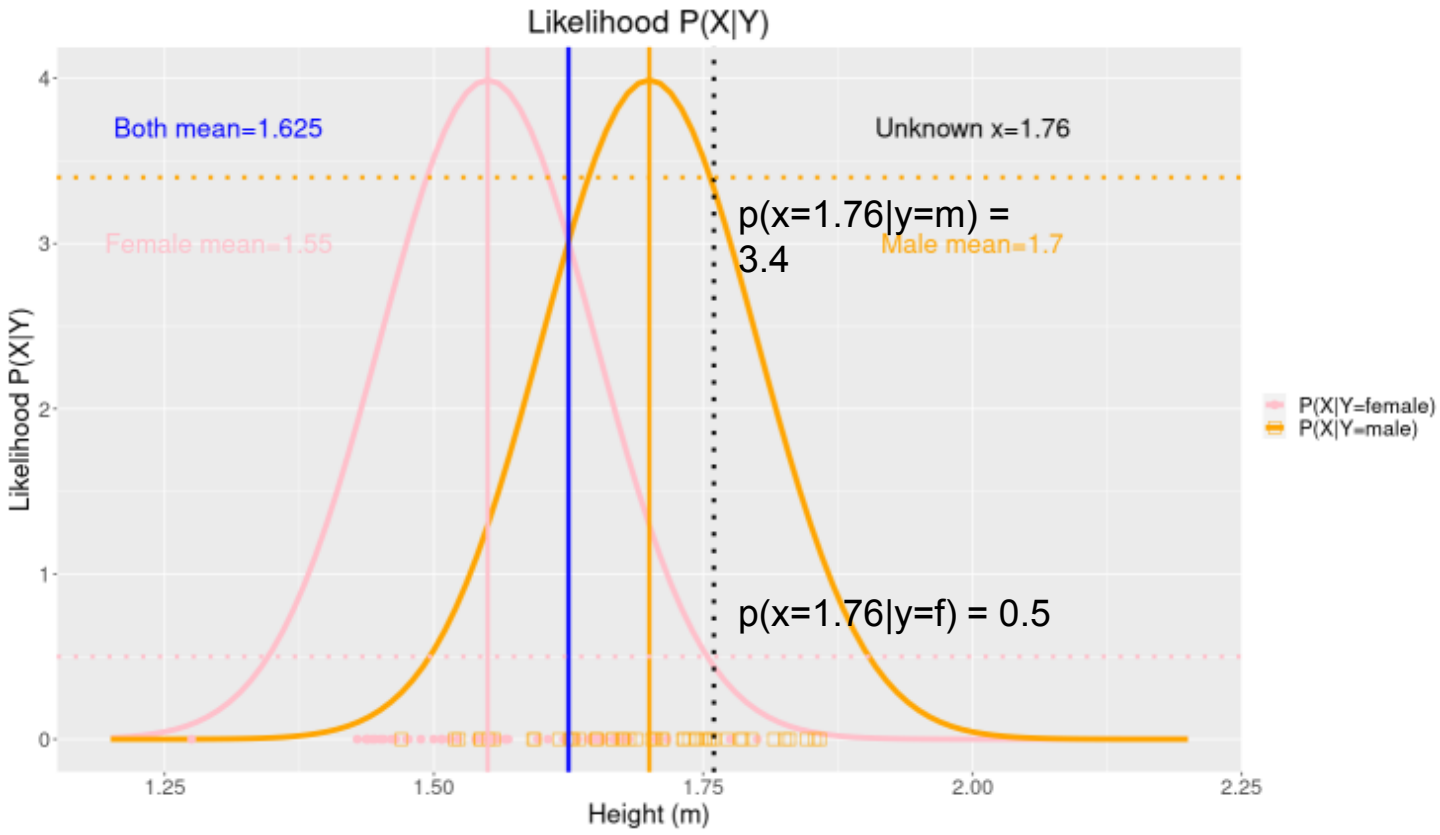
$$f_n(X) = \begin{cases} 1 & if\ P(x|y=1) > P(x|y=0) \\ 0 & otherwise \end{cases}$$

- Example: $P(x=1.76|y=m) = 3.4 > P(x=1.76|y=f)=0.5$ -> y=male
- Log-likelihood test

$$\log\frac{P(x|y=m)}{P(x|y=f)} = log\frac{3.4}{0.5} > 0$$ -> y = male

- BUT what if female are more likely than male in Nu Vuong country -> take into account the Prior propability P(Y=male) and P(Y=female)

$$\mu_1 = \mu_f = 155,\ \mu_2 = \mu_m = 170,\ \sigma_1 = \sigma_2 = \sigma = 01$$



Decision Boundary: 2) Using Maximum Likelihood

$$f_n(X) = \begin{cases} 1 & if\ P(x|y=1) > P(x|y=0) \\ 0 & otherwise \end{cases}$$

$$P(x|y=1) > P(x|y=0)$$

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma^2}} > \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu_2)^2}{\sigma^2}}$$

$$ln\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma^2}}\right) > ln\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu_2)^2}{\sigma^2}}\right)$$

$$ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)\left(-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma^2}\right) > ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)\left(-\frac{1}{2}\frac{(x-\mu_2)^2}{\sigma^2}\right)$$

$$(x-\mu_1)^2 > (x-\mu_2)^2$$

When $P(x|y=1) = P(x|y=0)$ *at* x_0:

$$(x_0-\mu_1)^2 = (x_0-\mu_2)^2$$

$$x_0^2 - 2x_0\mu_1 + \mu_1^2 = x_0^2 - 2x_0\mu_2 + \mu_2^2$$

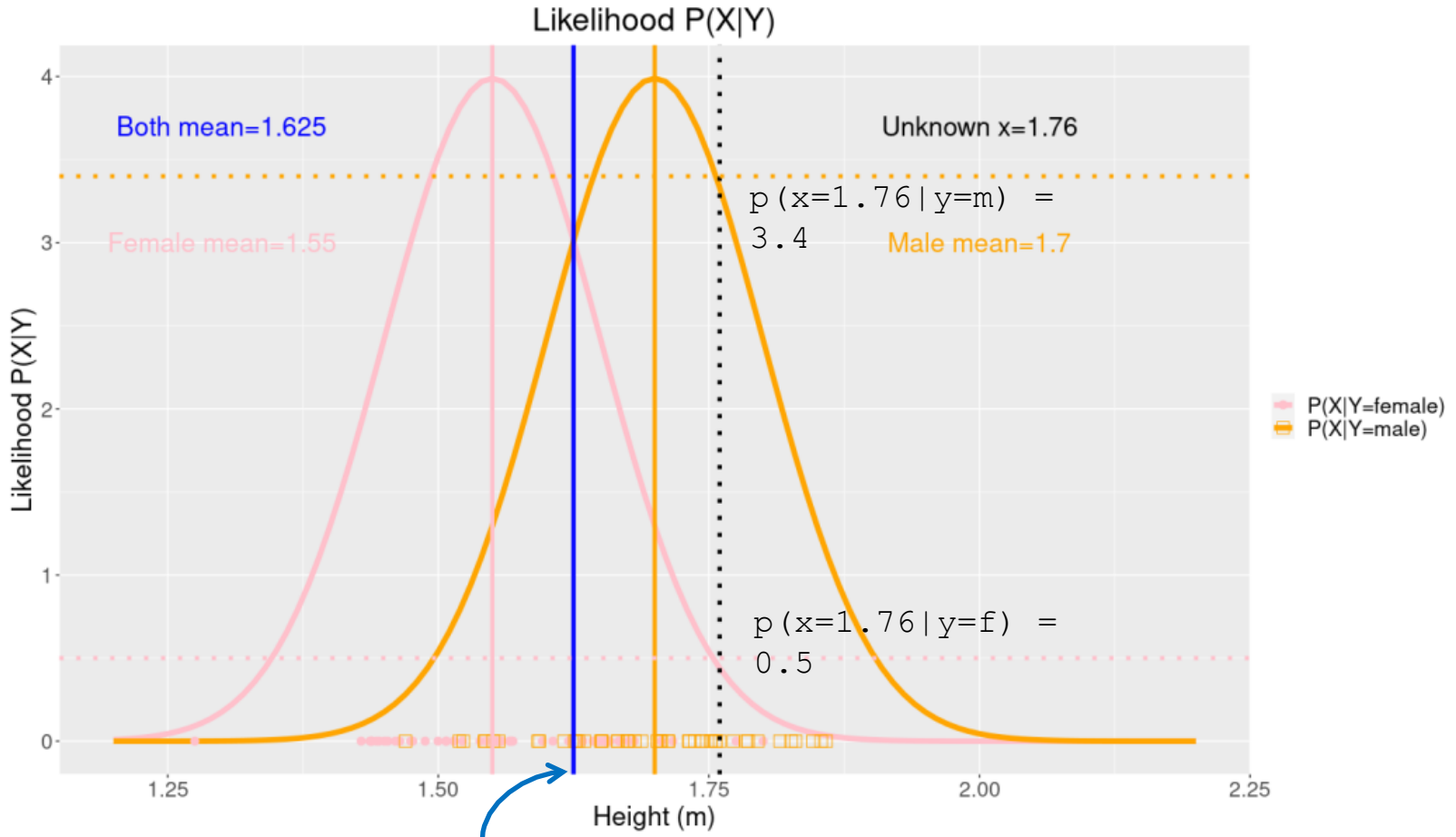
$$-2x_0\mu_1 + \mu_1^2 = -2x_0\mu_2 + \mu_2^2$$

