7a - Feature Selection I

COM2004/3004

Jon Barker

Department of Computer Science
University of Sheffield

Autumn Semester

Lecture Objectives

In this lecture we will

- Review the three main data-driven computing problem types
- Review the structure of a data-driven classification system
- Explain the importance of good feature selection
- Introduce the concept of divergence

Data Driven Computing

Three major classes of data-driven problem:

- **Classification** – attaching a label.
  "animal, vegetable or mineral?"; "male or female?" ; "healthy or cancerous"; "whose face is this?"; "which letter is this?"
- **Regression** – estimating or predicting a response.
  Forecasting tomorrow's weather; tracking a moving object; forecasting financial markets
- **Clustering/Visualisation** – looking for patterns.
  Interactive discovery of meaningful structure in complex data sets.

Data Driven Computing

Three major classes of data-driven problem:

- **Classification** – attaching a label.
  "animal, vegetable or mineral?"; "male or female?" ; "healthy or cancerous"; "whose face is this?"; "which letter is this?"
- **Regression** – estimating or predicting a response.
  Forecasting tomorrow's weather; tracking a moving object; forecasting financial markets
- **Clustering/Visualisation** – looking for patterns.
  Interactive discovery of meaningful structure in complex data sets.

Data Driven Computing

Three major classes of data-driven problem:

- **Classification** – attaching a label.
  "animal, vegetable or mineral?"; "male or female?" ; "healthy or cancerous"; "whose face is this?"; "which letter is this?"
- **Regression** – estimating or predicting a response.
  Forecasting tomorrow's weather; tracking a moving object; forecasting financial markets
- **Clustering/Visualisation** – looking for patterns.
  Interactive discovery of meaningful structure in complex data sets.

Classification - central to many applications

- Speech recognition
  - Input: acoustic signal from microphone
  - Output: word labels
- Speaker verification
  - Input: acoustic signal from microphone
  - Output: impostor / not-impostor
- Handwriting recognition
  - Input: sequence of pen movements
  - Output: word labels
- Smile detection
  - Input: image
  - Output: face smiling / face not smiling
- Finger print recognition
  - Input: image (optical, ultrasound or capacitance sensor)
  - Output: person ID
- Traffic sign detection
  - Input: video stream from car
  - Output: traffic sign labels
- Counter terrorism
  - Input: intercepted emails / telephone calls
  - Output: suspicious / not suspicious

Classification - central to many applications

- Speech recognition
  - Input: acoustic signal from microphone
  - Output: word labels
- Speaker verification
  - Input: acoustic signal from microphone
  - Output: impostor / not-impostor
- Handwriting recognition
  - Input: sequence of pen movements
  - Output: word labels
- Smile detection
  - Input: image
  - Output: face smiling / face not smiling
- Finger print recognition
  - Input: image (optical, ultrasound or capacitance sensor)
  - Output: person ID
- Traffic sign detection
  - Input: video stream from car
  - Output: traffic sign labels
- Counter terrorism
  - Input: intercepted emails / telephone calls
  - Output: suspicious / not suspicious

Classification - central to many applications

- Speech recognition
  - Input: acoustic signal from microphone
  - Output: word labels
- Speaker verification
  - Input: acoustic signal from microphone
  - Output: impostor / not-impostor
- Handwriting recognition
  - Input: sequence of pen movements
  - Output: word labels
- Smile detection
  - Input: image
  - Output: face smiling / face not smiling
- Finger print recognition
  - Input: image (optical, ultrasound or capacitance sensor)
  - Output: person ID
- Traffic sign detection
  - Input: video stream from car
  - Output: traffic sign labels
- Counter terrorism
  - Input: intercepted emails / telephone calls
  - Output: suspicious / not suspicious

Classification - central to many applications

- Speech recognition
  - Input: acoustic signal from microphone
  - Output: word labels
- Speaker verification
  - Input: acoustic signal from microphone
  - Output: impostor / not-impostor
- Handwriting recognition
  - Input: sequence of pen movements
  - Output: word labels
- Smile detection
  - Input: image
  - Output: face smiling / face not smiling
- Finger print recognition
  - Input: image (optical, ultrasound or capacitance sensor)
  - Output: person ID
- Traffic sign detection
  - Input: video stream from car
  - Output: traffic sign labels
- Counter terrorism
  - Input: intercepted emails / telephone calls
  - Output: suspicious / not suspicious

Classification - central to many applications

- Speech recognition
  - Input: acoustic signal from microphone
  - Output: word labels
- Speaker verification
  - Input: acoustic signal from microphone
  - Output: impostor / not-impostor
- Handwriting recognition
  - Input: sequence of pen movements
  - Output: word labels
- Smile detection
  - Input: image
  - Output: face smiling / face not smiling
- Finger print recognition
  - Input: image (optical, ultrasound or capacitance sensor)
  - Output: person ID
- Traffic sign detection
  - Input: video stream from car
  - Output: traffic sign labels
- Counter terrorism
  - Input: intercepted emails / telephone calls
  - Output: suspicious / not suspicious

Classification - central to many applications

- Speech recognition
  - Input: acoustic signal from microphone
  - Output: word labels
- Speaker verification
  - Input: acoustic signal from microphone
  - Output: impostor / not-impostor
- Handwriting recognition
  - Input: sequence of pen movements
  - Output: word labels
- Smile detection
  - Input: image
  - Output: face smiling / face not smiling
- Finger print recognition
  - Input: image (optical, ultrasound or capacitance sensor)
  - Output: person ID
- Traffic sign detection
  - Input: video stream from car
  - Output: traffic sign labels
- Counter terrorism
  - Input: intercepted emails / telephone calls
  - Output: suspicious / not suspicious

Classification - central to many applications

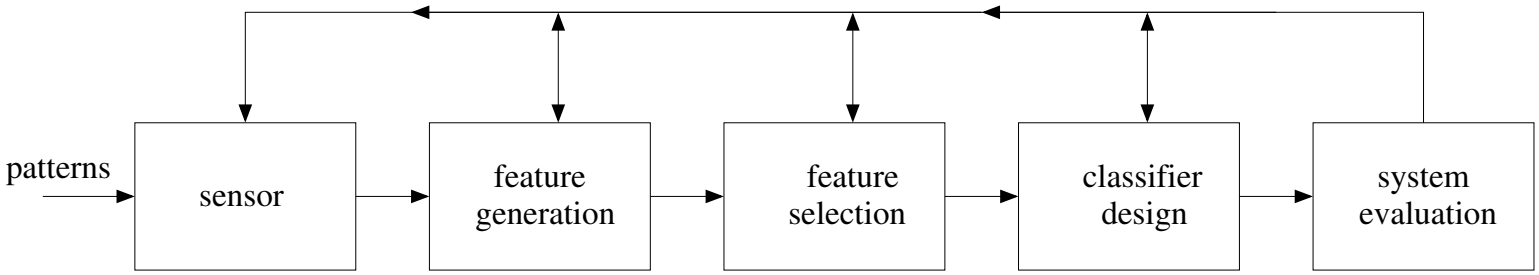
- Speech recognition
  - Input: acoustic signal from microphone
  - Output: word labels
- Speaker verification
  - Input: acoustic signal from microphone
  - Output: impostor / not-impostor
- Handwriting recognition
  - Input: sequence of pen movements
  - Output: word labels
- Smile detection
  - Input: image
  - Output: face smiling / face not smiling
- Finger print recognition
  - Input: image (optical, ultrasound or capacitance sensor)
  - Output: person ID
- Traffic sign detection
  - Input: video stream from car
  - Output: traffic sign labels
- Counter terrorism
  - Input: intercepted emails / telephone calls
  - Output: suspicious / not suspicious

Classification demos

Scrabble player

Traffic sign detection

Designing a classifier



- **sensor** – capture data from physical world, e.g. microphone, camera, array of temperature sensors, blood pressure monitor etc etc
- **feature generation** – generate candidate features from raw data, e.g. frequency analysis of sound wave; edge detection in image data
- **feature selection** – choose subset of features that carry most information
- **classifier** – optimal design depends on statistical properties of features
- **evaluation** – test system to measure performance, redesign

Feature Selection

- Given a large number of candidate features, how do we decide which features will be most useful?
- Why not just use **all** the candidate features?
- For a large set of “traditional” classifiers, given a fixed amount of training data, there is an optimum number of features. (T&K p.265)
- Adding features beyond this number leads to “overfitting” and loss of “generalisation”
- (This is not true of all classifiers and the cause has only been properly understood in the last 20 years)

Feature Selection

- Given a large number of candidate features, how do we decide which features will be most useful?
- Why not just use **all** the candidate features?
- For a large set of “traditional” classifiers, given a fixed amount of training data, there is an optimum number of features. (T&K p.265)
- Adding features beyond this number leads to “overfitting” and loss of “generalisation”
- (This is not true of all classifiers and the cause has only been properly understood in the last 20 years)

Feature Selection

- Given a large number of candidate features, how do we decide which features will be most useful?
- Why not just use **all** the candidate features?
- For a large set of “traditional” classifiers, given a fixed amount of training data, there is an optimum number of features. (T&K p.265)
- Adding features beyond this number leads to “overfitting” and loss of “generalisation”
- (This is not true of all classifiers and the cause has only been properly understood in the last 20 years)

Feature Selection

- Given a large number of candidate features, how do we decide which features will be most useful?
- Why not just use **all** the candidate features?
- For a large set of “traditional” classifiers, given a fixed amount of training data, there is an optimum number of features. (T&K p.265)
- Adding features beyond this number leads to “overfitting” and loss of “generalisation”
- (This is not true of all classifiers and the cause has only been properly understood in the last 20 years)

Feature Selection

- Given a large number of candidate features, how do we decide which features will be most useful?
- Why not just use **all** the candidate features?
- For a large set of “traditional” classifiers, given a fixed amount of training data, there is an optimum number of features. (T&K p.265)
- Adding features beyond this number leads to “overfitting” and loss of “generalisation”
- (This is not true of all classifiers and the cause has only been properly understood in the last 20 years)

Feature Selection

- Given that we can't use all the features, how do we decide which features will be most useful?
- e.g. Consider the 900 pixel values that make up the image in your assignment. Are some pixels more informative than others?
- To motivate our discussion we consider the following toy problem
  - Task: To classify students as **male** or **female** according to some survey data.
  - Our survey will collect a large number of measurements such as height, weight, shoe size.

Feature Selection

- Given that we can't use all the features, how do we decide which features will be most useful?
- e.g. Consider the 900 pixel values that make up the image in your assignment. Are some pixels more informative than others?
- To motivate our discussion we consider the following toy problem
  - Task: To classify students as **male** or **female** according to some survey data.
  - Our survey will collect a large number of measurements such as height, weight, shoe size.

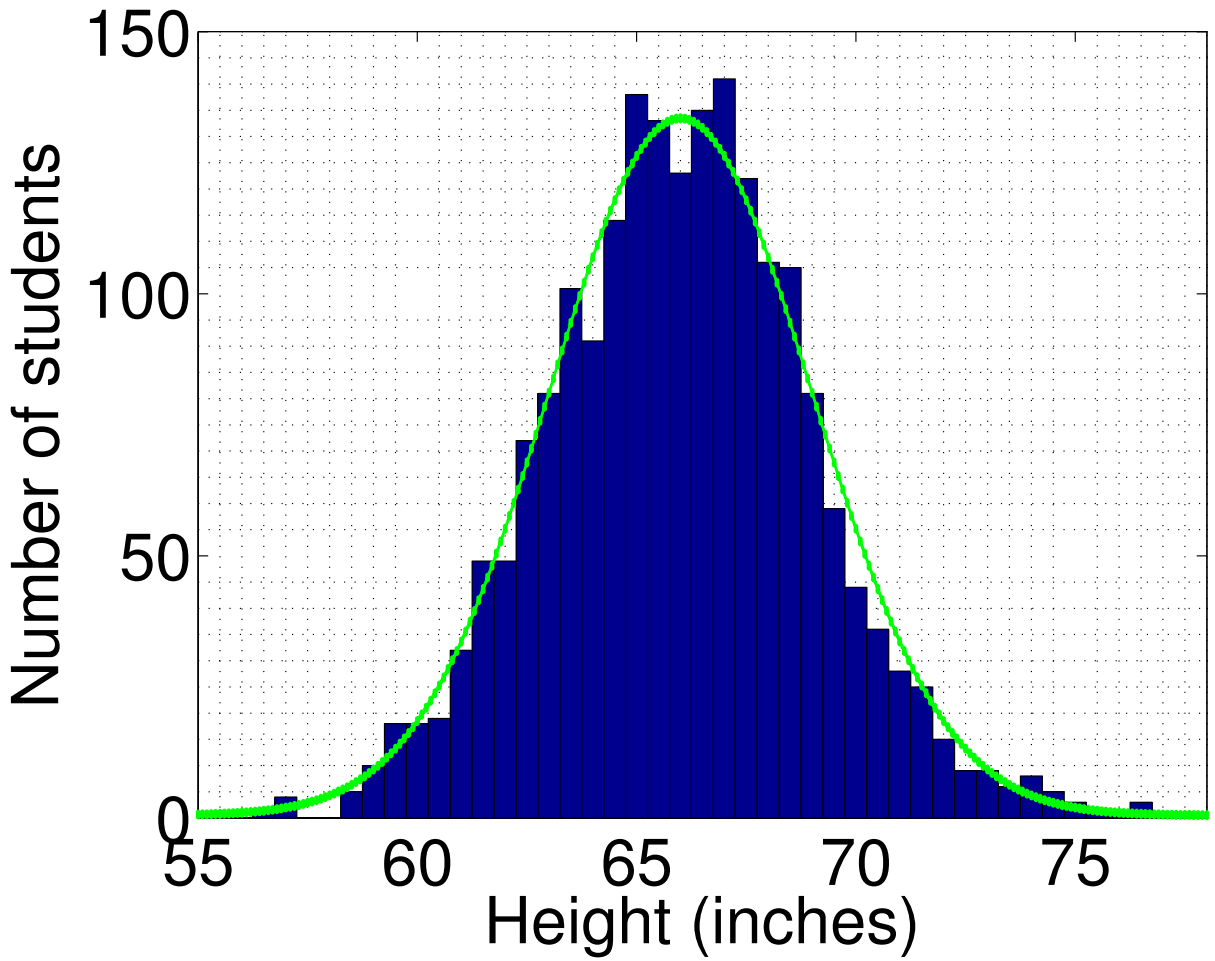
Feature Selection

- Given that we can't use all the features, how do we decide which features will be most useful?
- e.g. Consider the 900 pixel values that make up the image in your assignment. Are some pixels more informative than others?
- To motivate our discussion we consider the following toy problem
  - Task: To classify students as **male** or **female** according to some survey data.
  - Our survey will collect a large number of measurements such as height, weight, shoe size.