$$2x_0(\mu_2-\mu_1) = \mu_2^2 - \mu_1^2$$

$$x_0 = \frac{(\mu_2+\mu_1)}{2}$$

$$x_0 = \frac{(1.55+1.7)}{2} = 1.625$$

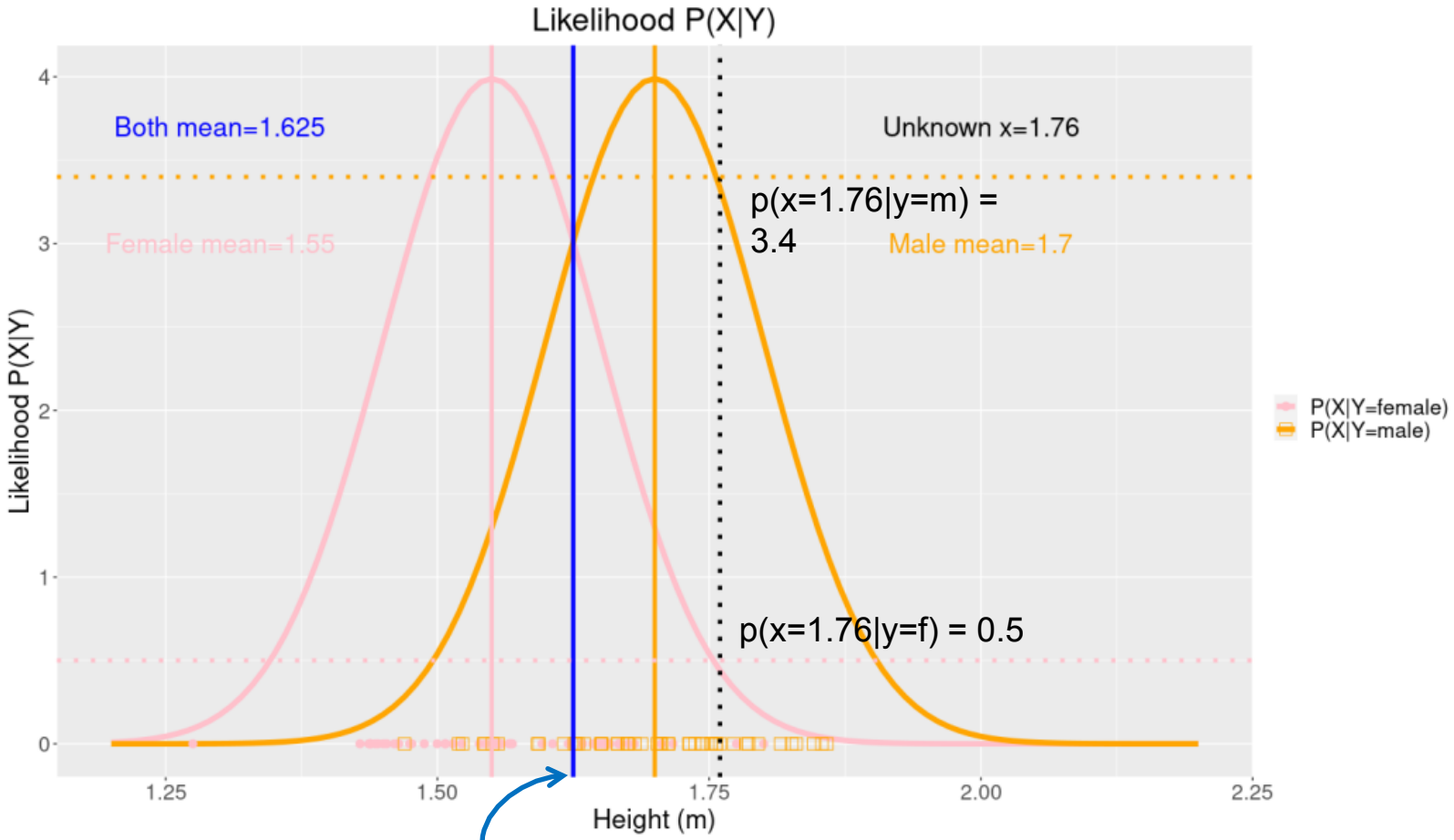
$$\mu_1 = \mu_f = 155,\ \mu_2 = \mu_m = 170,\ \sigma_1 = \sigma_2 = \sigma = 01$$



Decision boudary at $x_0 = \frac{(1.55+1.7)}{2} = 1.625$

Decision Boundary: 2) Using Maximum Likelihood

$$\mu_1 = \mu_f = 155,\ \mu_2 = \mu_m = 170,\ \sigma_1 = \sigma_2 = \sigma = 01$$



Decision boudary at $x_0 = \frac{(1.55+1.7)}{2} = 1.625$

How to decide male of female: 3) Bayesian a posteria criterion

- Using Posteria with Maximum A Posterior (MAP) of Bayes Rule

$$\hat{y} = \max_y P(y|x)$$

$\sigma_1 = \sigma_2 = \sigma = 0.4, \mu_1 = \mu_f = 155, \mu_2 = \mu_m = 170$

$$P(Y = female|X) = \frac{P(X|Y = female)P(Y = female)}{P(X)} = \frac{Likelihood * Prior}{Evidence} = Posterior$$

$P(x \mid \hat{y}) >= P(x \mid y)$

$\alpha(x)$: $P(y=male \mid x) > P(y=female \mid x)$ -> y = male

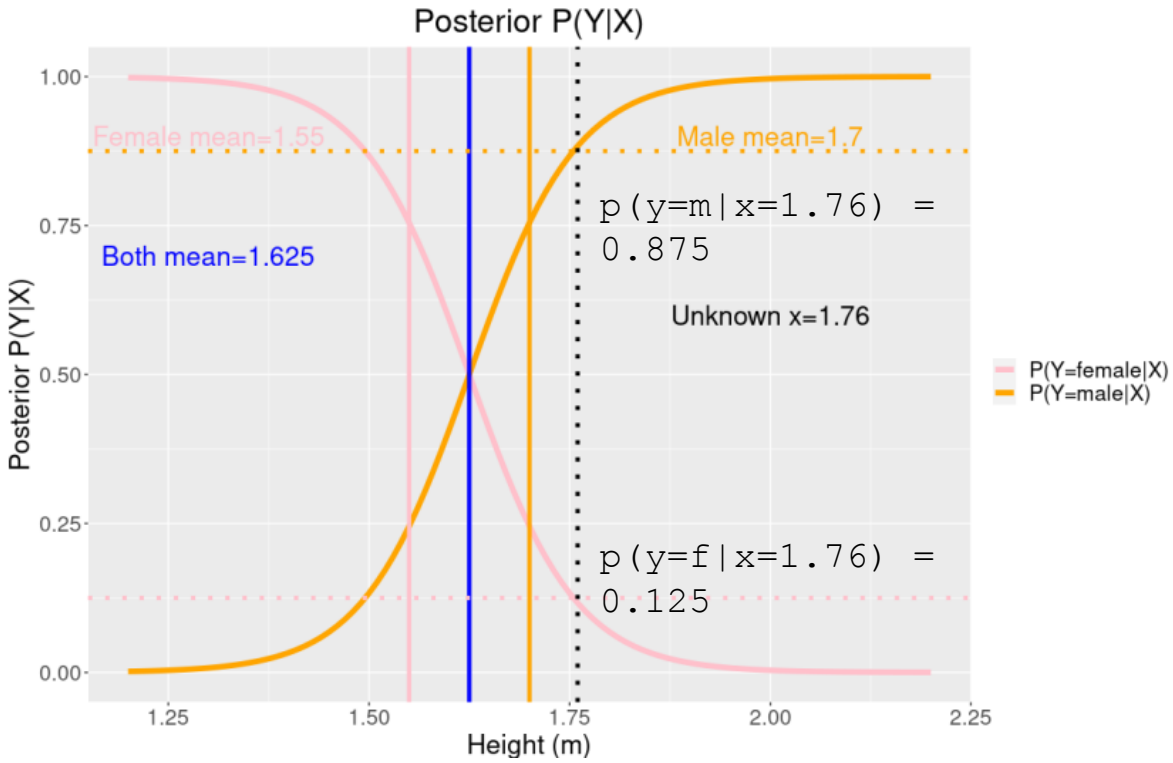
$P(y=male \mid x) < P(y=female \mid x)$ -> y = female

-> or $\alpha(x)$ can be write as a classifier $f_n(X)$:

$$f_n(X) = \begin{cases} f \\ m \end{cases} \quad \begin{matrix} if\ P(y = m|X = x) > P(y = f|X = x) \\ otherwise \end{matrix}$$

$$f_n(X) = \begin{cases} 1 \\ 0 \end{cases} \quad \begin{matrix} if\ P(y = 1|X = x) > P(y = 0|\ X = x) \\ otherwise \end{matrix}$$

- Example: $P(y=m \mid x=1.76) = 0.875 > P(y=f \mid x=1.76) = 0.125$ -> y=male

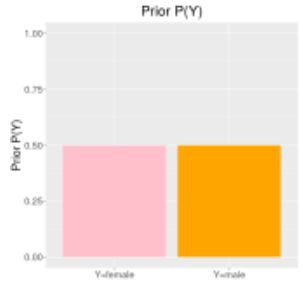


Decision Boundary: 3) Bayesian a posteria criterion

$$f_n(X) = \begin{cases} 1 & if\ P(y=1|x) > P(y=0|x) \\ 0 & otherwise \end{cases}$$

$$P(y=1|x) > P(y=0|x)$$

$$P(x|y=1)P(y=1) > P(x|y=0)P(y=0)$$



1) $P(y=1) = P(y=0) = 0.5$, *then it is the same as Maximum Likelihood case*

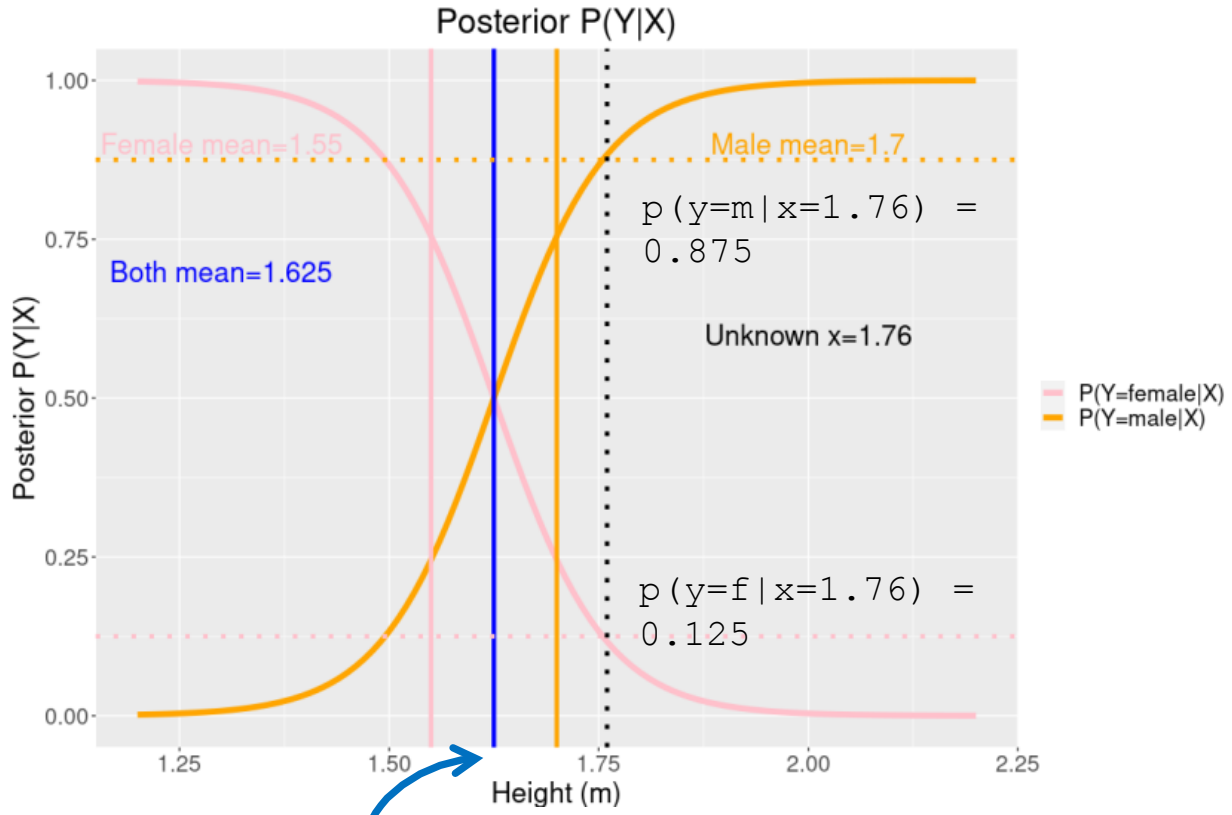
$$P(x|y=1) > P(x|y=0)$$

Therefore, the decision boundary at

$$x_0 = \frac{(\mu_2 + \mu_1)}{2}$$

$$x_0 = \frac{(1.55 + 1.7)}{2} = 1.625$$

$$\mu_1 = \mu_f = 155,\ \mu_2 = \mu_m = 170,\ \sigma_1 = \sigma_2 = \sigma = 01$$



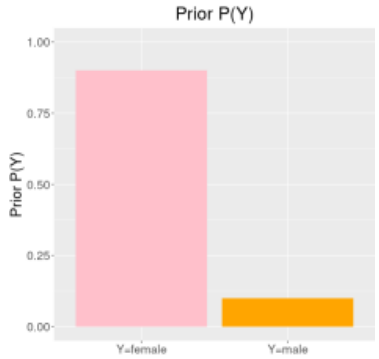
$$x_0 = \frac{(1.55 + 1.7)}{2} = 1.625$$

Decision Boundary: 3) Bayesian a posteria criterion

$$f_n(X) = \begin{cases} 1 & if\ P(y=1|x) > P(y=0|x) \\ 0 & otherwise \end{cases}$$

$$P(y=1|x) > P(y=0|x)$$

$$P(x|y=1)P(y=1) > P(x|y=0)P(y=0)$$



2) $P(y=1) = p$ and $P(y=0) = 1-p$

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma^2}} p > \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu_2)^2}{\sigma^2}} (1-p)$$

$$-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma^2} + lnp > -\frac{1}{2}\frac{(x-\mu_2)^2}{\sigma^2} + ln(1-p)$$

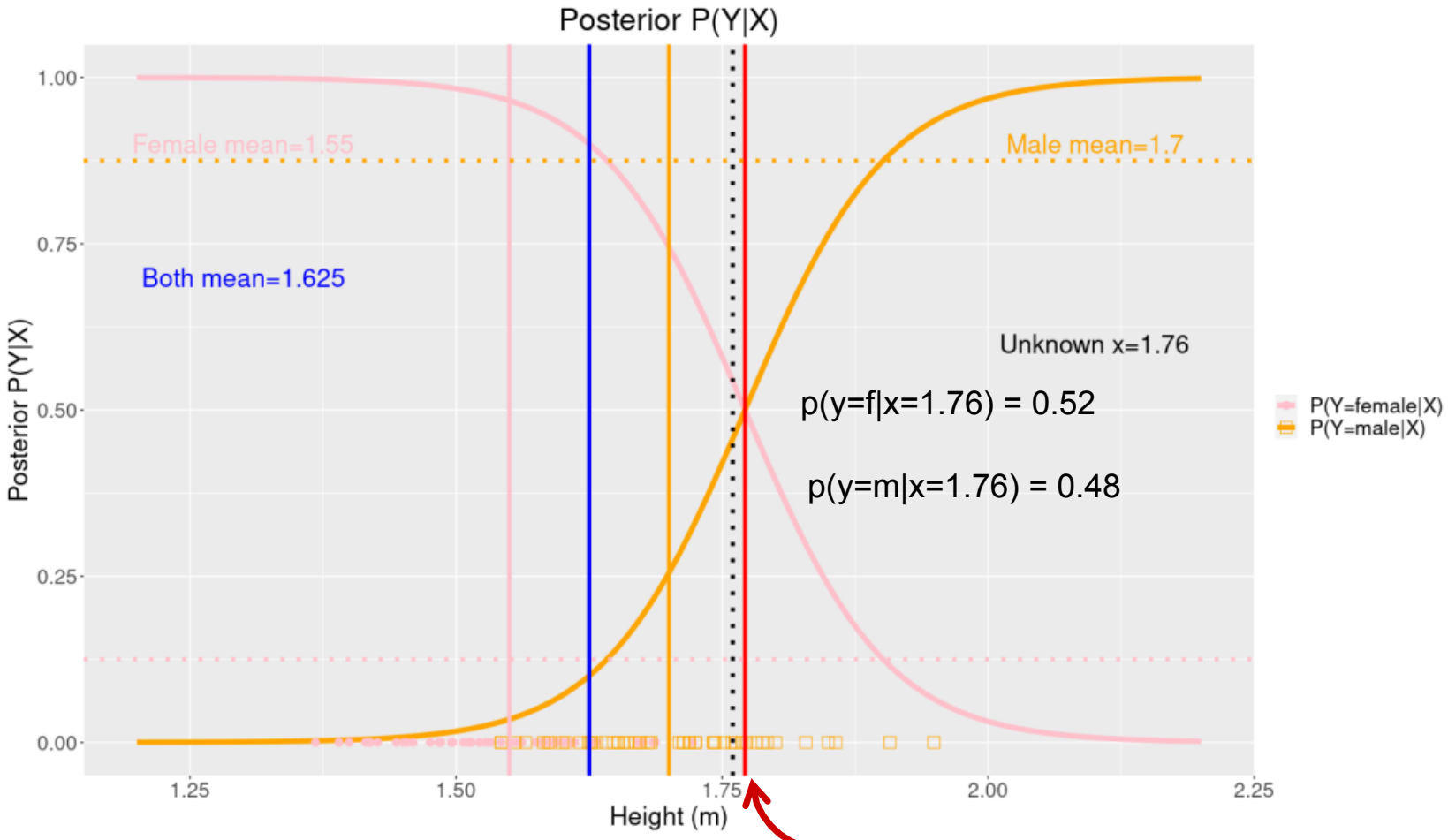
When $P(x|y=1) = P(x|y=0)$ *at* x_0:

$$x_0 = \frac{\mu_1+\mu_2}{2} + \frac{\sigma^2}{\mu_1-\mu_2} ln\left(\frac{1-p}{p}\right)$$

```
x0 <- ((1.55+1.7)/2) + (0.1^2/(1.55-1.7))*log(0.1/0.9)
```

```
[1] 1.771482
```

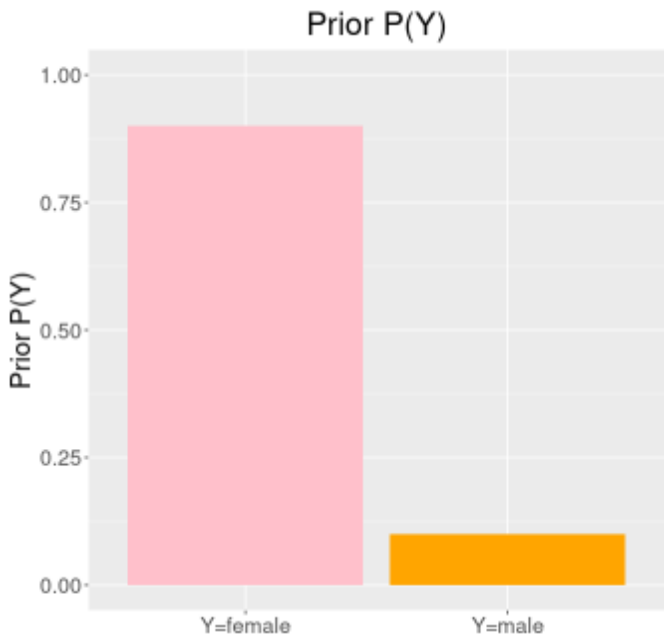
$$\mu_1 = \mu_f = 155,\ \mu_2 = \mu_m = 170,\ \sigma_1 = \sigma_2 = \sigma = 01$$