Gaussian distribution

Consider measurements such as weight, height, arm length. Such measurements tend to have a Gaussian distribution.

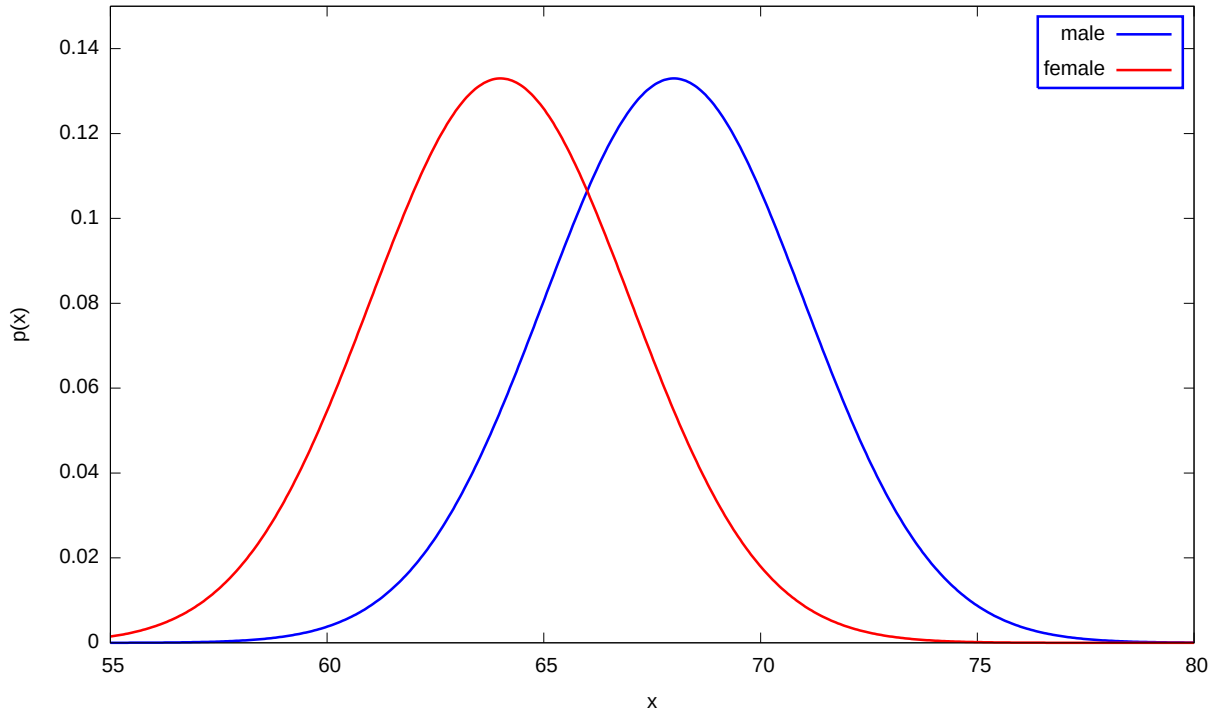
e.g. histogram of the height of 2,000 university students.



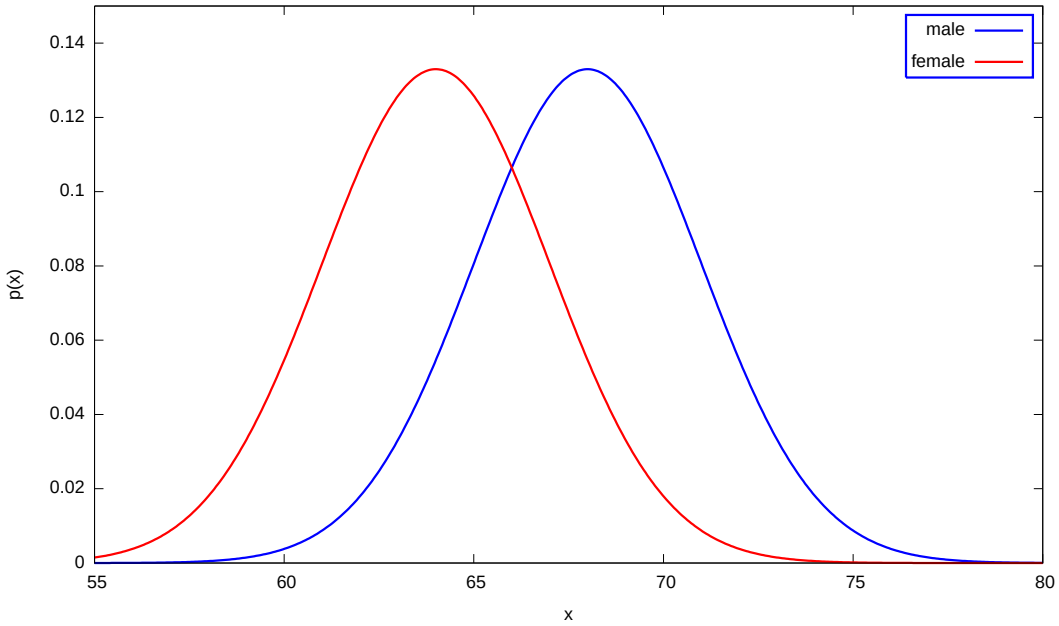
A simple classifier

Consider a classifier based on one feature: height.

- Both men and women's' heights will have a Gaussian distribution.
- But, men are *'on average'* a little taller than women.
- The mean of the male height distribution is a little higher.
- e.g. the distribution may look like this, with women spread around 64 inches and mean around 68.



A simple classifier



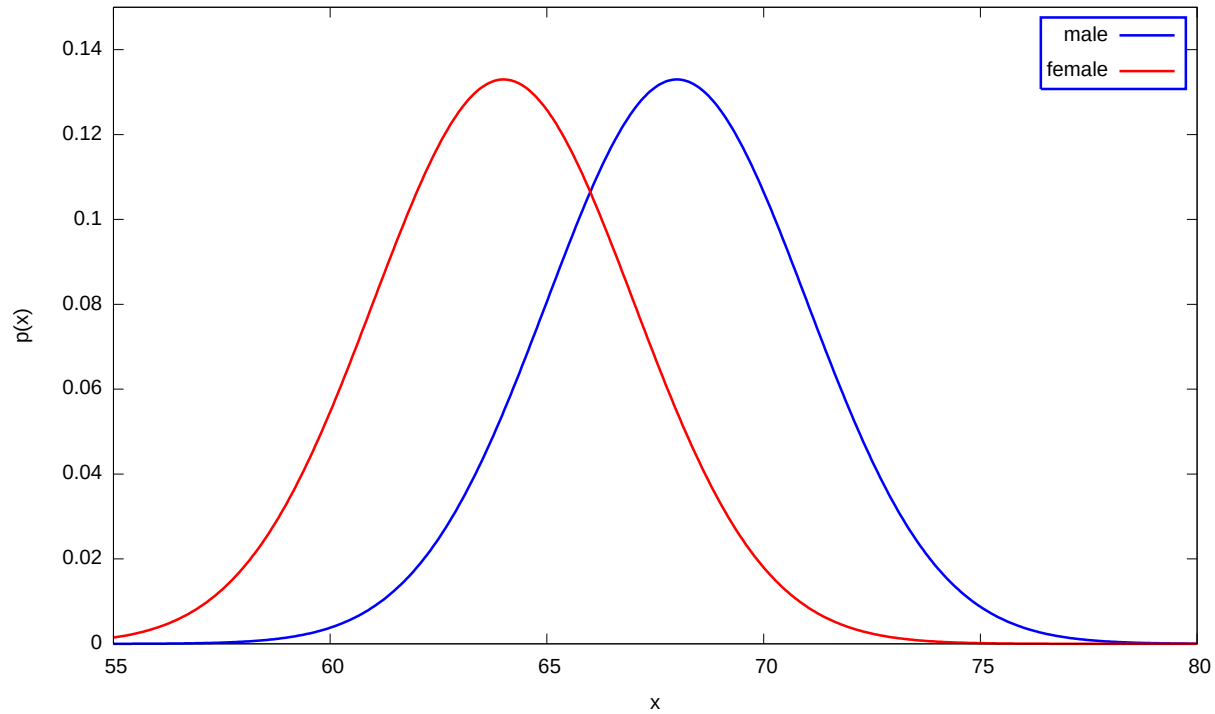
We can use the following classification procedure:

- Using our knowledge of the distributions choose some threshold T
- We measure someone's height, x.
- We compare x to the threshold value, T.
- If $x \geq T$ output *male* else output *female*.

A simple classifier

How do we set the threshold value?

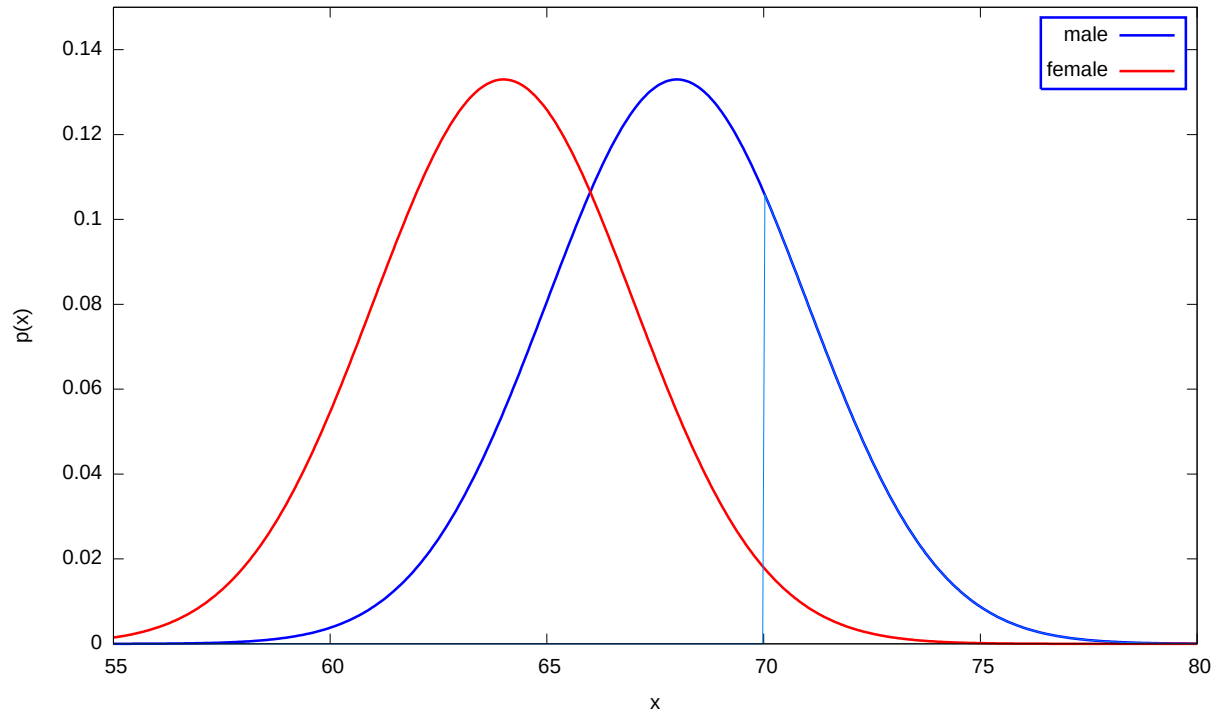
- Imagine setting threshold to 70 inches.
- Very tall women will be misclassified.
- Many shorter men will be misclassified.
- Shaded regions represent misclassifications.



A simple classifier

How do we set the threshold value?

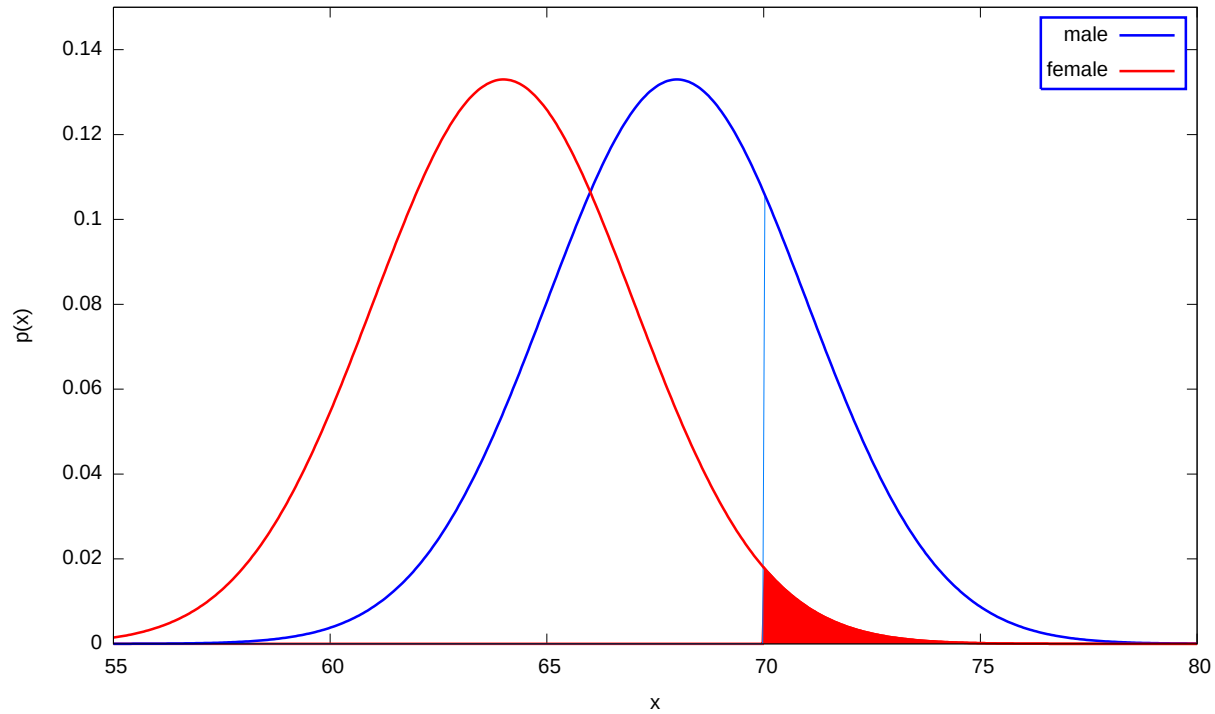
- Imagine setting threshold to 70 inches.
- Very tall women will be misclassified.
- Many shorter men will be misclassified.
- Shaded regions represent misclassifications.