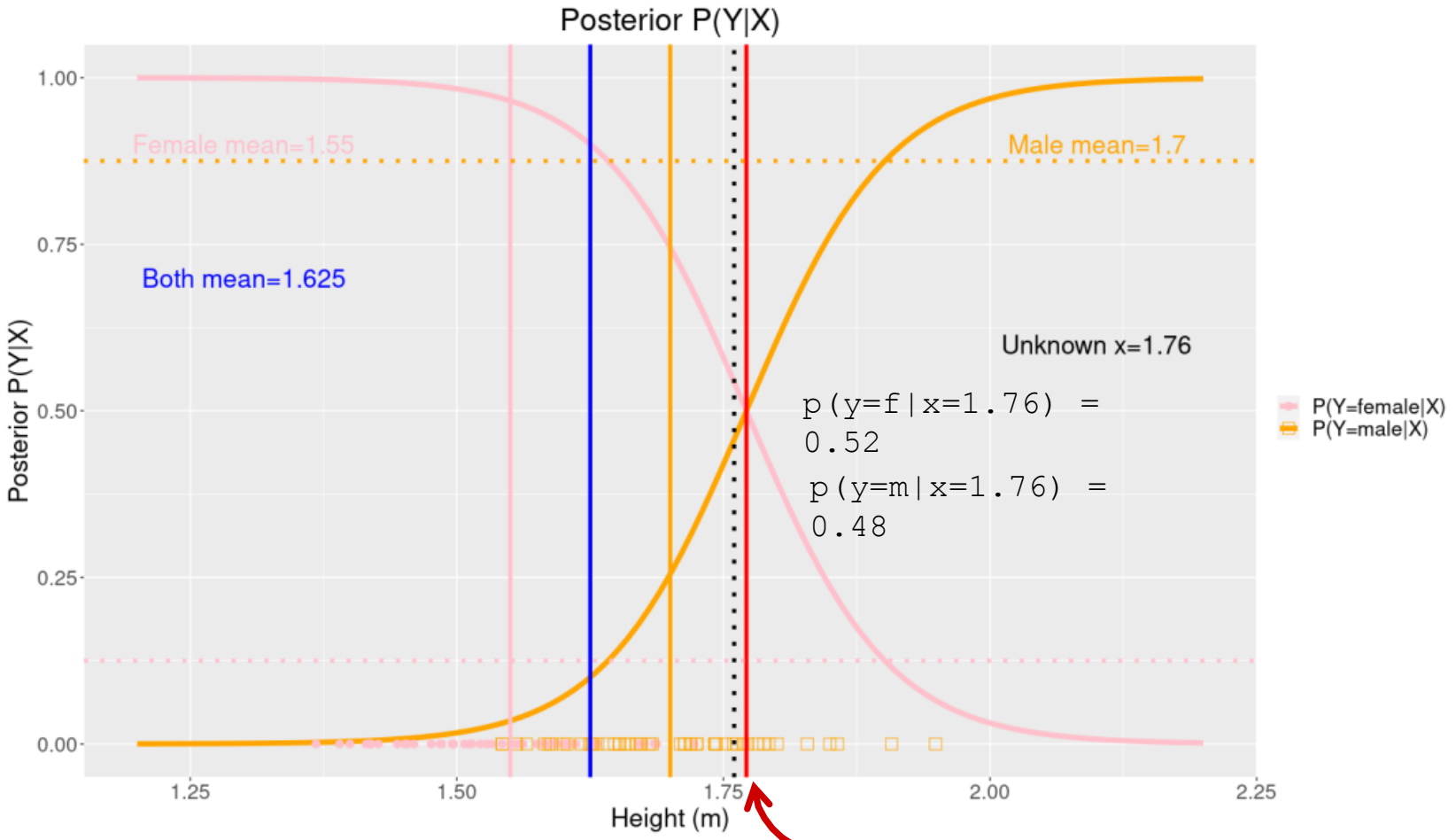


Decision boudary at $x_0 = 1.77$

Decision Boundary: 3) Bayesian a posteria criterion

$$\mu_1 = \mu_f = 155,\ \mu_2 = \mu_m = 170,\ \sigma_1 = \sigma_2 = \sigma = 01$$





Decision boudary at $x_0 = 1.77$

Loss/Cost function

❖What does it cost if you make a mistake?

- i.e. suppose you decide y = male, but really y = female
- i.e. you may pay a big penalty if you decide it is a female when it is a male

❖ Loss function L(x, y, $\hat{y}$): how much loss you incur by classifying the label of x as f(x) = $\hat{y}$ if the true label is y

- All errors penalisied the same: $L(f(x), y) = L(\hat{y}, y) = 1_{f(x)\neq y} = 1_{\hat{y}\neq y}$= I(Y ≠ f(X))
  - l(f(x), y) = 0, if f(x) = y
  - l(f(x), y) = 1, if f(x) ≠ y
- All errors penalisied NOT the same
  - l(f(x)=0, y=0) = l(f(x)=1, y=1) = 0
  - l(f(x)=1, y=0) = 10
  - l(f(x)=0, y=1) = 1000000

Risk

- The risk R is the expected loss

$$R(\hat{y} \mid X = x) = E(L(f(X), Y)) = E[1_{f(x) \neq y}] = E[I(Y \neq f(X))]$$

- The expected conditional risk at point x:

$$R(\hat{y} \mid X = x) = E(L(f(X), Y)) = E[E[L(f(X), Y) \mid X]]$$

$$= \sum_Y l(f(x), Y) P(Y \mid X = x)$$

$$= l(Y = m, f(x)) P(Y = m \mid X = x) + l(Y = f, f(x)) P(Y = f \mid X = x)$$

- Use Bayes decision rule: select the label $f_n(X)$ for which the conditional risk is minimal

Summary: Bayes Decision Theory

Bayes risk is the best you can do if

(a) You know P(x | y)P(y) and Loss function

(b) You can compute $\hat{\alpha} = Arg\min_{\alpha \in A} R(\alpha)$

(c) You can afford the losses (e.g. gambling, poker)

(d) you can make decision for a sequence of data x_1, x_2 ... x_n with states y_1, y_2 ... $y_n$where each point (x_i, y_i) are independently identically distributed from P(x,y)

Bayesian Decision Theory with actions, example

Example 7.6 (Milleritis)**.** We will apply Bayesian decision theory to the problem of diagnosing and treating a fictional disease which we will call milleritis. Suppose the level of uric acid (X) in the blood is affected by this disease as follows:

- Patients without milleritis: $X \sim \mathrm{N}(\mu = 6, \sigma = 1)$ mg/100ml blood
- Patients with milleritis: $X \sim \mathrm{N}(\mu = 10, \sigma = 1.5)$ mg/100ml blood

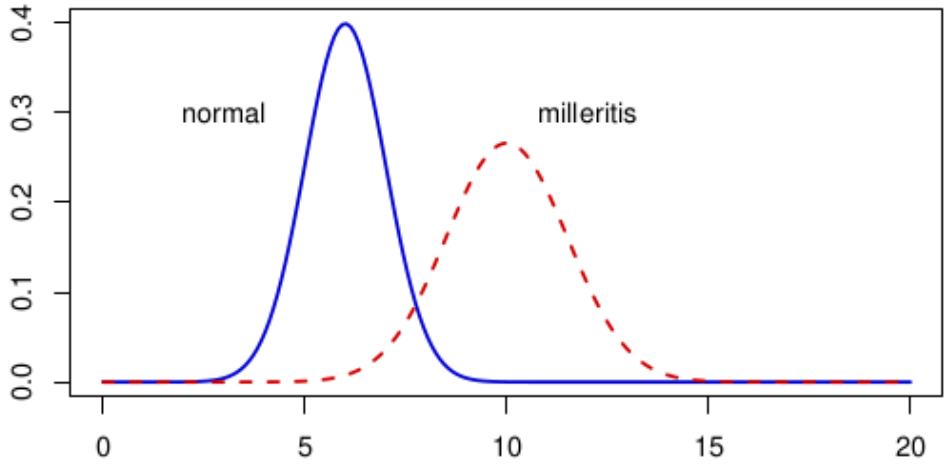


Figure 7.2: Densities of "normal" *vs.* milleritis.

Bayesian Decision Theory with actions, example

For a measured value of X, we can use the following version of Bayes theorem to find the **conditional probability** of a patient being milleritis positive (M^+):

$$\mathrm{P}(M^+|X=x) = \frac{f(x|M^+) \times \mathrm{P}(M^+)}{f(x|M^+) \times \mathrm{P}(M^+) + f(x|M^-) \times \mathrm{P}(M^-)}$$

where

$$f(x|M^+) = \frac{1}{\sqrt{4.5\pi}} e^{-\frac{(x-10)^2}{4.5}} \quad \text{and} \quad f(x|M^-) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-6)^2}{2}}.$$

Suppose the prevalence of milleritis is 20%. Therefore we have $\mathrm{P}(M^+) = 0.2$ and $\mathrm{P}(M^-) = 0.8$. The conditional probability of milleritis $\mathrm{P}(M^+|X=x)$ is shown in Figure 7.3.

Suppose when a patient comes to a doctor, their uric acid level is measured (this costs \$5) and (depending on the result) one of the following three courses of action is taken:

- a_1: Do not treat. This incurs no additional cost for a M^- patient. But if the patient is M^+, the condition will worsen and the cost of treating an acute case is \$200.

Bayesian Decision Theory with actions, example

For a measured value of X, we can use the following version of Bayes theorem to find the **conditional probability** of a patient being milleritis positive (M^+):

$$\mathrm{P}(M^+|X=x) = \frac{f(x|M^+) \times \mathrm{P}(M^+)}{f(x|M^+) \times \mathrm{P}(M^+) + f(x|M^-) \times \mathrm{P}(M^-)}$$

where

$$f(x|M^+) = \frac{1}{\sqrt{4.5\pi}} e^{-\frac{(x-10)^2}{4.5}} \quad \text{and} \quad f(x|M^-) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-6)^2}{2}}.$$

Suppose the prevalence of milleritis is 20%. Therefore we have $\mathrm{P}(M^+) = 0.2$ and $\mathrm{P}(M^-) = 0.8$. The conditional probability of milleritis $\mathrm{P}(M^+|X=x)$ is shown in Figure 7.3.