A simple classifier

How do we set the threshold value?

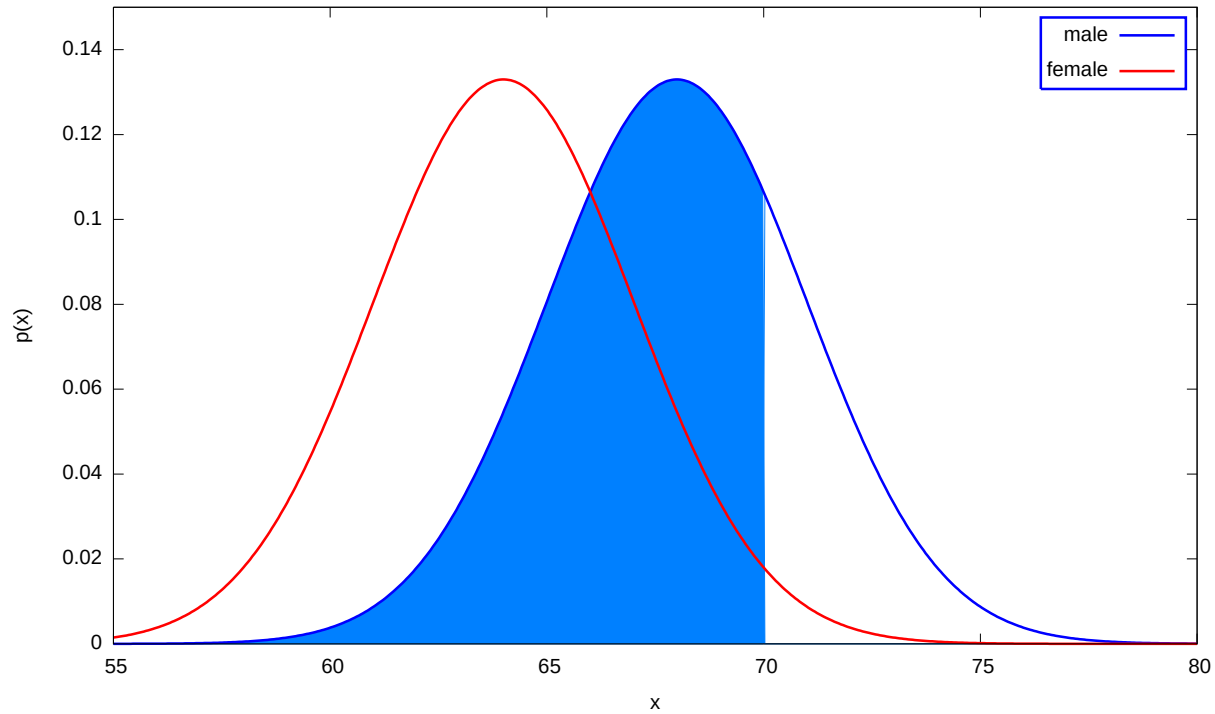
- Imagine setting threshold to 70 inches.
- Very tall women will be misclassified.
- Many shorter men will be misclassified.
- Shaded regions represent misclassifications.



A simple classifier

How do we set the threshold value?

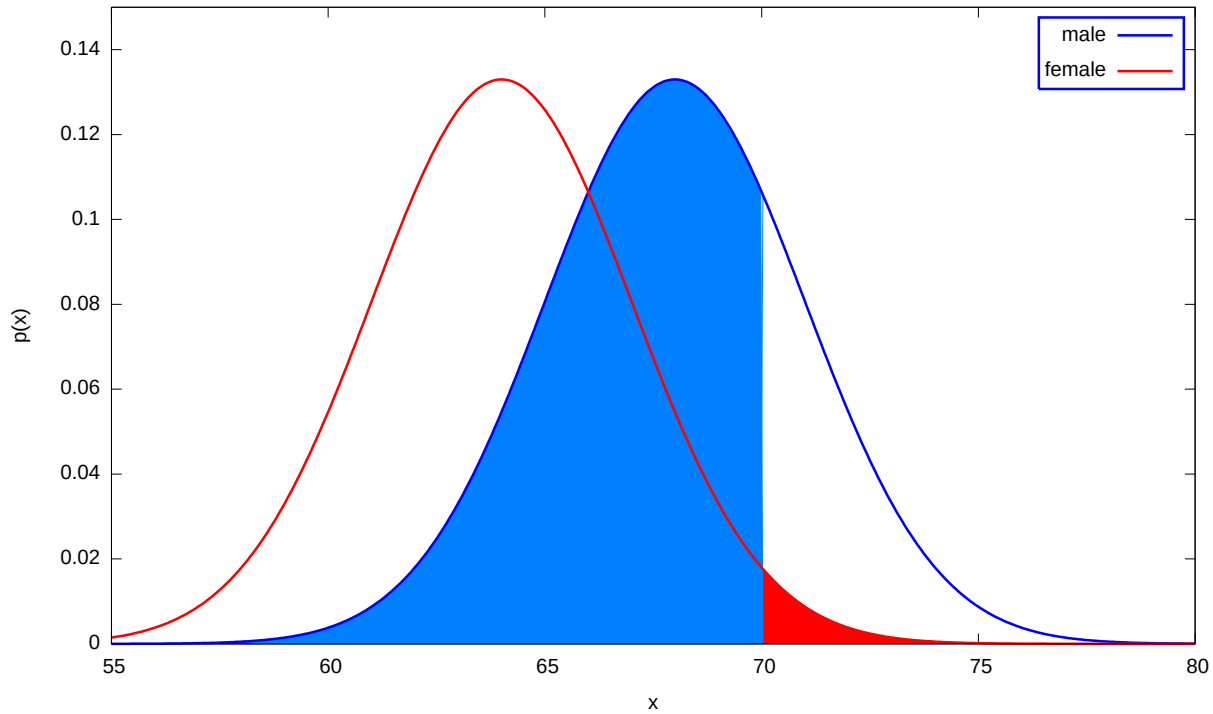
- Imagine setting threshold to 70 inches.
- Very tall women will be misclassified.
- Many shorter men will be misclassified.
- Shaded regions represent misclassifications.



A simple classifier

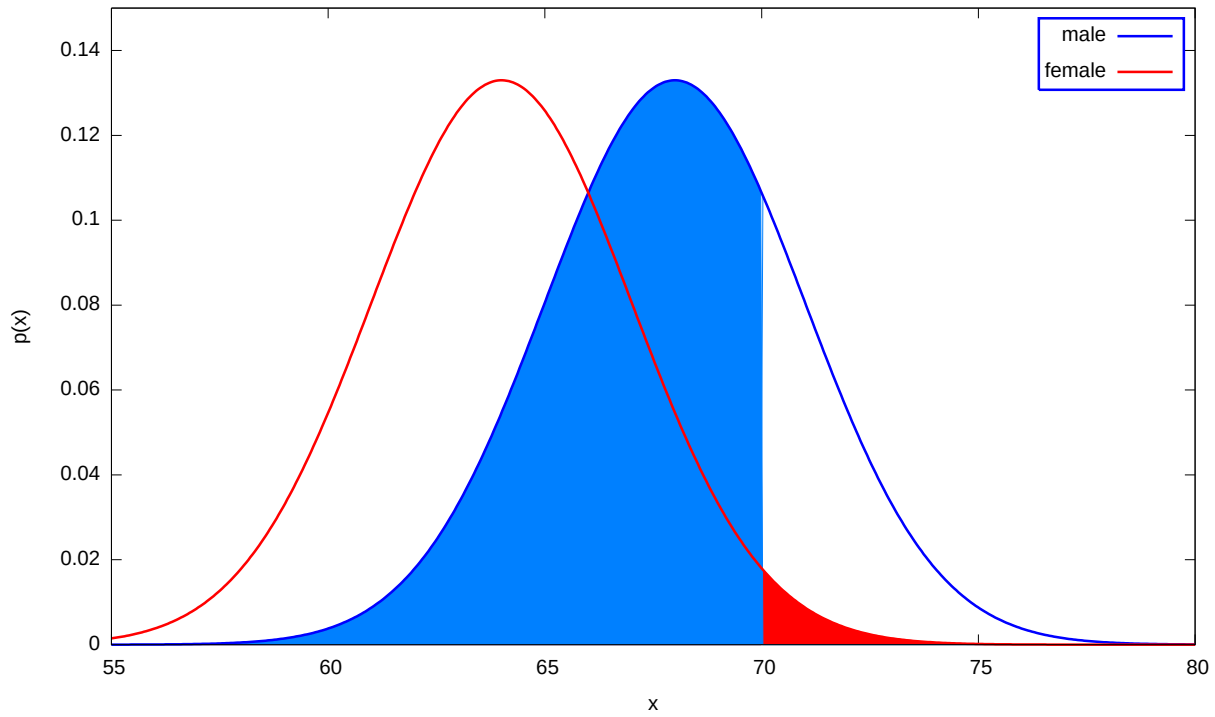
How do we set the threshold value?

- Imagine setting threshold to 70 inches.
- Very tall women will be misclassified.
- Many shorter men will be misclassified.
- Shaded regions represent misclassifications.



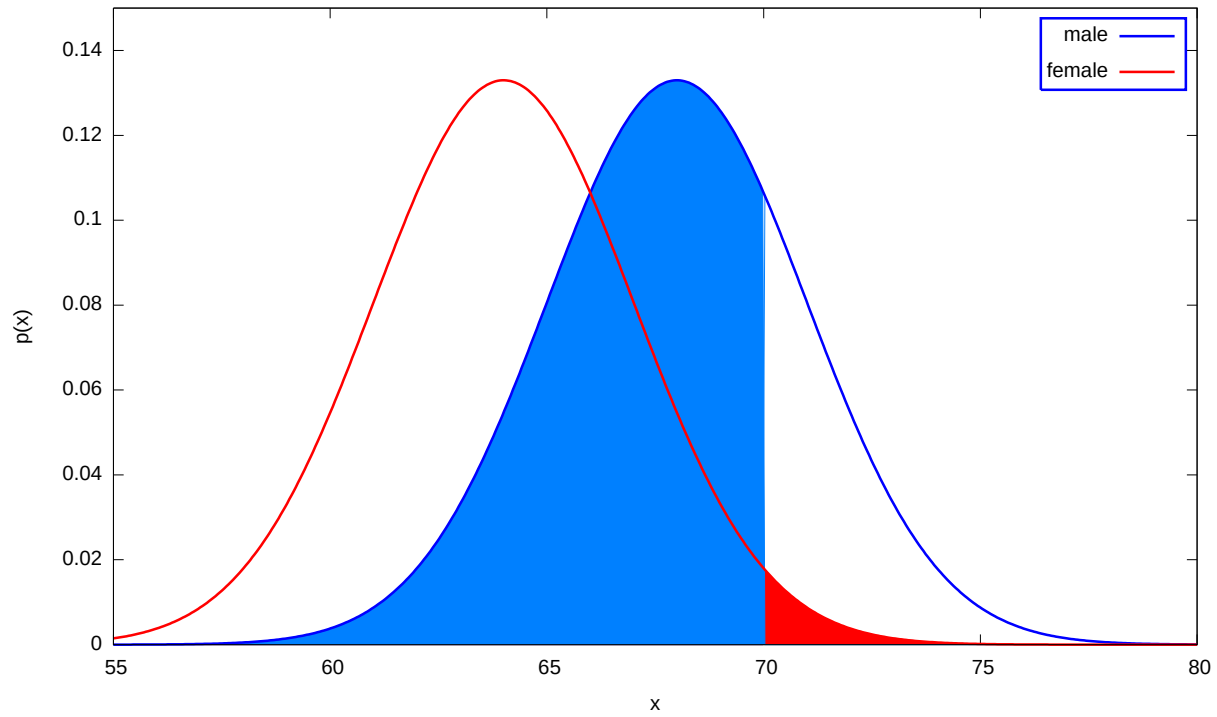
Setting the threshold

- How do we choose the threshold?
- Easy. Pick threshold that minimises errors. How?
- Easy. Need to minimise shared area.
- Optimum threshold is where the distributions cross.
- If distributions have equal variance, they cross at the midpoint between their mean values.
- In our example the optimum threshold will be 66 inches.



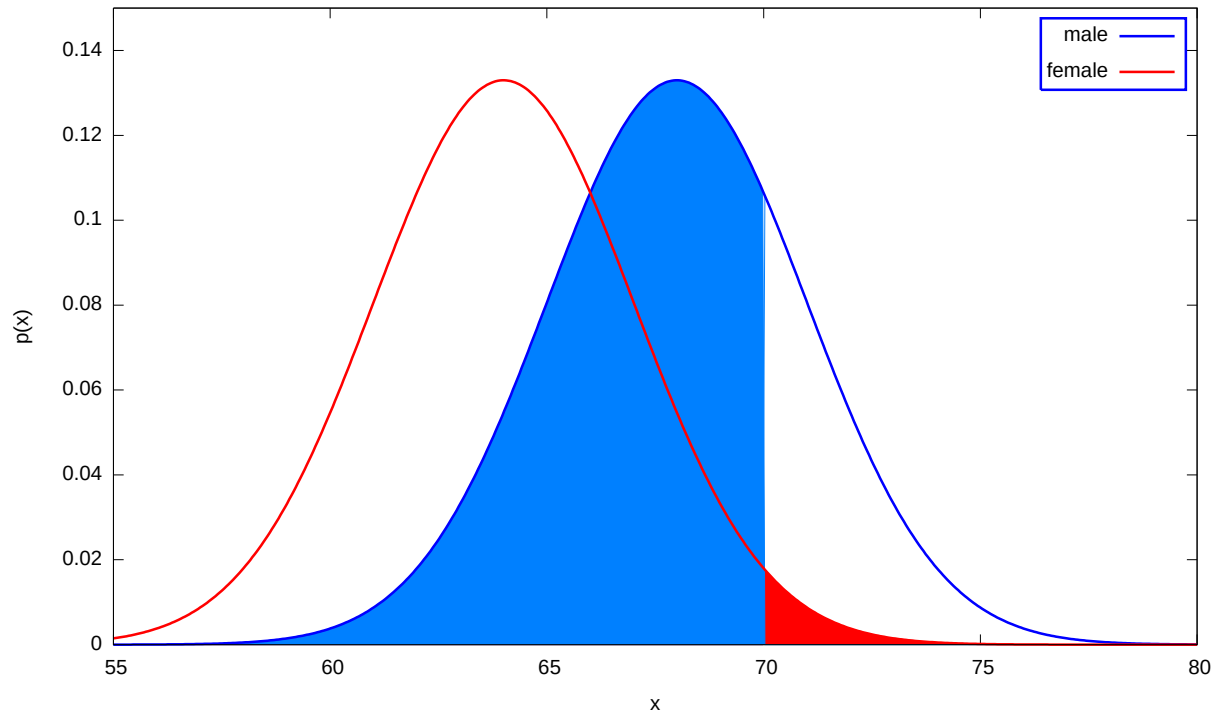
Setting the threshold

- How do we choose the threshold?
- Easy. Pick threshold that minimises errors. How?
- Easy. Need to minimise shared area.
- Optimum threshold is where the distributions cross.
- If distributions have equal variance, they cross at the midpoint between their mean values.
- In our example the optimum threshold will be 66 inches.



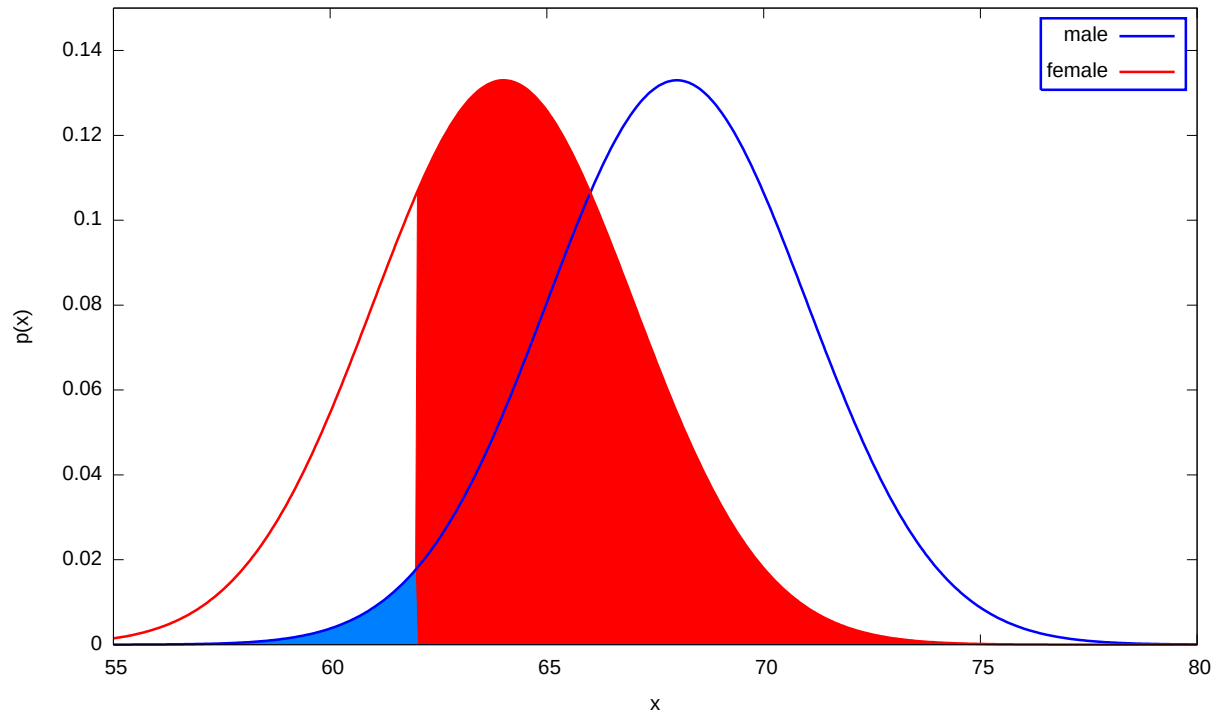
Setting the threshold

- How do we choose the threshold?
- Easy. Pick threshold that minimises errors. How?
- Easy. Need to minimise shared area.
- Optimum threshold is where the distributions cross.
- If distributions have equal variance, they cross at the midpoint between their mean values.
- In our example the optimum threshold will be 66 inches.



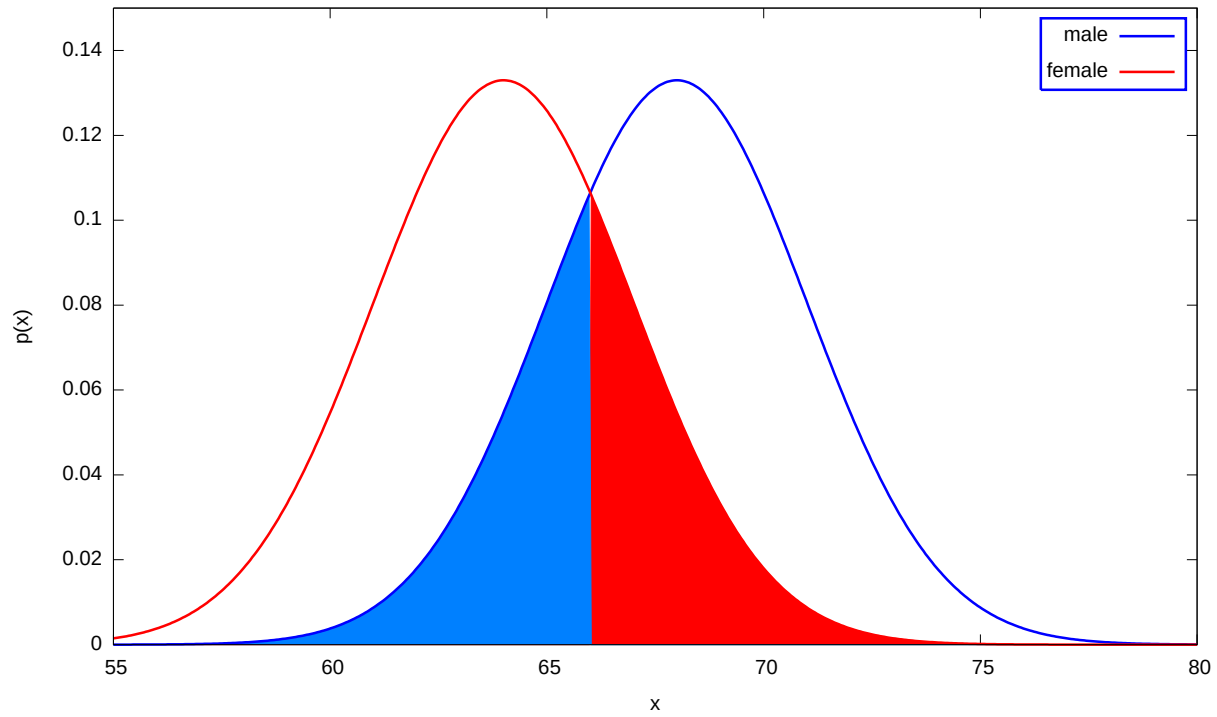
Setting the threshold

- How do we choose the threshold?
- Easy. Pick threshold that minimises errors. How?
- Easy. Need to minimise shared area.
- Optimum threshold is where the distributions cross.
- If distributions have equal variance, they cross at the midpoint between their mean values.
- In our example the optimum threshold will be 66 inches.



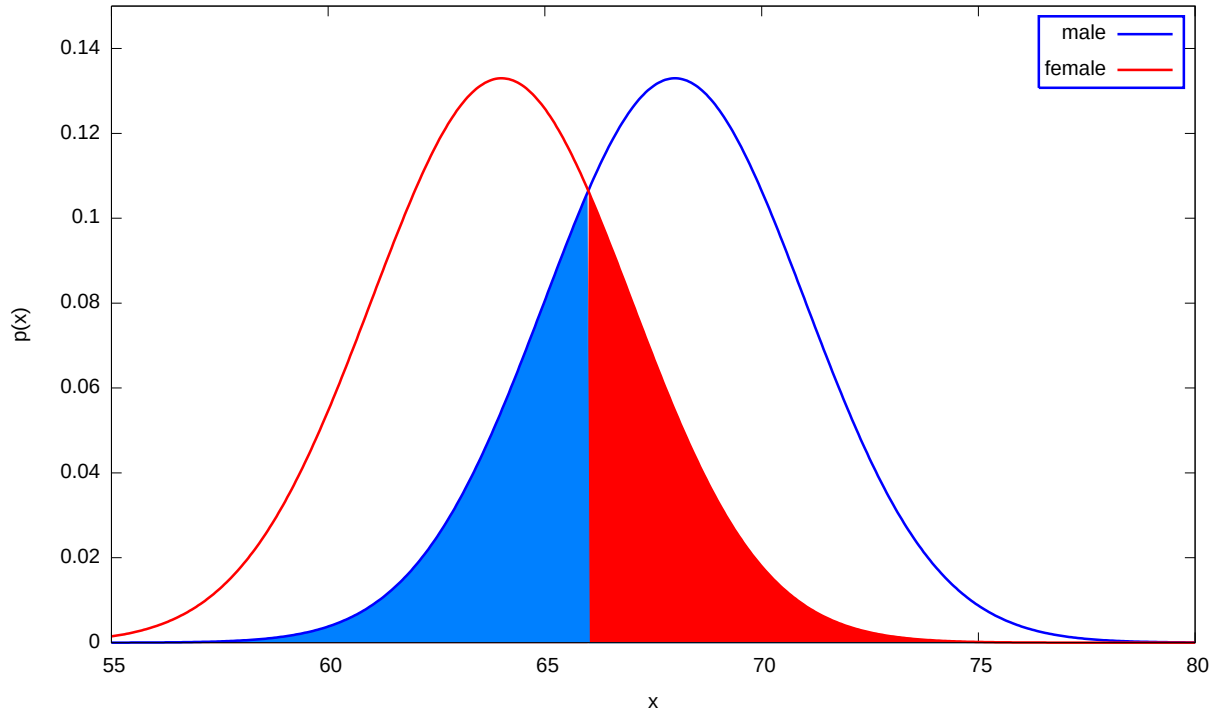
Setting the threshold

- How do we choose the threshold?
- Easy. Pick threshold that minimises errors. How?
- Easy. Need to minimise shared area.
- Optimum threshold is where the distributions cross.
- If distributions have equal variance, they cross at the midpoint between their mean values.
- In our example the optimum threshold will be 66 inches.



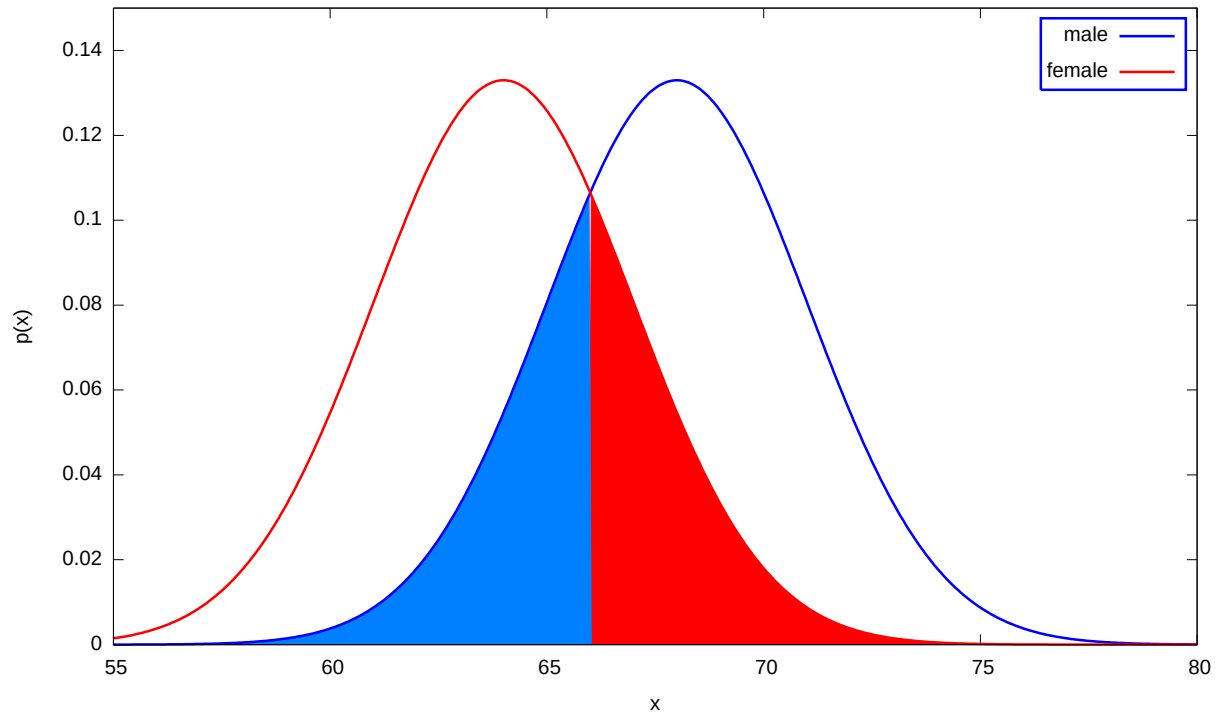
Setting the threshold

- How do we choose the threshold?
- Easy. Pick threshold that minimises errors. How?
- Easy. Need to minimise shared area.
- Optimum threshold is where the distributions cross.
- If distributions have equal variance, they cross at the midpoint between their mean values.
- In our example the optimum threshold will be 66 inches.



Setting the threshold

- How do we choose the threshold?
- Easy. Pick threshold that minimises errors. How?
- Easy. Need to minimise shared area.
- Optimum threshold is where the distributions cross.
- If distributions have equal variance, they cross at the midpoint between their mean values.
- In our example the optimum threshold will be 66 inches.



What makes a good feature?