Suppose when a patient comes to a doctor, their uric acid level is measured (this costs \$5) and (depending on the result) one of the following three courses of action is taken:

- a_1: Do not treat. This incurs no additional cost for a M^- patient. But if the patient is M^+, the condition will worsen and the cost of treating an acute case is \$200.

Bayesian Decision Theory with actions, example

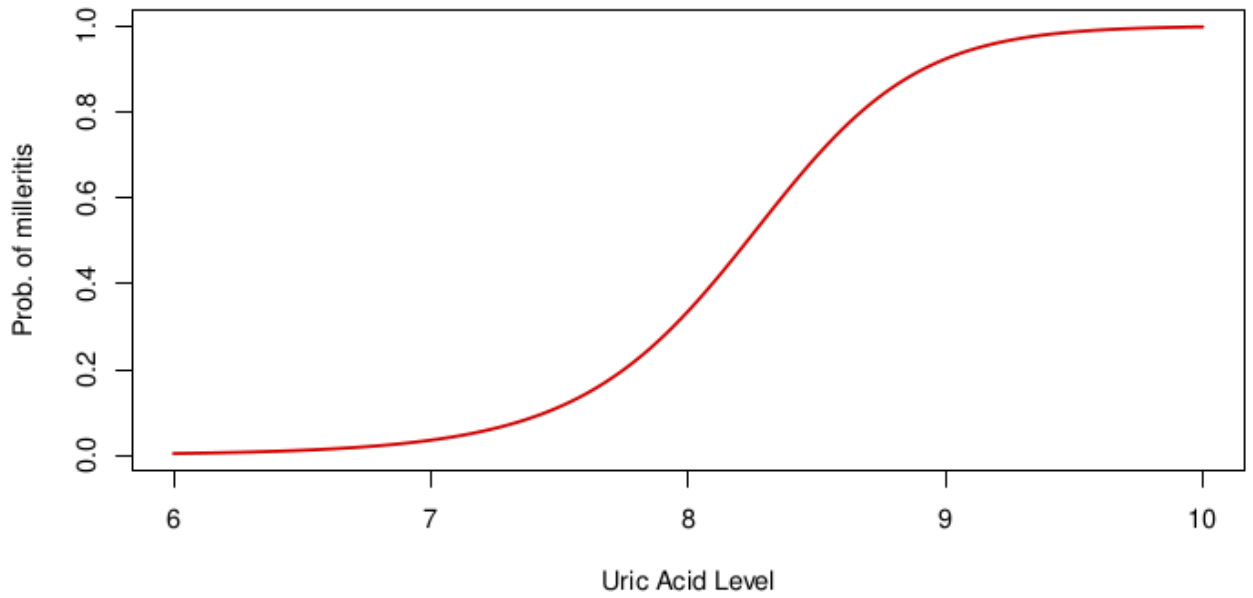


Figure 7.3: Conditional probability of milleritis $\mathrm{P}(M^+|X = x)$.

a_2: Treat for milleritis. The cost of treatment is \$55.

a_3: Do a second test which costs an additional \$25 but gives a definitive result. Only treat if the patient is M^+.

Bayesian Decision Theory with actions, example

If we define the **loss** as the cost of treatment, then we have the following table of losses:

	M^-	M^+
a_1	\$5	\$205
a_2	\$60	\$60
a_3	\$30	\$85

To get the **expected loss** (as a function of x) for each action, we simply apply $\mathrm{P}(M^+|X = x)$ to these losses. As an example for a_1:

$$\mathrm{E}(\text{loss}|X = x) = 5 \times \mathrm{P}(M^-|X = x) + 205 \times \mathrm{P}(M^+|X = x),$$

where $\mathrm{P}(M^-|X = x) = 1 - \mathrm{P}(M^+|X = x)$.

Bayesian Decision Theory with actions, example

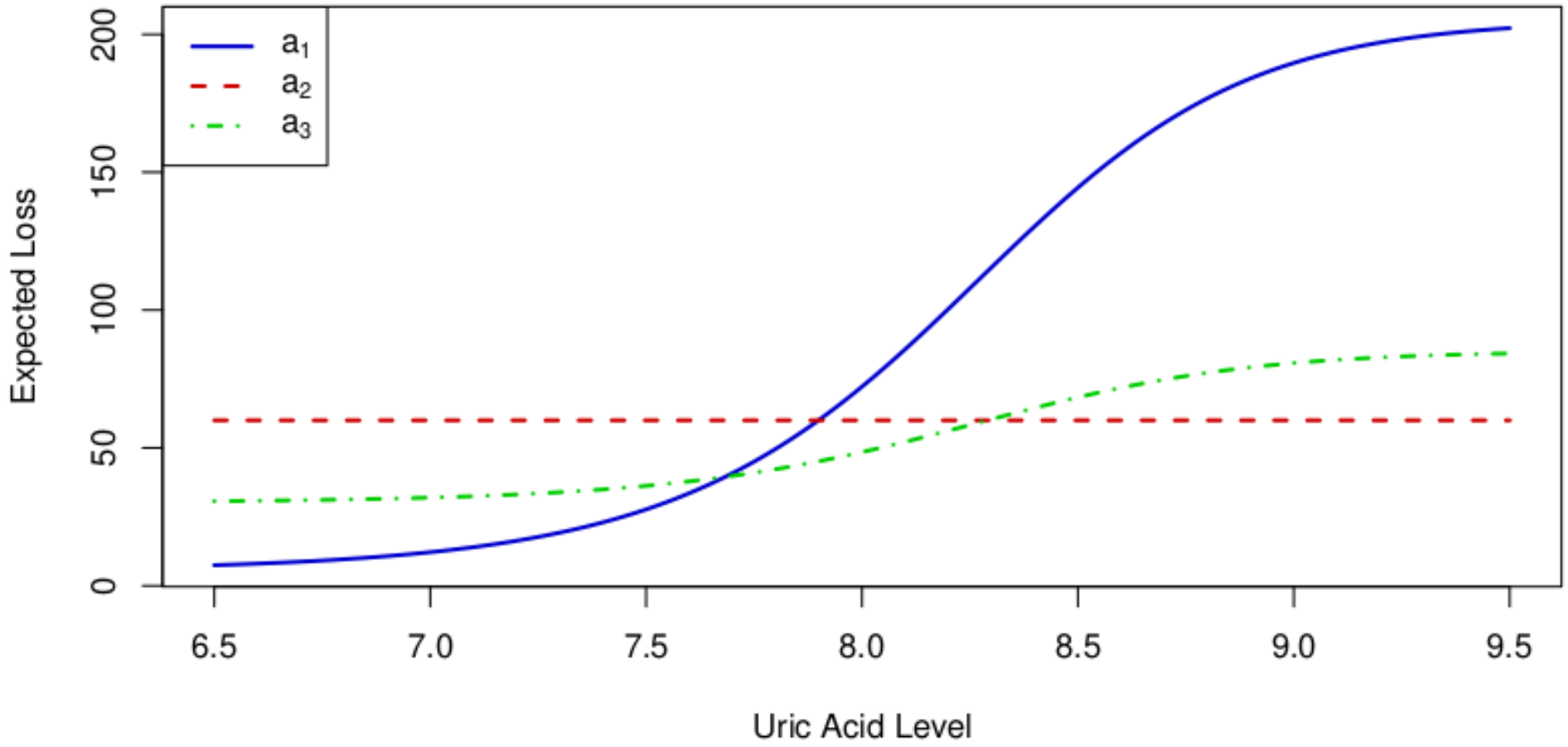


Figure 7.4: Expected losses for all three actions

Bayesian Decision Theory with actions, example

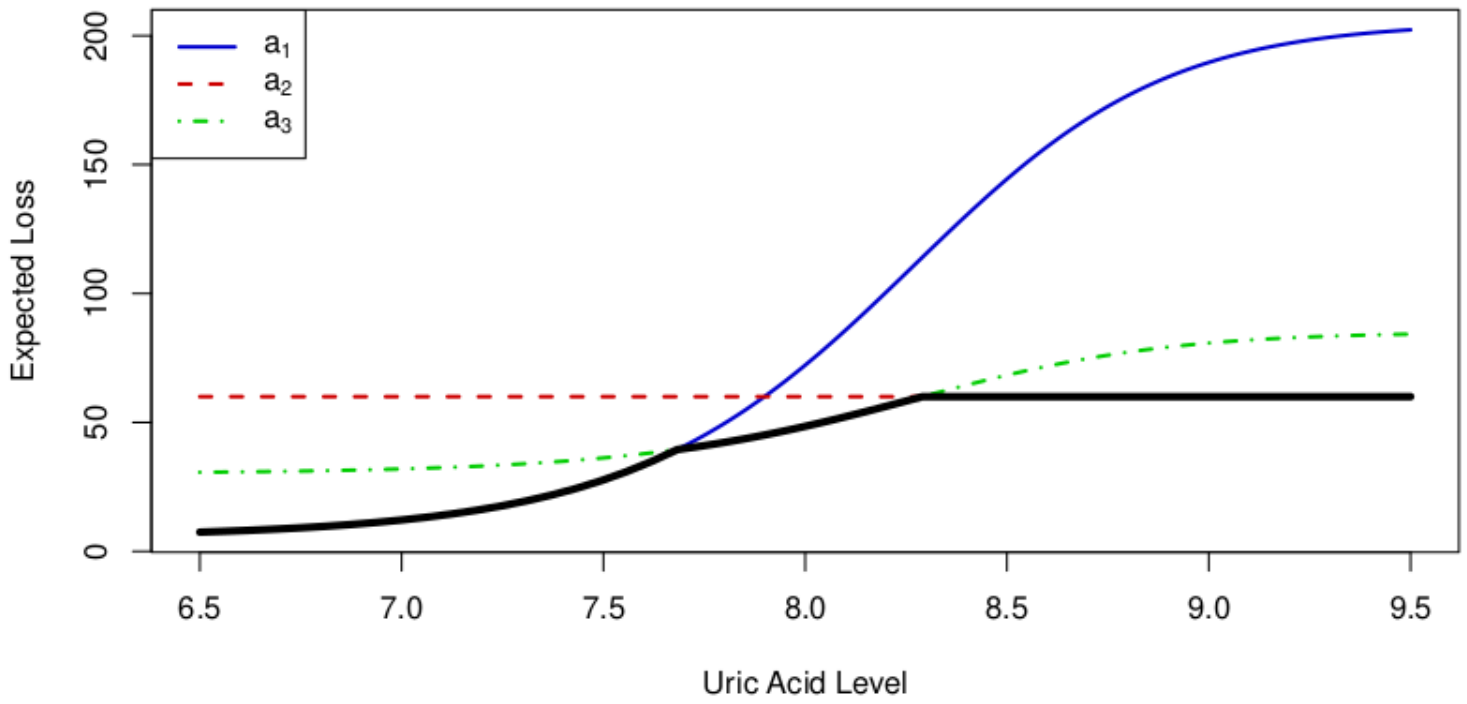


Figure 7.5: Expected losses for all three actions

Bayes rule is given by

$$a_1 \text{ for } x < 7.68, \qquad a_3 \text{ for } 7.68 < x < 8.29, \qquad \text{and } a_2 \text{ for } x > 8.29. \qquad \Diamond$$