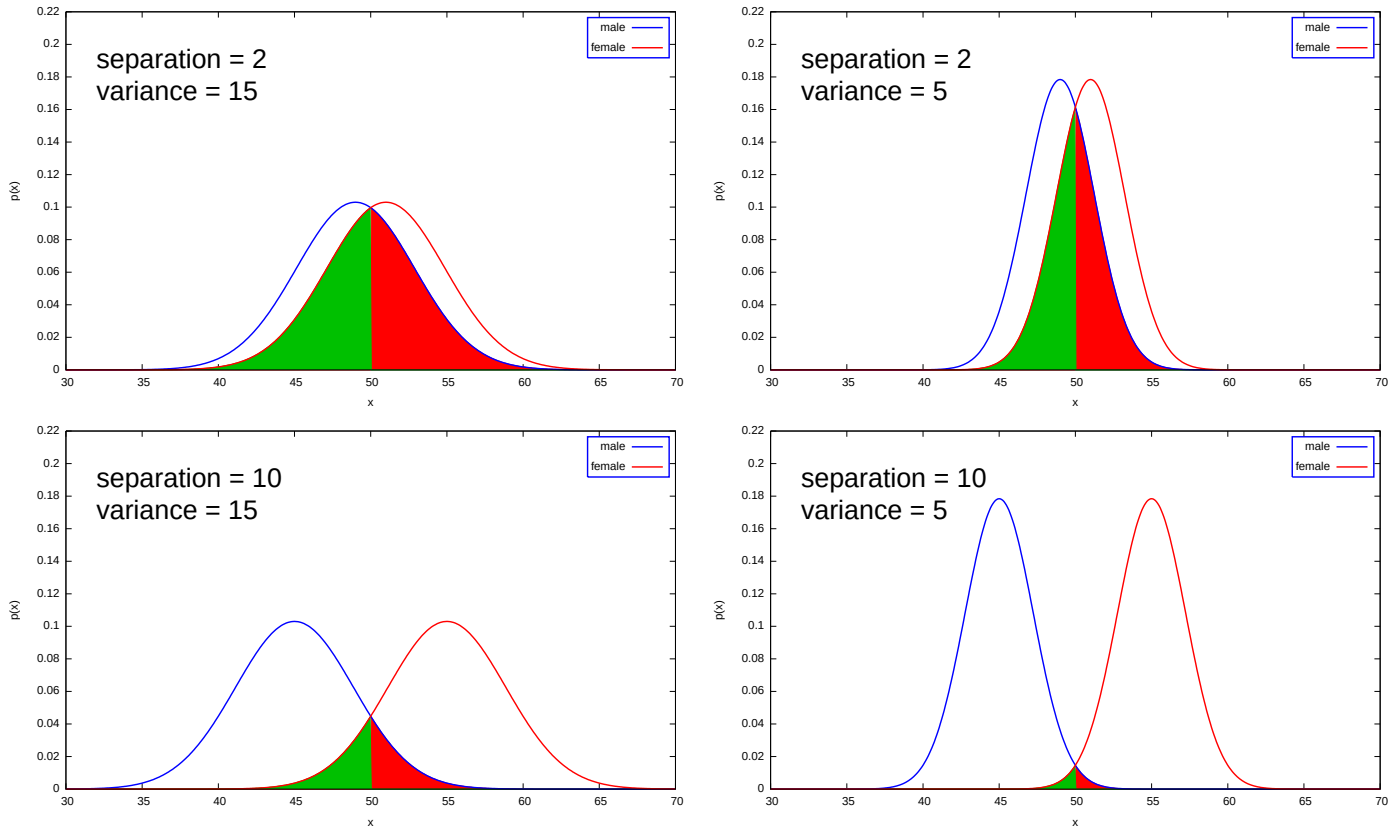
Consider a system for classifying adults as male or female.

Which of the following might be a useful feature:

- hair length
- shoe size
- eye colour
- height
- weight
- length of left arm
- length of right arm

What makes a good feature?

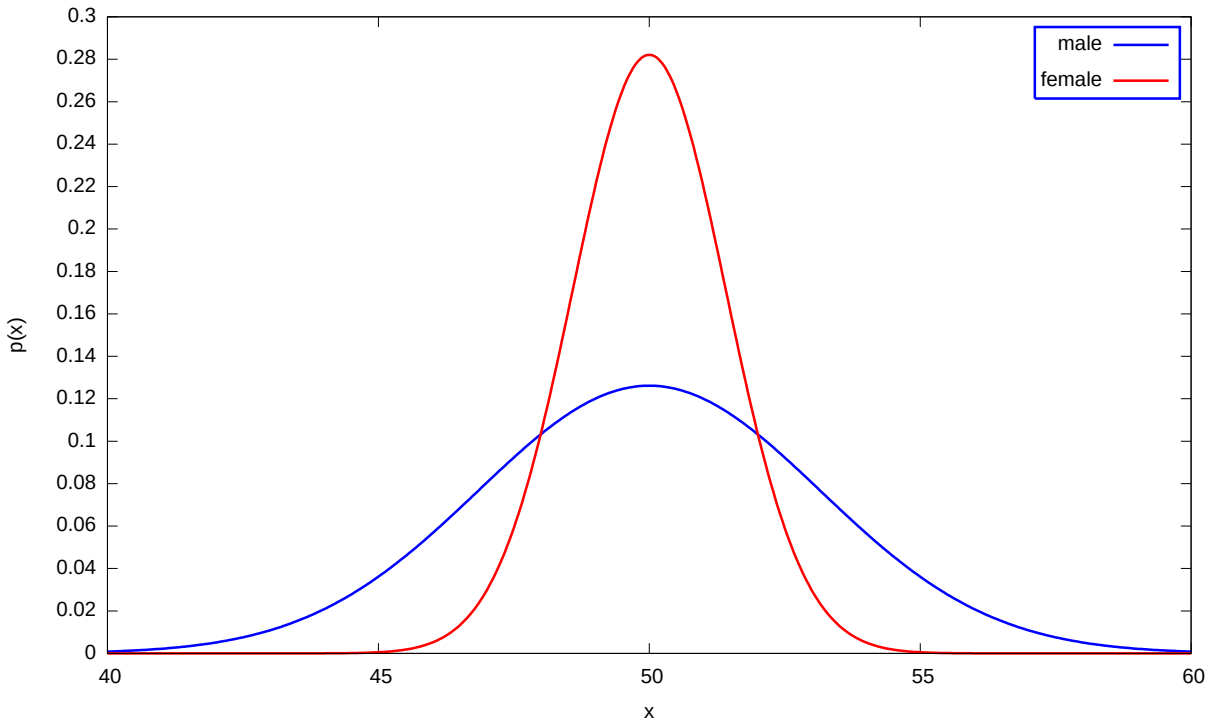
We are looking for features whose distributions have the smallest overlap.



Generally, looking for features where classes have a **large separation between means** and a **small variance**.

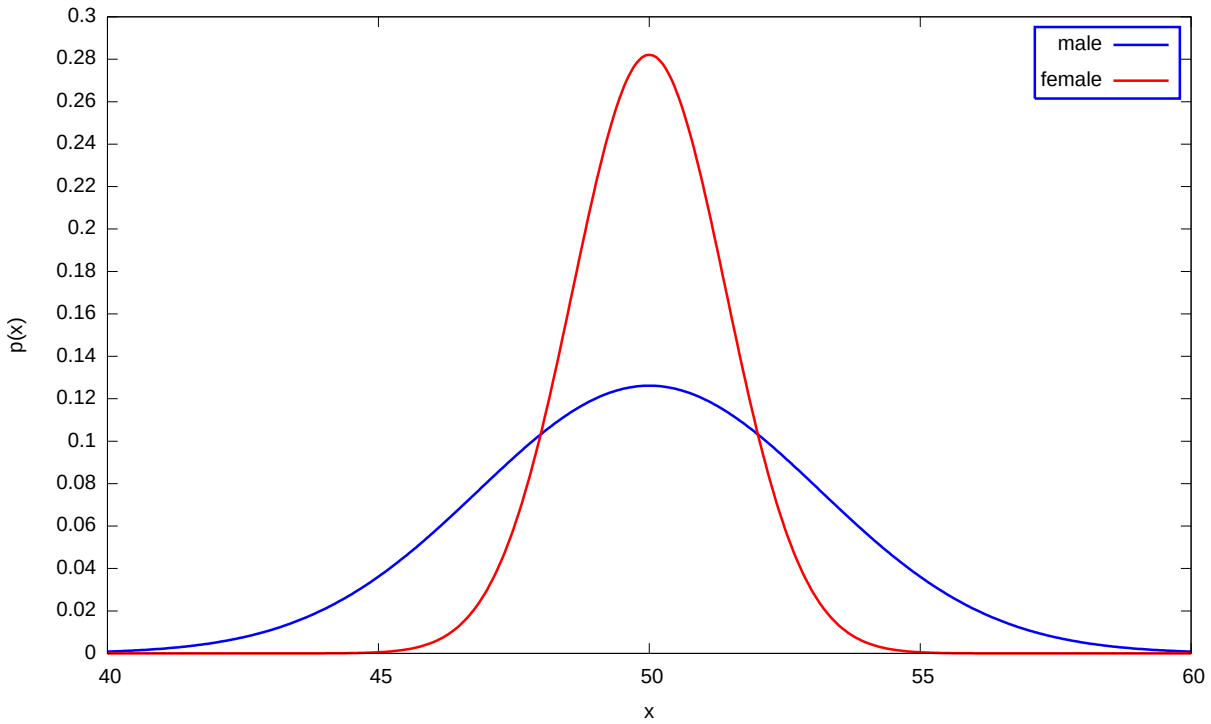
Equal means

- Consider the following feature: the male and female mean is the same, but the variances (i.e., spread) are different.
- Can we classifying using this feature?
- Yes. The extreme measurements are more likely to belong to the male class.
- Again, shaded regions represent misclassifications.



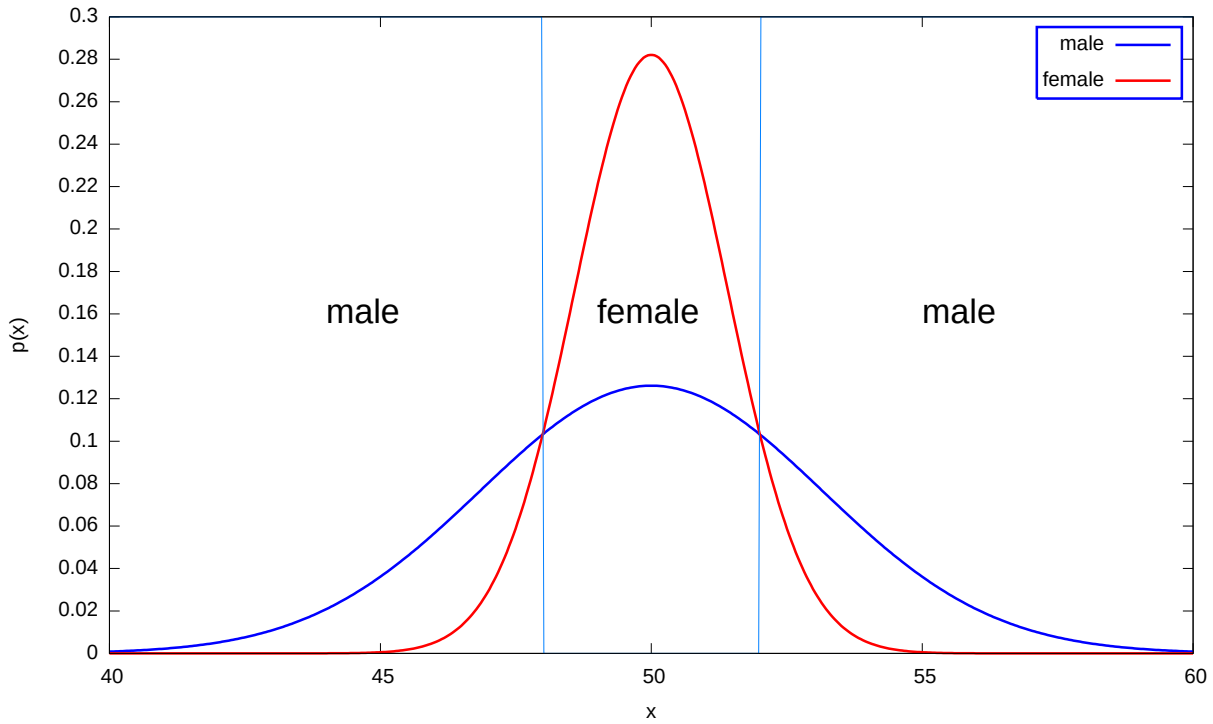
Equal means

- Consider the following feature: the male and female mean is the same, but the variances (i.e., spread) are different.
- Can we classifying using this feature?
- Yes. The extreme measurements are more likely to belong to the male class.
- Again, shaded regions represent misclassifications.



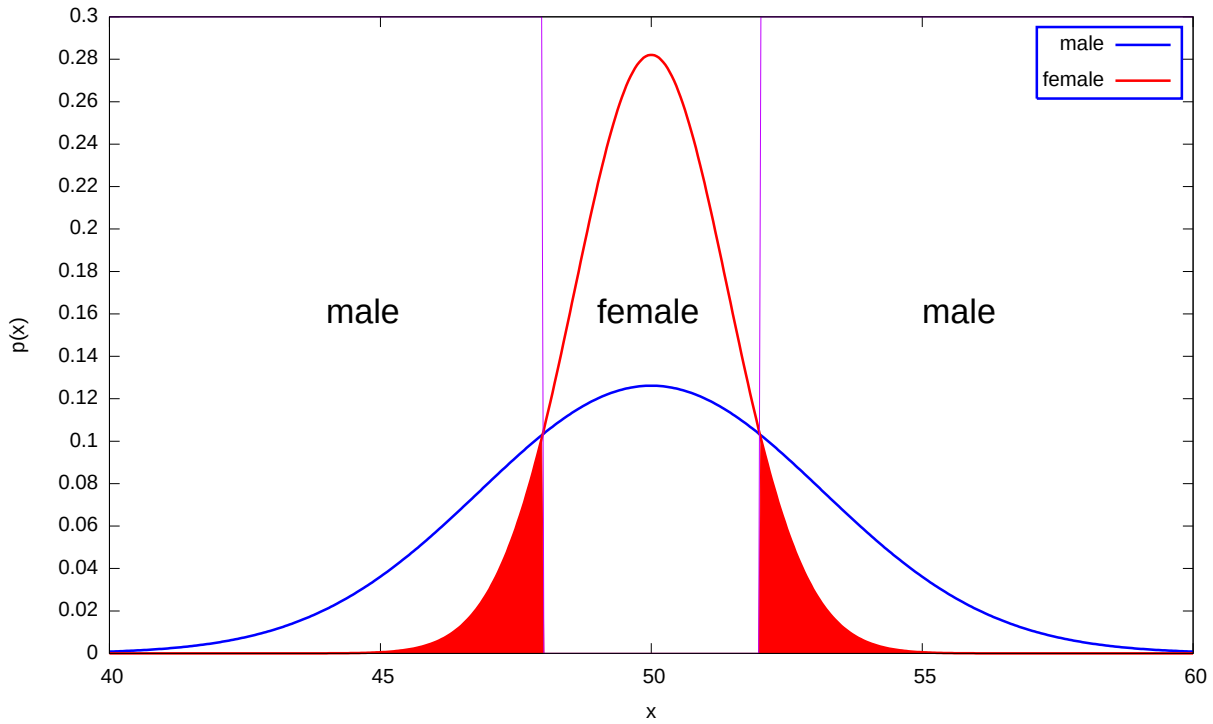
Equal means

- Consider the following feature: the male and female mean is the same, but the variances (i.e., spread) are different.
- Can we classifying using this feature?
- Yes. The extreme measurements are more likely to belong to the male class.
- Again, shaded regions represent misclassifications.