Bayesian Decision Theory, exercise

- Create a shinyApp for gender prediction example with adding plot of expected loss 1/0

Bayesian Decision Theory, homework

- Create a shinyApp for Milleritis example with the slider allowing to change a_1, a_2 and a_3

Probability as an Expectation: indicator function

Let A be any event. We can write $\mathbb{P}(A)$ as an expectation, as follows.
Define the ***indicator function:***

$$I_A = \begin{cases} 1 & \textit{if event } A \textit{ occurs,} \\ 0 & \textit{otherwise.} \end{cases}$$

$$\mathrm{I}(\mathrm{f}(\mathrm{X}) \neq \mathrm{Y}) = 1_{f(X) \neq Y} = \begin{cases} 1 \; if \; f(X) \neq Y \\ 0 \; if \; f(X) = Y \end{cases}$$

Then I_A is a ***random variable***, and

$$\begin{aligned} \mathbb{E}(I_A) &= \sum_{r=0}^{1} r\mathbb{P}(I_A = r) \\ &= 0 \times \mathbb{P}(I_A = 0) + 1 \times \mathbb{P}(I_A = 1) \\ &= \mathbb{P}(I_A = 1) \\ &= \mathbb{P}(A). \end{aligned}$$

Thus

$$\boxed{\mathbb{P}(A) = \mathbb{E}(I_A) \textit{ for any event } A.}$$

Bayesian Classifier: Optimal prediction functions in closed form

- The optimal classifier f^* with 0-1 loss is:

$f^* = \text{argmin } R(f) = \text{argmin } E(L(f(X), Y))$

$= \text{argmin } E[E[L(f(X), Y)|X]]$

$= \text{argmin } P(y = f|x)1_{f(x)=m} + P(y = m|x)1_{f(x)=f}$

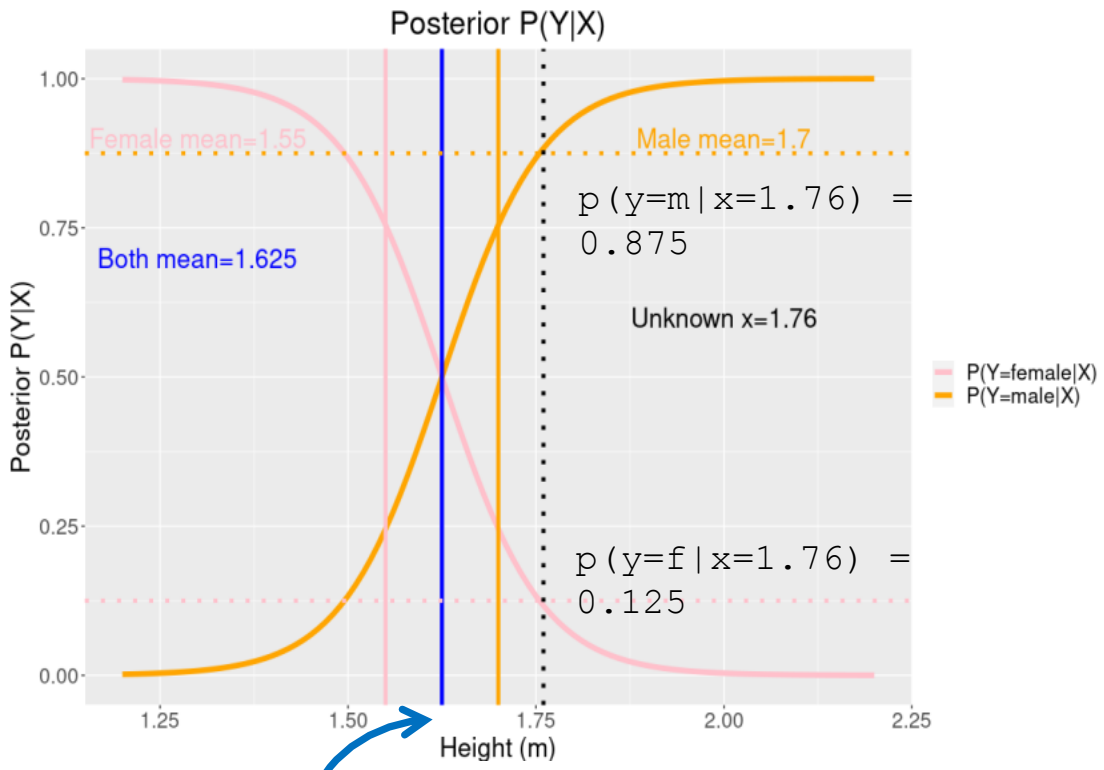
$= \text{argmin } P(y = f|x)1_{f(x)=m} + P(y = m|x)(1-1_{f(x)=m})$

$= \text{argmin } P(y = m|x) + [P(y = f|x) - P(y = m|x)]1_{f(x)=m}$

$= \begin{cases} m & \text{if } P(y = m|x) > P(y = f|x) \\ f & Otherwise \end{cases}$

- which is exactly $f^*(x) = \text{argmax } P(Y=y|x)$

$\mu_1 = \mu_f = 155,\ \mu_2 = \mu_m = 170,\ \sigma_1 = \sigma_2 = \sigma = 01$



$$x_0 = \frac{(1.55 + 1.7)}{2} = 1.625$$

Bayesian Classifier

- The optimal classifier f^* with 0-1 loss is:

$f^* = \text{argmin } P(y = f \mid x)1_{f(x)=m} + P(y = m \mid x)1_{f(x)=f}$ (M)

for example x=1.76 then $p(y=m \mid x=1.76) = 0.875$

and $p(y=f \mid x=1.76) = 0.125$. So with pointwise, (M) can be wrote

$f^* = \text{argmin } P(y=f \mid x=1.76)1_{f(x)=m} + P(y=m \mid x=1.76)1_{f(x)=f}$

$= \text{argmin } 0.125 * 1_{f(x)=m} + 0.875 * 1_{f(x)=f}$

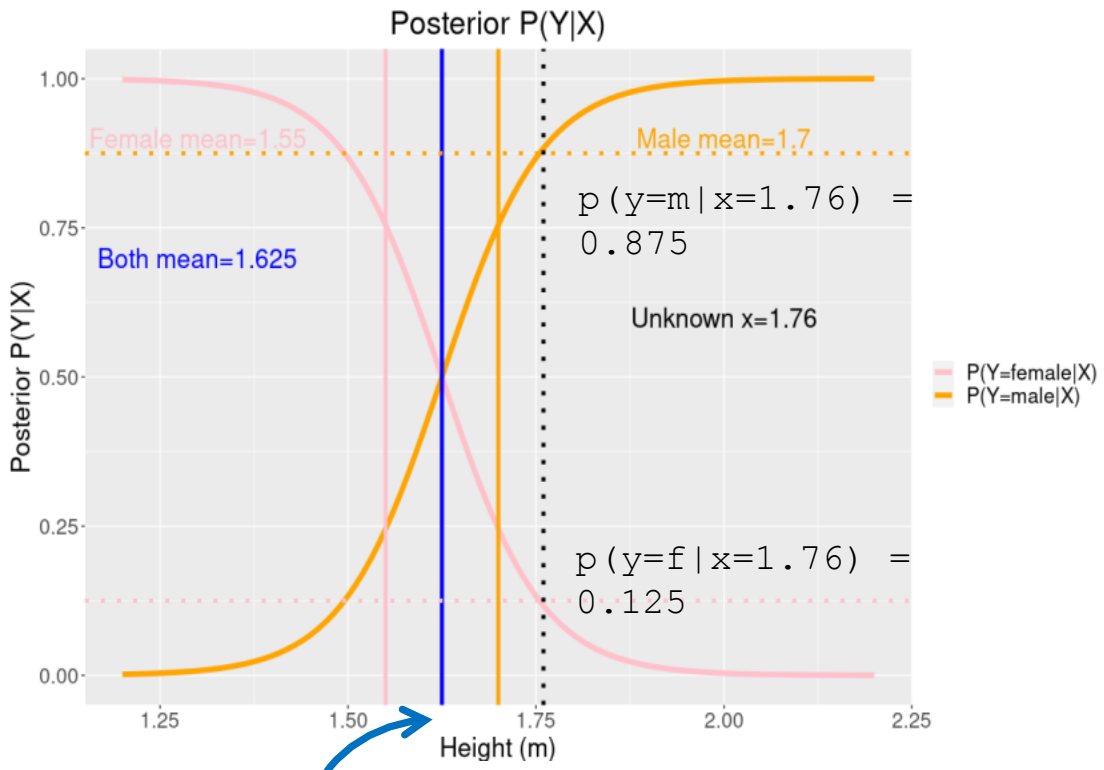
$= m$ and $R_{min} = 0.125$

$$\text{so } f^* = \begin{cases} m & \text{if } P(y = m|x) > P(y = f|x) \\ f & Otherwise \end{cases}$$

$$= \begin{cases} 0 & \text{if } P(y = 0|x) > P(y = 1|x) \\ 1 & Otherwise \end{cases}$$

which is exactly $f^*(x) = \text{argmax } P(Y=y \mid x)$

$\mu_1 = \mu_f = 155,\ \mu_2 = \mu_m = 170,\ \sigma_1 = \sigma_2 = \sigma = 01$



$$x_0 = \frac{(1.55 + 1.7)}{2} = 1.625$$

Toy example 2: Predict gender of a person based on height and hair length

- Setting: binary classification, that is Y = {male, female} = {M, F} = {0, 1} and vector X (X1 is height of a person and X2 is hair length of a person) and its mean (μ), covariance ($\sum$)