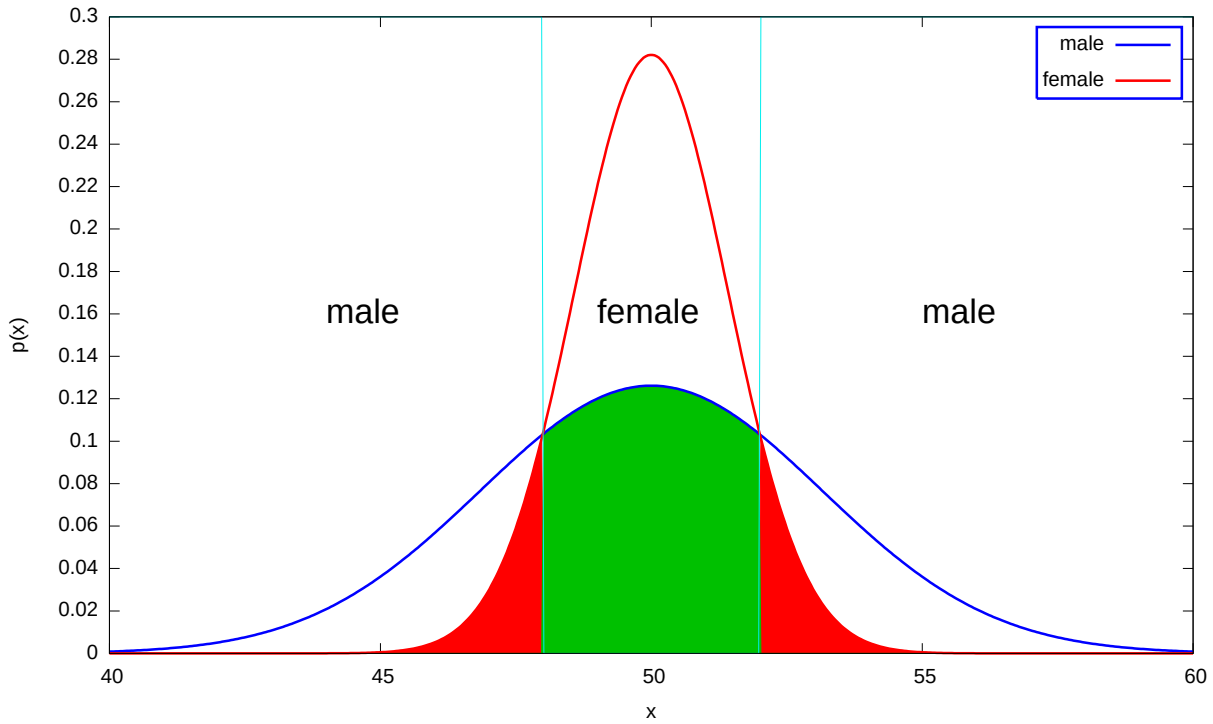
Equal means

- Consider the following feature: the male and female mean is the same, but the variances (i.e., spread) are different.
- Can we classifying using this feature?
- Yes. The extreme measurements are more likely to belong to the male class.
- Again, shaded regions represent misclassifications.



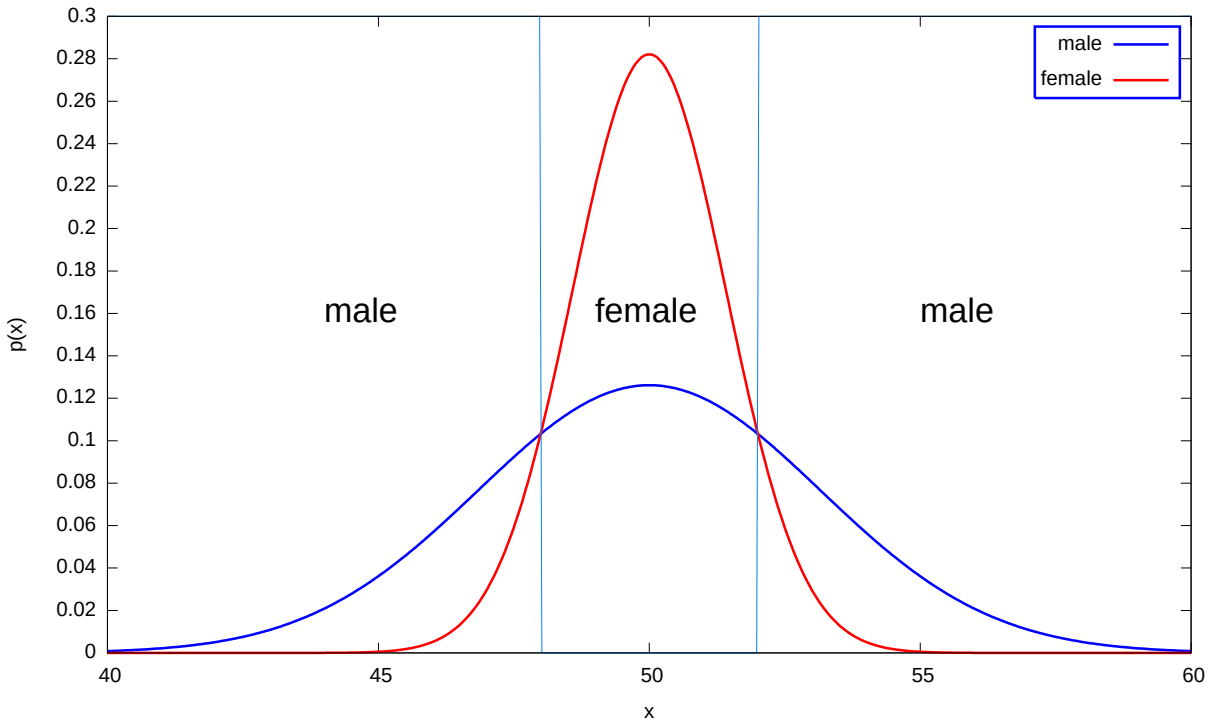
Equal means

- Consider the following feature: the male and female mean is the same, but the variances (i.e., spread) are different.
- Can we classifying using this feature?
- Yes. The extreme measurements are more likely to belong to the male class.
- Again, shaded regions represent misclassifications.



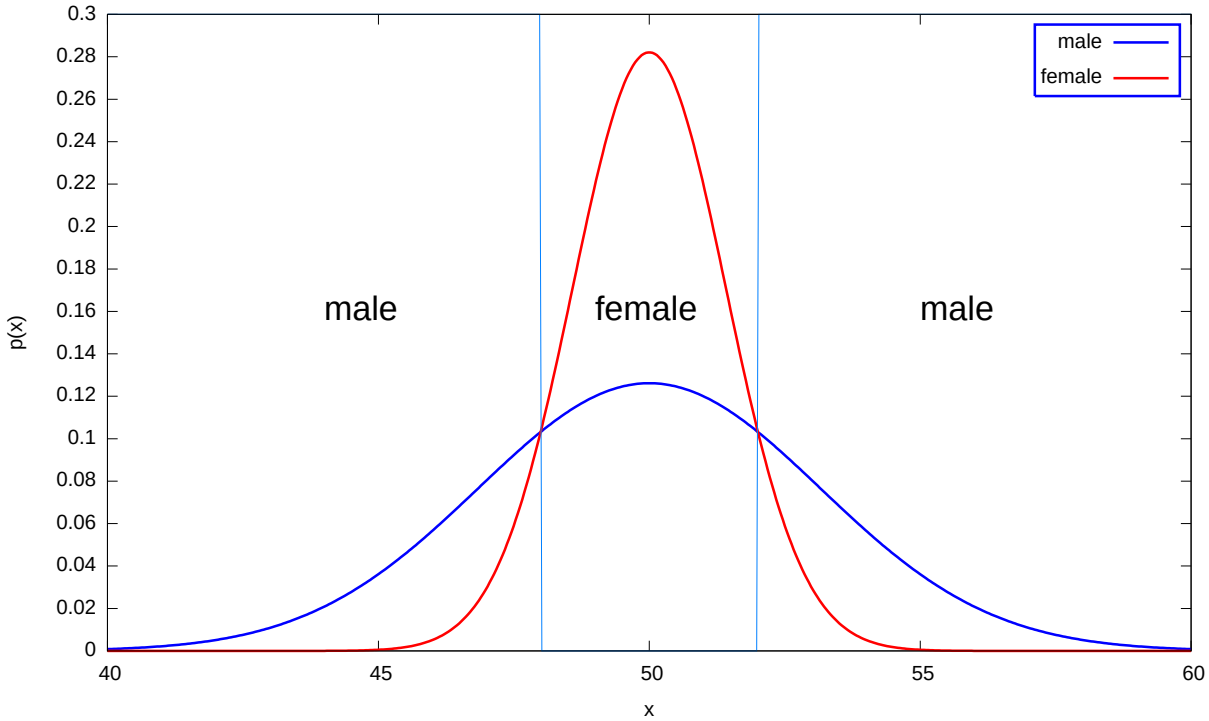
Bayes' Classification Rule

- If $p(male|x) > p(female|x)$ output *male* else output *female*.
- $p(x|male)p(male) > p(x|female)p(female)$ output *male* else output *female*.
- If priors are equal (i.e. $p(male) = p(female)$
  - $p(x|male) > p(x|female)$ output *male* else output *female*.



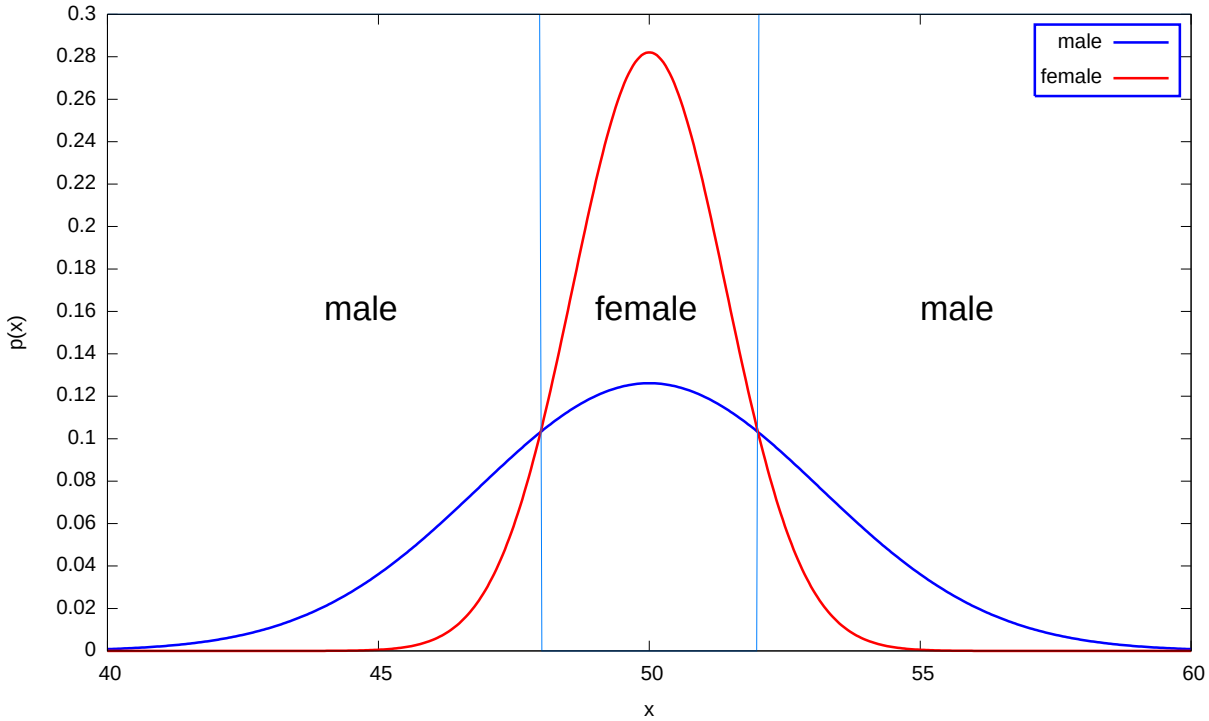
Bayes' Classification Rule

- If $p(male|x) > p(female|x)$ output *male* else output *female*.
- $p(x|male)p(male) > p(x|female)p(female)$ output *male* else output *female*.
- If priors are equal (i.e. $p(male) = p(female)$
  - $p(x|male) > p(x|female)$ output *male* else output *female*.



Bayes' Classification Rule

- If $p(male|x) > p(female|x)$ output *male* else output *female*.
- $p(x|male)p(male) > p(x|female)p(female)$ output *male* else output *female*.
- If priors are equal (i.e. $p(male) = p(female)$
  - $p(x|male) > p(x|female)$ output *male* else output *female*.



Divergence

- So how might we define the 'separatedness' of two distributions.
- Define,

$$D_{12} = \int_{-\infty}^{+\infty} p(x|\omega_1) \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} dx \tag{1}$$

- Note, this is a sort of average of the log of ratio of the two distributions. It is largest when the distributions are most different.
- D_{12} in not equal to D_{21}. The **divergence**, written d_{12}, is thee symmetrical measure defined as,

$$d_{12} = D_{12} + D_{21} \tag{2}$$

- Properties: i/ $d_{12} \geq 0$; ii/ $d_{12} = 0$ only if and only if $p(x|\omega_1) = p(x|\omega_2)$; iii/ $d_{12} = d_{21}$

Divergence

- So how might we define the 'separatedness' of two distributions.
- Define,

$$D_{12} = \int_{-\infty}^{+\infty} p(x|\omega_1) \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} dx \tag{1}$$

- Note, this is a sort of average of the log of ratio of the two distributions. It is largest when the distributions are most different.
- D_{12} in not equal to D_{21}. The **divergence**, written d_{12}, is thee symmetrical measure defined as,

$$d_{12} = D_{12} + D_{21} \tag{2}$$

- Properties: i/ $d_{12} \geq 0$; ii/ $d_{12} = 0$ only if and only if $p(x|\omega_1) = p(x|\omega_2)$; iii/ $d_{12} = d_{21}$

Divergence

- So how might we define the 'separatedness' of two distributions.
- Define,

$$D_{12} = \int_{-\infty}^{+\infty} p(x|\omega_1) \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} dx \tag{1}$$

- Note, this is a sort of average of the log of ratio of the two distributions. It is largest when the distributions are most different.
- D_{12} in not equal to D_{21}. The **divergence**, written d_{12}, is thee symmetrical measure defined as,

$$d_{12} = D_{12} + D_{21} \tag{2}$$

- Properties: i/ $d_{12} \geq 0$; ii/ $d_{12} = 0$ only if and only if $p(x|\omega_1) = p(x|\omega_2)$; iii/ $d_{12} = d_{21}$

Divergence

- So how might we define the 'separatedness' of two distributions.
- Define,

$$D_{12} = \int_{-\infty}^{+\infty} p(x|\omega_1) \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} dx \tag{1}$$

- Note, this is a sort of average of the log of ratio of the two distributions. It is largest when the distributions are most different.
- D_{12} in not equal to D_{21}. The **divergence**, written d_{12}, is thee symmetrical measure defined as,

$$d_{12} = D_{12} + D_{21} \tag{2}$$

- Properties: i/ $d_{12} \geq 0$; ii/ $d_{12} = 0$ only if and only if $p(x|\omega_1) = p(x|\omega_2)$; iii/ $d_{12} = d_{21}$

Divergence

- So how might we define the 'separatedness' of two distributions.
- Define,

$$D_{12} = \int_{-\infty}^{+\infty} p(x|\omega_1) \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} dx \tag{1}$$

- Note, this is a sort of average of the log of ratio of the two distributions. It is largest when the distributions are most different.
- D_{12} in not equal to D_{21}. The **divergence**, written d_{12}, is thee symmetrical measure defined as,

$$d_{12} = D_{12} + D_{21} \tag{2}$$

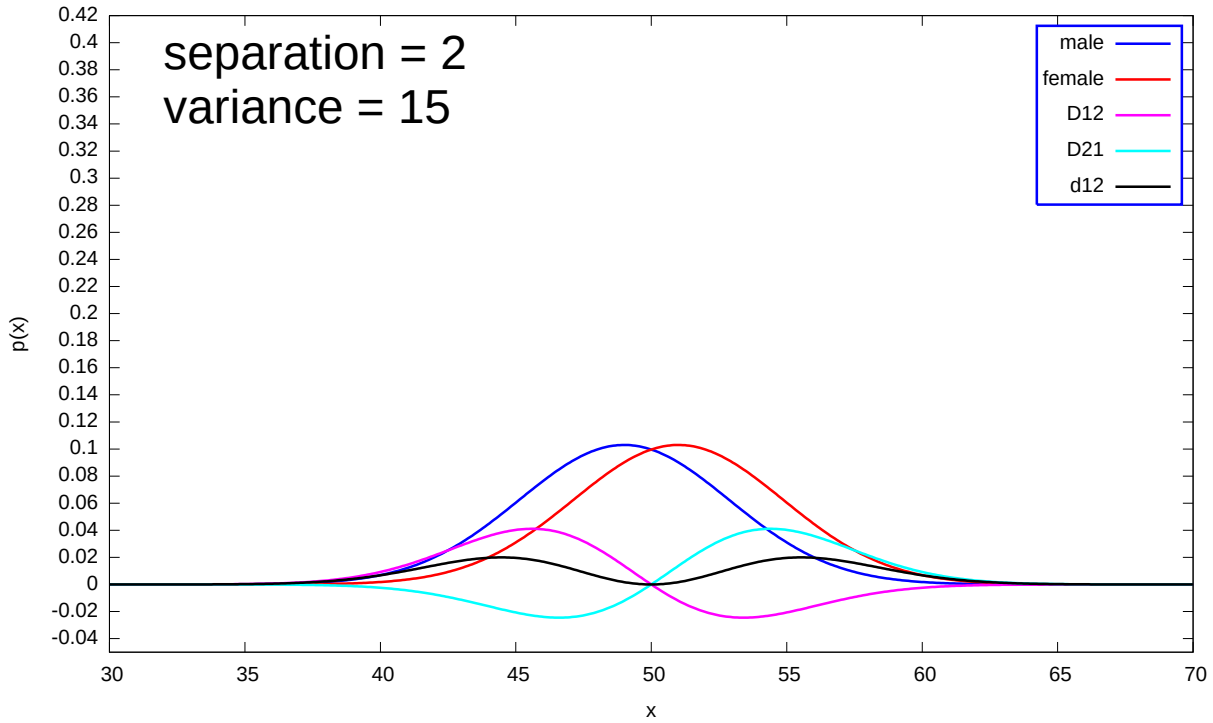
- Properties: i/ $d_{12} \geq 0$; ii/ $d_{12} = 0$ only if and only if $p(x|\omega_1) = p(x|\omega_2)$; iii/ $d_{12} = d_{21}$

Why does this definition work?

$$D_{12} = \int_{-\infty}^{+\infty} p(x|\omega_1) \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} dx \tag{3}$$

Remember,

- $\ln(1) = 0$, so D_{12} is small for x where $p(x|\omega_1)$ and $p(x|\omega_2)$ have similar values.
- Integrating just mean measuring the area under the curve.

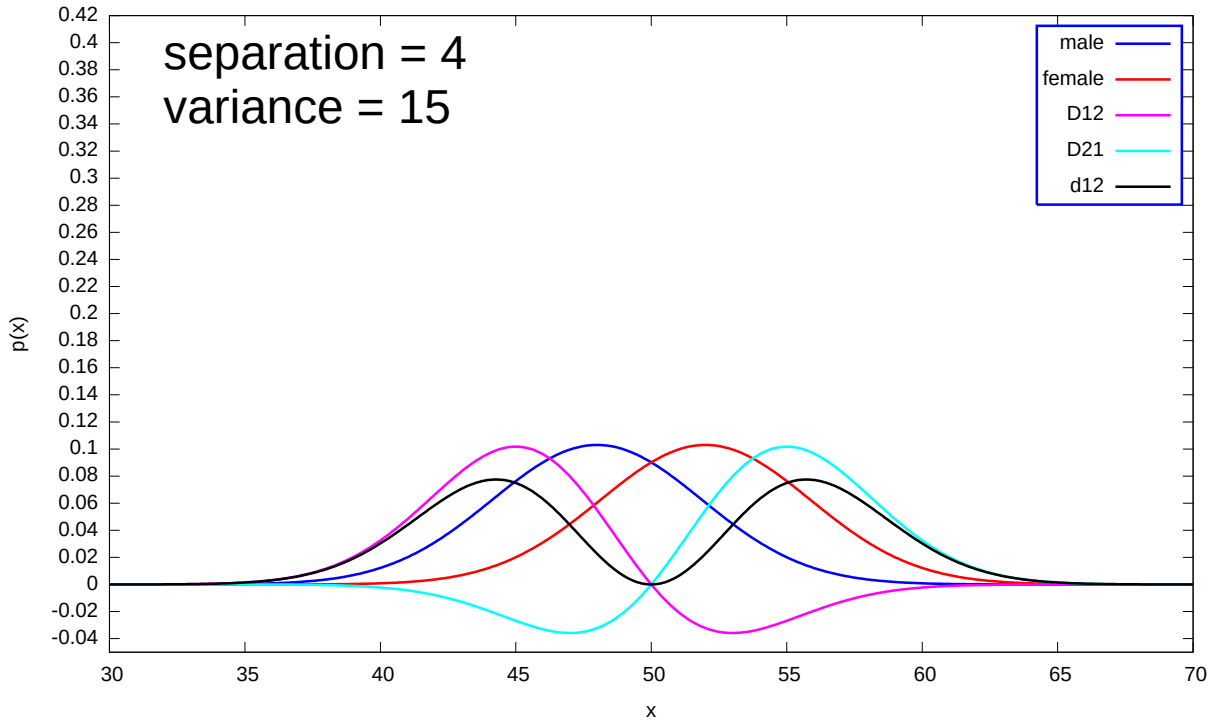


Why does this definition work?

$$D_{12} = \int_{-\infty}^{+\infty} p(x|\omega_1) \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} dx \tag{3}$$

Remember,

- $\ln(1) = 0$, so D_{12} is small for x where $p(x|\omega_1)$ and $p(x|\omega_2)$ have similar values.
- Integrating just mean measuring the area under the curve.

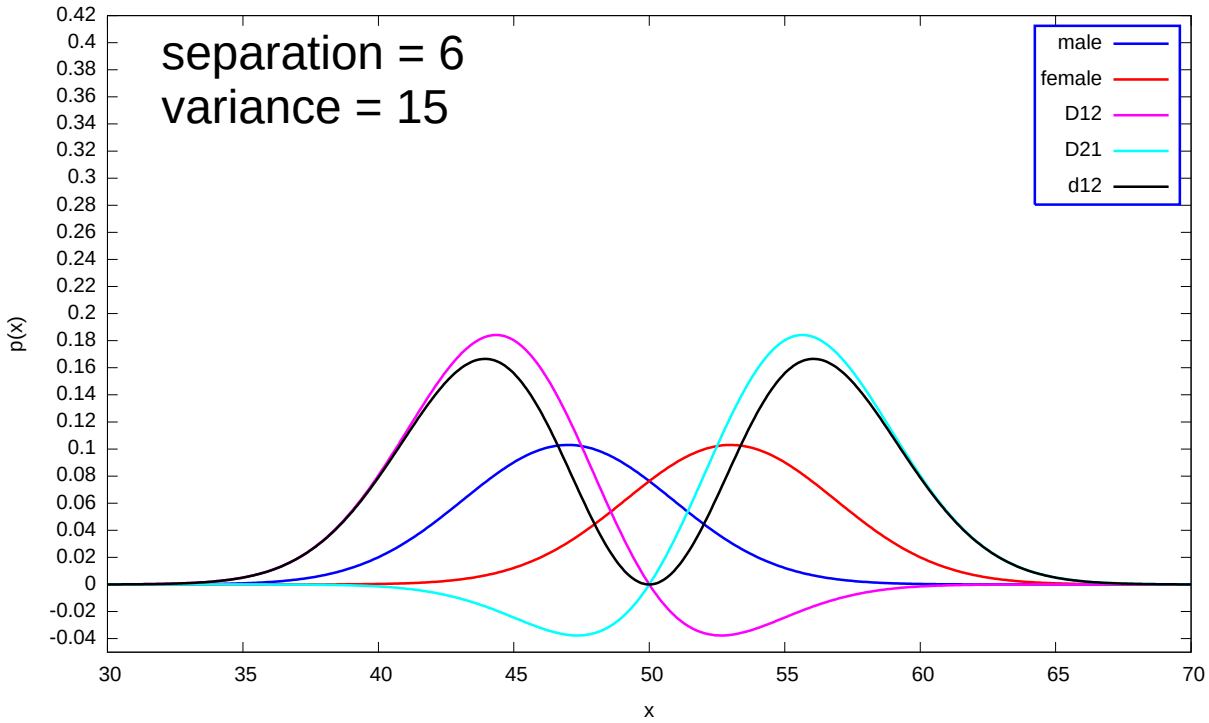


Why does this definition work?

$$D_{12} = \int_{-\infty}^{+\infty} p(x|\omega_1) \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} dx \tag{3}$$

Remember,

- $\ln(1) = 0$, so D_{12} is small for x where $p(x|\omega_1)$ and $p(x|\omega_2)$ have similar values.
- Integrating just mean measuring the area under the curve.

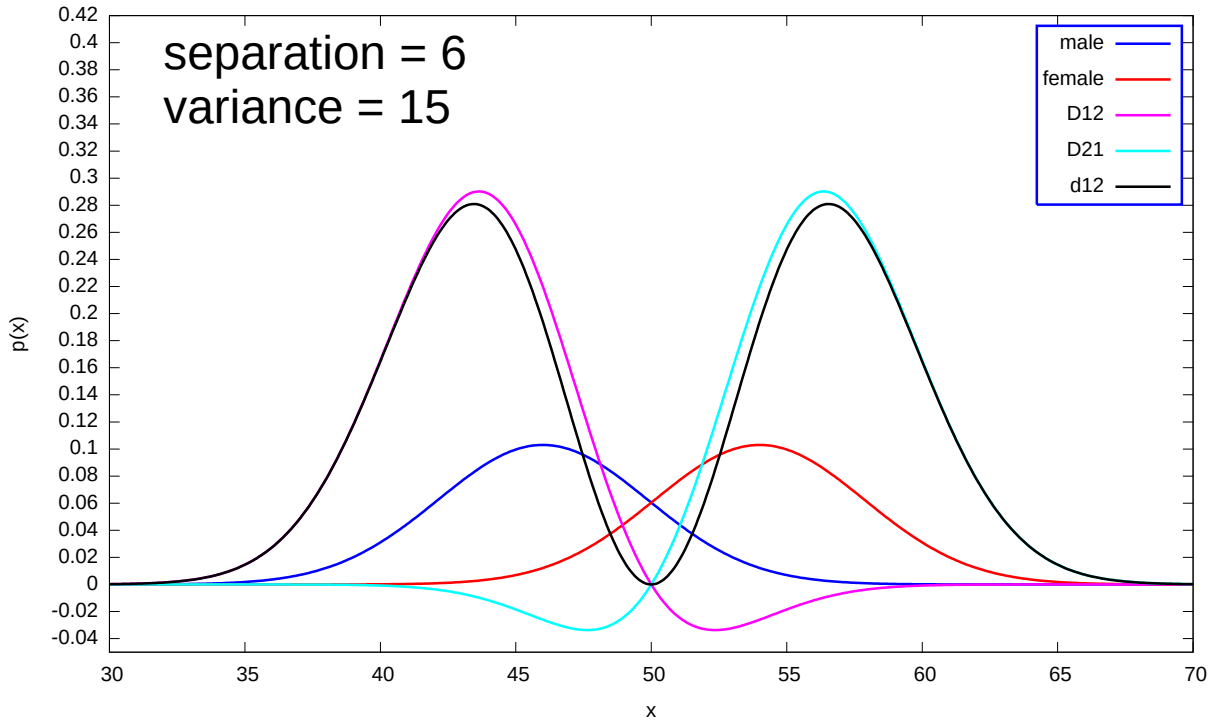


Why does this definition work?

$$D_{12} = \int_{-\infty}^{+\infty} p(x|\omega_1) \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} dx \tag{3}$$

Remember,

- $\ln(1) = 0$, so D_{12} is small for x where $p(x|\omega_1)$ and $p(x|\omega_2)$ have similar values.
- Integrating just mean measuring the area under the curve.