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_{X_1} \\ \mu_{X_2} \end{pmatrix} \qquad \Sigma = \begin{pmatrix} V[X_1] & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_1, X_2) & V[X_2] \end{pmatrix}$$

- The joint density p(x1, x2, y) of the probability measure P on X × Y can be decomposed as follows
  - The **class-conditional density** or **likelihood** p(x1, x2 | y). It models the occurrence of the features x of class y.
  - The **conditional probability** p(y | x1, x2) or **Posterior**. The probability that we observe y given that the input is x. The most probable class y for the features x is then used for prediction.
  - The **marginal distribution or evidence** p(x1) and p(x2). It models the cumulated occurrence of features x1, x2 over all classes.
  - The class probabilities p(y). The total probability of a class y or **Prior**.

https://www.byclb.com/TR/Tutorials/neural_networks/ch4_1.htm

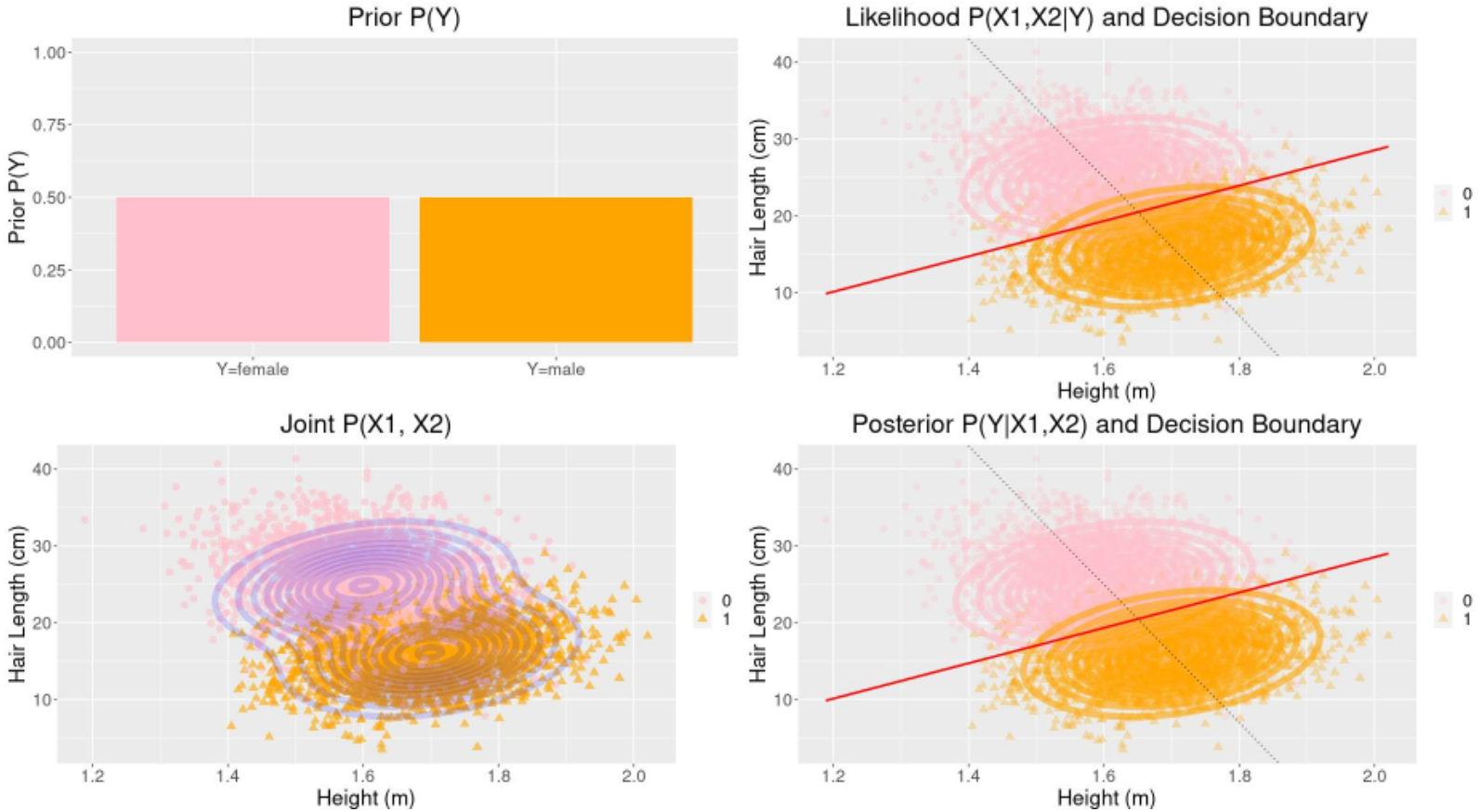
Toy example 2: Predict gender of a person based on height and hair length

P(Y=male) == P(Y=female) = p = 0.5

$$\mu_{X1|Y=female} = 1.60;\ \mu_{X1|Y=male} = 1.70$$
$$\mu_{X2|Y=female} = 25;\ \mu_{X1|Y=male} = 16$$

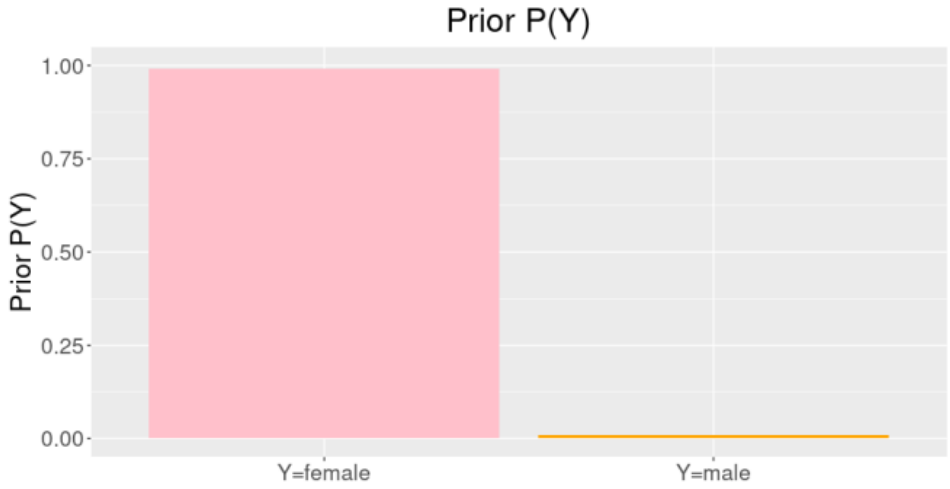
$$\Sigma = \begin{pmatrix} V[X_1] & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_1, X_2) & V[X_2] \end{pmatrix}$$

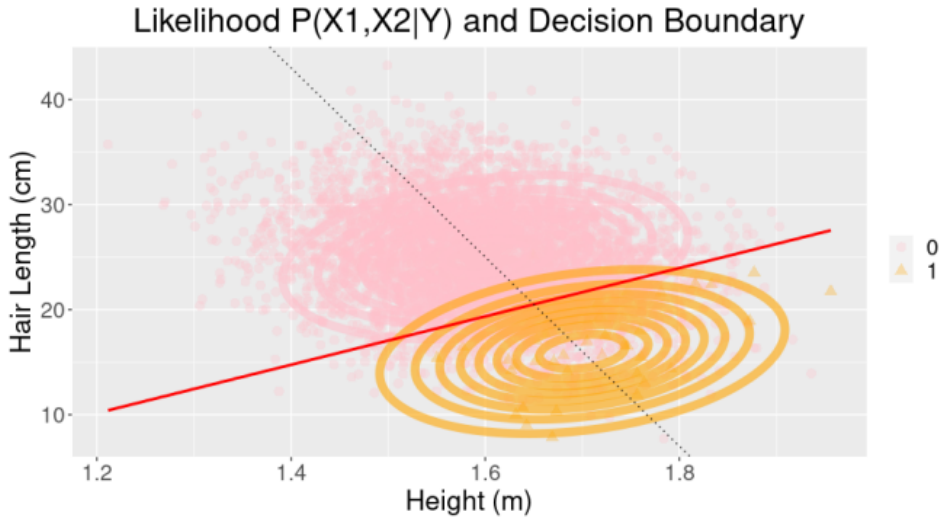
$$\sum_{X1|Y\,=\,female} = \begin{bmatrix} 0.01 & -0.1 \\ -0.1 & 25 \end{bmatrix} \qquad \sum_{X2|Y\,=\,male} = \begin{bmatrix} 0.01 & 0.1 \\ 0.1 & 14 \end{bmatrix}$$

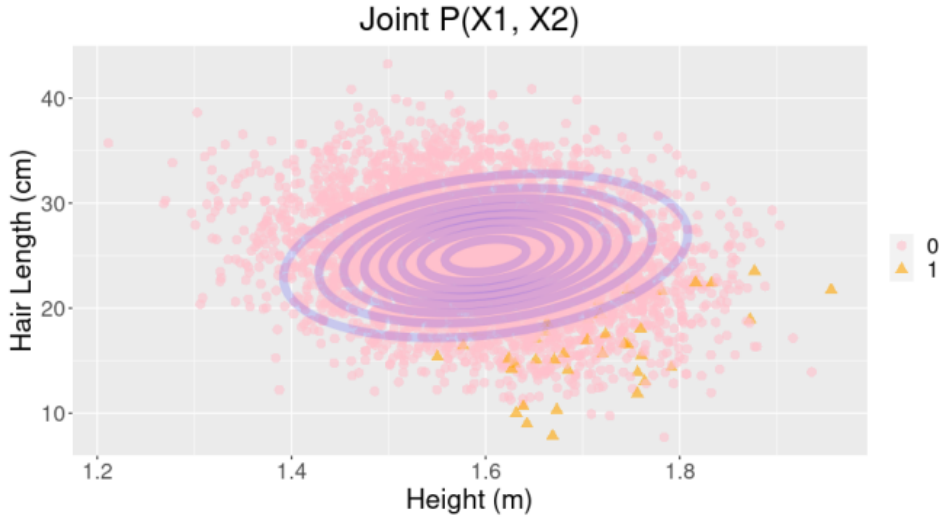


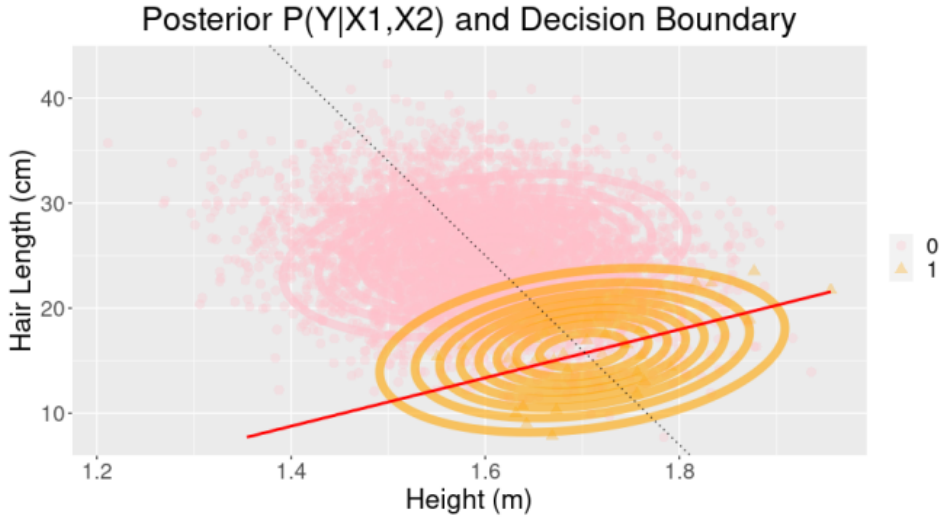
Toy example 2: Predict gender of a person based on height and hair length

P(Y=female) = 0.99
P(Y=male) = 0.01



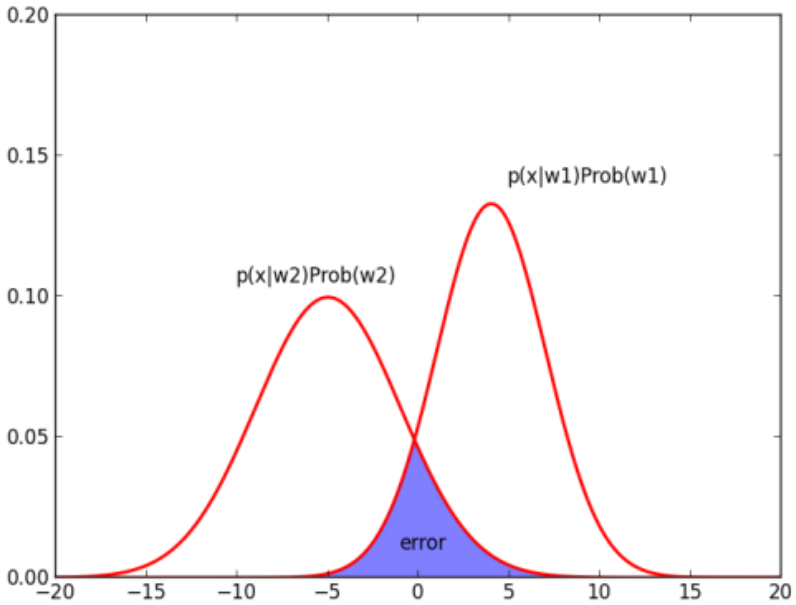






Exercises

1. Let Y={w1, w2} and a continuous random variable X. Please explain why the violet region is error?



2. How to caculate the error?

https://www.projectrhea.org/rhea/index.php/Bayes_Rule_for_1-dimensional_and_N-dimensional_feature_spaces

Regression function (context of classification)

Consider (X, Y) drawn according to a probability distribution P on the product space $\mathcal{X} \times \{0, 1\}$. We want to describe the distribution P in terms of two other quantities:

- Let μ be the marginal distribution of X, that is $\mu(A) = P(X \in A)$.

- Define the so-called regression (!)-function:

$$\eta(x) := E(Y \mid X = x)$$

- In the special case of classification, the regression function can be rewritten as

$$\begin{aligned} \eta(x) &= 0 \cdot P(Y = 0 \mid X = x) + 1 \cdot P(Y = 1 \mid X = x) \\ &= P(Y = 1 \mid X = x) \end{aligned}$$

Regression function (context of classification) (2)

Intuition:

- If $\eta(x)$ is close to 0 or close to 1, then classifying x is easy.
- If $\eta(x)$ is close to 0.5, then classifying x is difficult.

WHY?

Regression function (context of classification) (3)

Proposition 1 (Unique decomposition)

The probability distribution P is uniquely determined by μ and η.

Intuition (discrete case): We can rewrite

$$\begin{aligned} P(X = x, Y = 1) &= P(Y = 1 | X = x) P(X = x) \\ &= \eta(x)\mu(x) \end{aligned}$$

and similarly

$$\begin{aligned} P(X = x, Y = 0) &= P(Y = 0 | X = x) P(X = x) \\ &= (1 - \eta(x))\mu(x) \end{aligned}$$

So we can express the probability of any event (x, y) in terms of η and μ.

Explicit form of the Bayes classifier

Consider the 0-1-loss function. Recall:

- the risk of a classifier under the 0-1-loss counts "how often" the classifier fails, that is

$$R(f) = E(\ell(X, Y, f(X))) = E(\mathbb{1}_{f(X)\neq Y}) = P(f(X) \neq Y).$$

- The Bayes classifier f^* was defined as the classifier that minimizes the true risk. This is an implicit definition, we don't yet have a formula for it.

Now consider the following classifier:

$$f^\circ(x) := \begin{cases} 1 & \text{if } \eta(x) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

Explicit form of the Bayes classifier (2)

Theorem 2 (f° is the Bayes classifier)

Consider classification with 0-1-loss. Let $f : \mathcal{X} \to \{0, 1\}$ be any (measurable) classifier and f° the classifier defined above. Then $R(f) \geq R(f^o)$.

Before we prove it, digest what this means:

- The theorem shows that $f^\circ = f^*$ (WHY?)
- Consequence: in the particular case of classification with the 0-1-loss, we have an explicit formula for the Bayes classifier.
- In practice, this doesn't help, WHY?

Explicit form of the Bayes classifier

Remarks:

- If we work with 0-1-loss and if we know the underlying probability distribution and hence the regression function, then we don't need to "learn", we can simply write down what the optimal classifier is.
- For many other loss functions one can also explicitly compute the optimal classifier. We will see one more example in a minute: regression with squared loss.
- Problem in practice: we don't know the regression function.

Plug-in classifier

Simple idea: If we don't know the underlying distribution, but are given some training data, simply estimate the regression function $\eta(x)$ by some quantity $\eta_n(x)$ and build the plugin-classifier

$$f_n := \begin{cases} 1 & \text{if } \eta_n(x) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

- In theory: It can be shown that the plugin-approach is universally consistent. That is, in the limit of infinitely many training points, the classifier is going to converge to the best one out there. ☺
- In practice: Estimating densities is notoriously hard, in particular for high-dimensional input spaces. We would need a ridiculous amount of training data. So unfortunately, the plugin-approach is useless for practice. ☹

Loss functions for regression

While in classification, there is a "natural loss function" (the 0-1-loss), there exist many loss functions for regression and it is not so obvious which one is the most useful one.

In the following, let's look at the classic case, the squared loss function:

Squared loss (L_2-loss): $\ell(x, y, f(x)) = (f(x) - y)^2$

ID	f(x)	y	f(x) - y	|f(x)-y|	$(f(x)-y)^2$
1	99	100	-1	1	1
2	100	98	2	2	4
3	100	90	10	10	100
Sum			11	13	105

ID	f1(x)	y	f1(x) - y	|f1(x)-y|	$(f1(x)-y)^2$
1	95	100	-5	5	25
2	100	98	2	2	4
3	100	90	10	10	100
Sum			7	17	129

Regression function (for L_2 regression)

As in the classification setting, we define the regression function:

$$\eta(x) = E(Y \mid X = x)$$

We now want to show an explicit formula for the Bayes learner as well. As in the classification case, we fix a particular loss function, this time it is the squared loss.

We need one more intermediate result:

Regression function (for L_2 regression) (2)

Proposition 3 (Decomposition)

We always have

$$E\big(|f(X) - Y|^2\big) = E\big(|f(X) - \eta(X)|^2\big) + E\big(|\eta(X) - Y|^2\big).$$

Note: Getting a related inequality with $\leq$ is trivial (by the triangle inequality), but the equality in this statement is not trivial.

Explicit form of optimal solution under L_2 loss

Define the following learning rule that predicts the real-valued output based on the regression function η:

$$f^\circ : \mathcal{X} \to \mathbb{R}, \; f^\circ(x) := \eta(x)$$

Theorem 4 (Explicit form of optimal L_2-solution)

The function f° minimizes the L_2-risk.

Proof. Follows directly from Proposition 3:

- Second expectation on the rhs does not depend on f.
- First expectation is always ≥ 0, and it is $= 0$ for $f(X) = \eta(X)$.
- So the whole right hand side is minimized by $f(X) = \eta(X)$.

Two major principles

- Assume we operate in the standard setup, and are given a set of training points (X_i, Y_i).
- Based on these points we want to "learn" a function $f : \mathcal{X} \to \mathcal{Y}$ that has as small true loss as possible.

There are two major approaches to supervised learning:

- Empirical risk minimization (ERM)
- Regularized risk minimization (RRM)