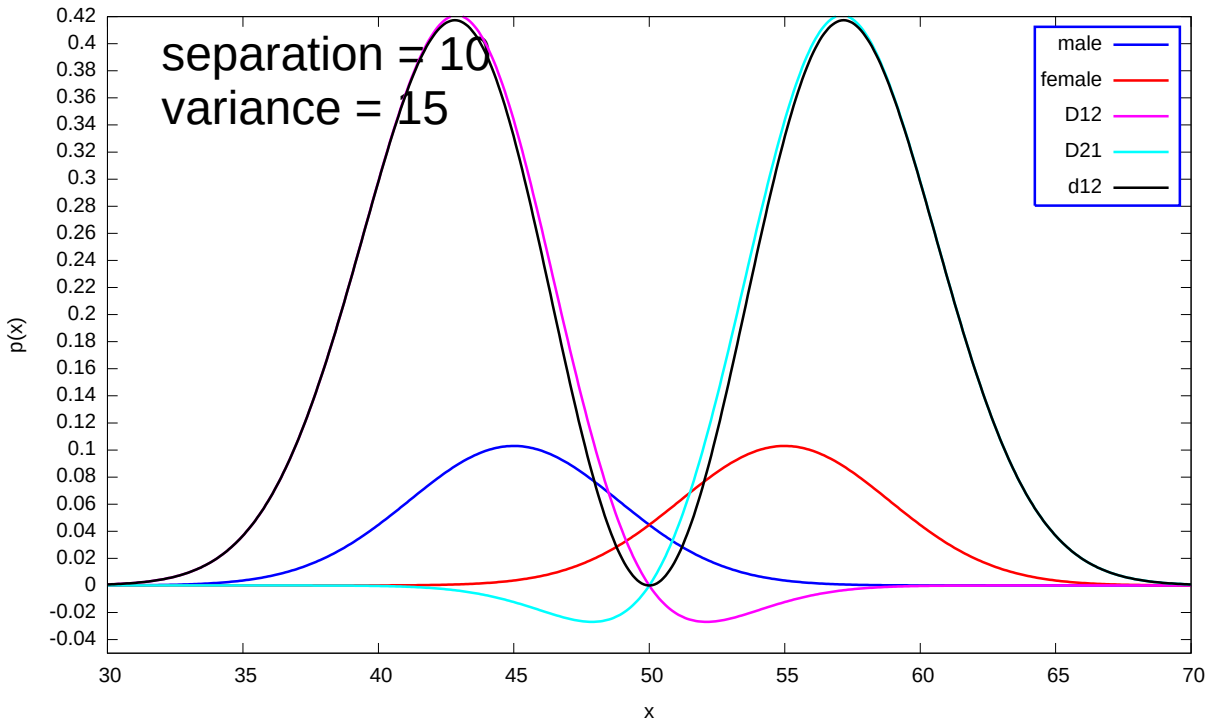


Why does this definition work?

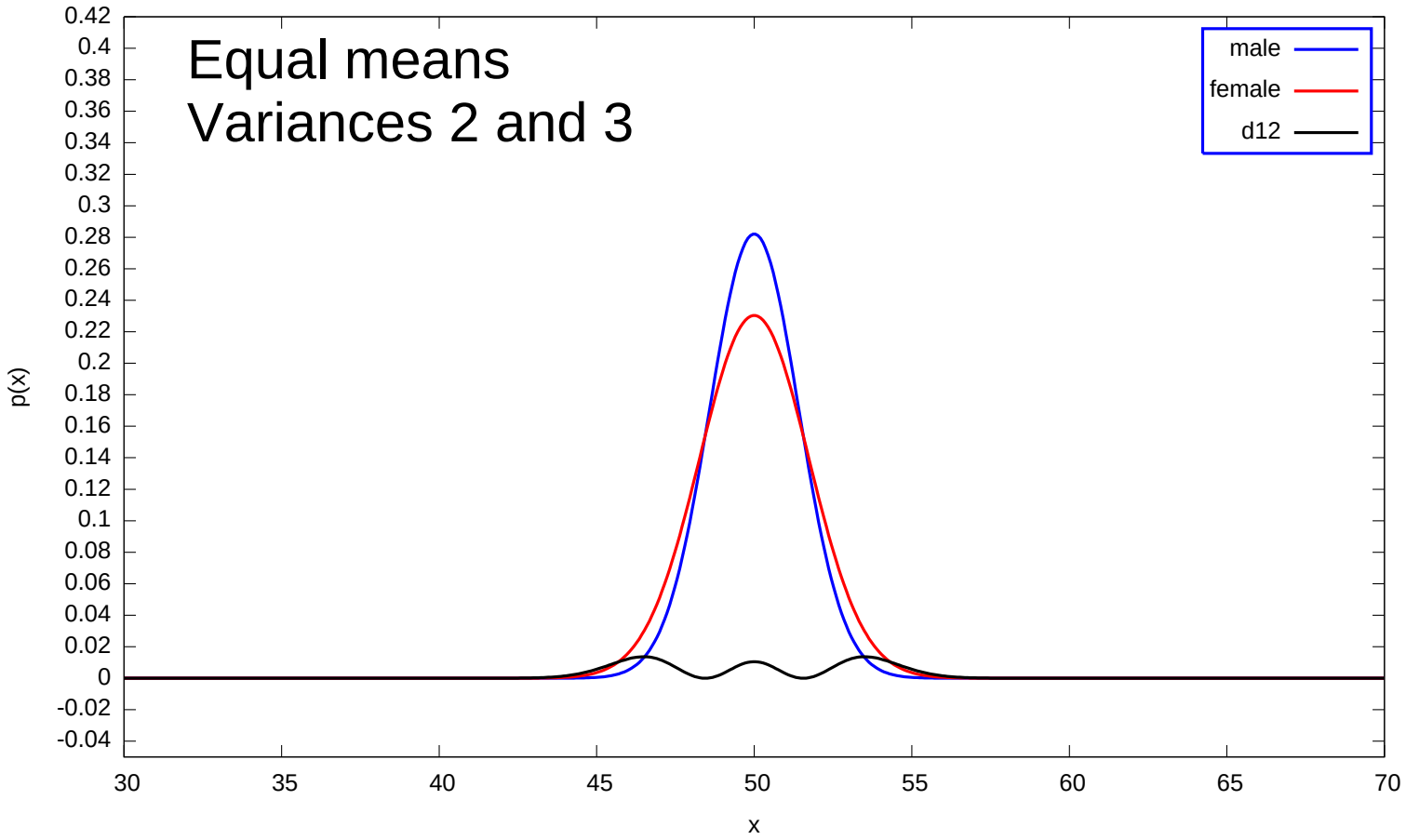
$$D_{12} = \int_{-\infty}^{+\infty} p(x|\omega_1) \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} dx \tag{3}$$

Remember,

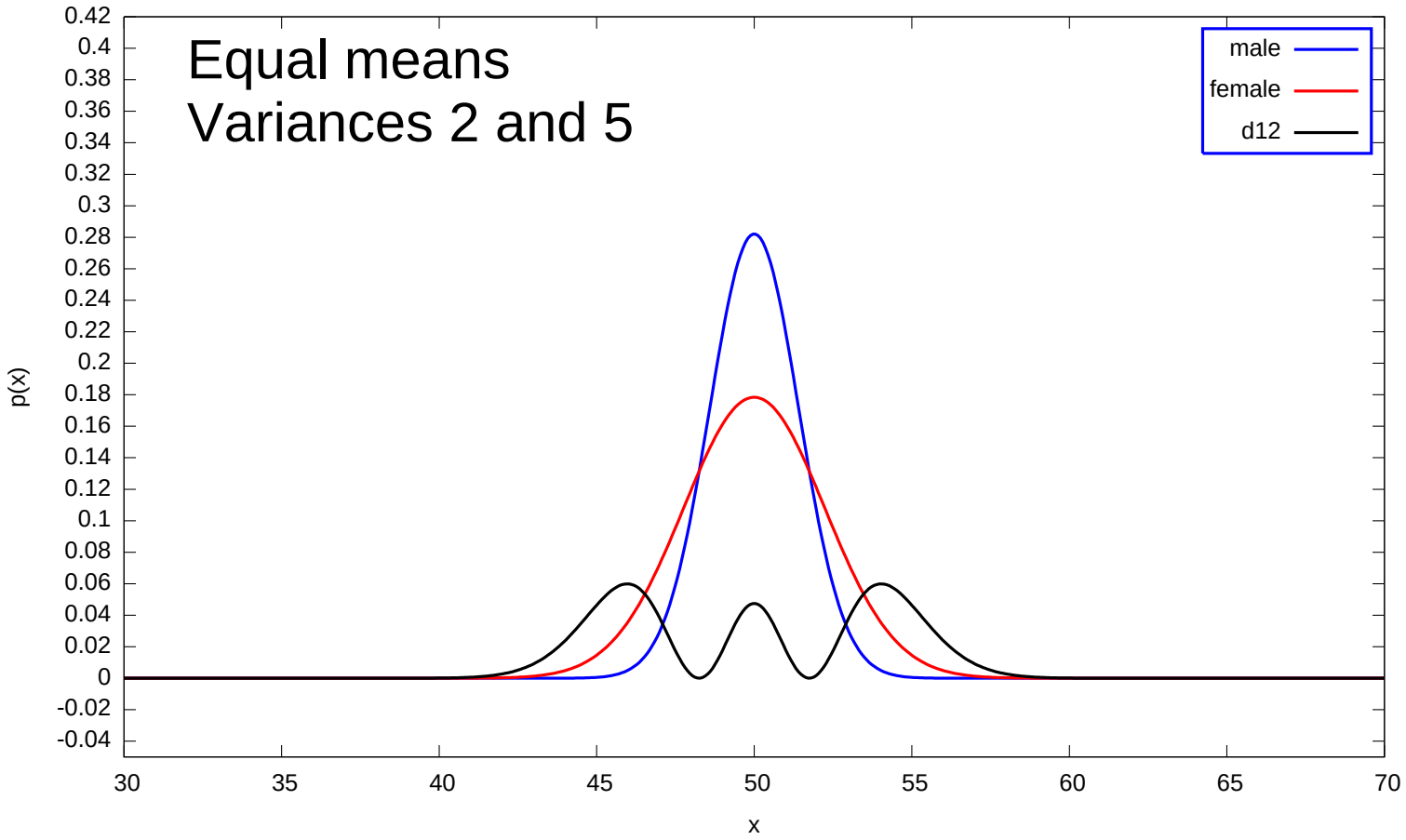
- $\ln(1) = 0$, so D_{12} is small for x where $p(x|\omega_1)$ and $p(x|\omega_2)$ have similar values.
- Integrating just mean measuring the area under the curve.



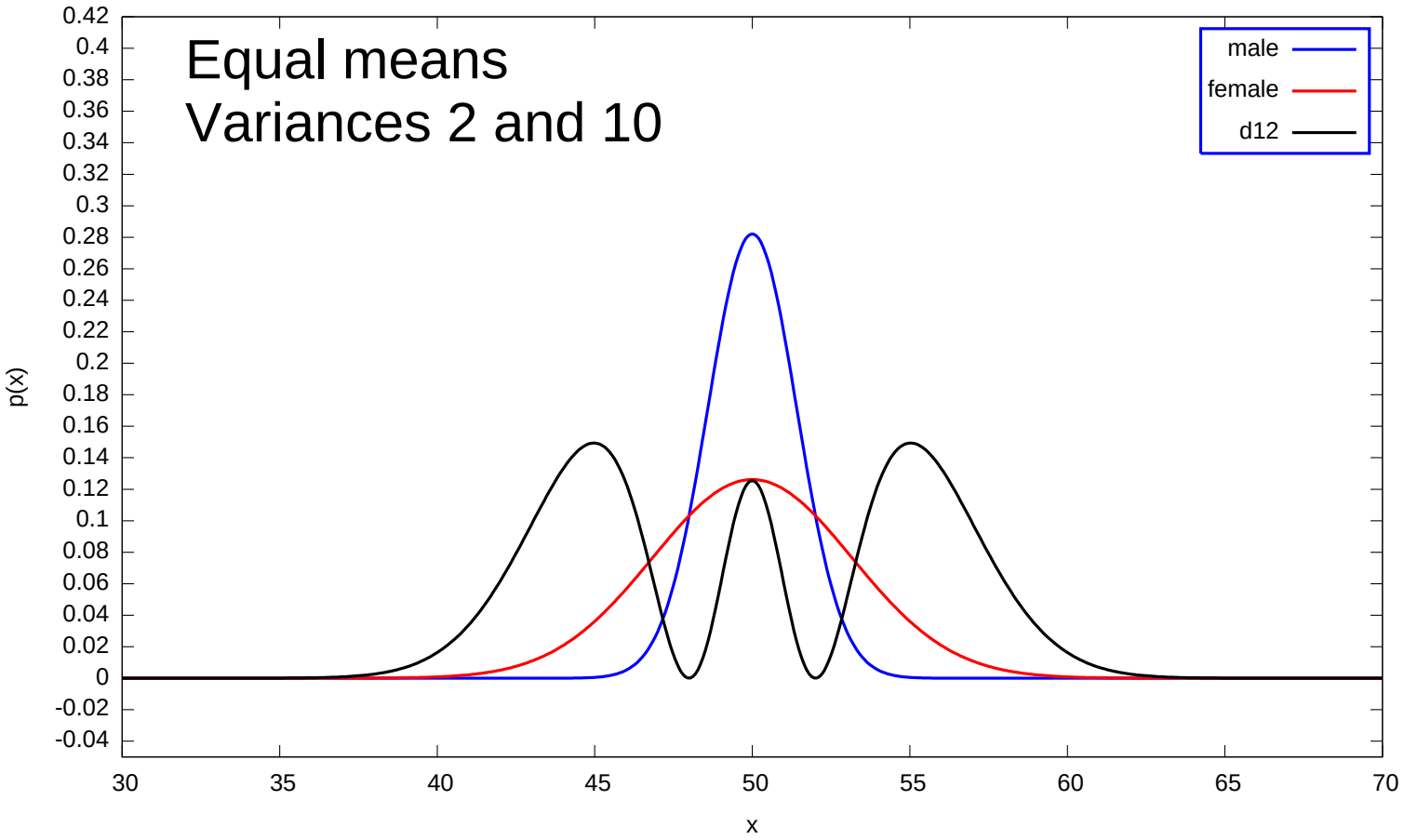
Divergence: The equal means case



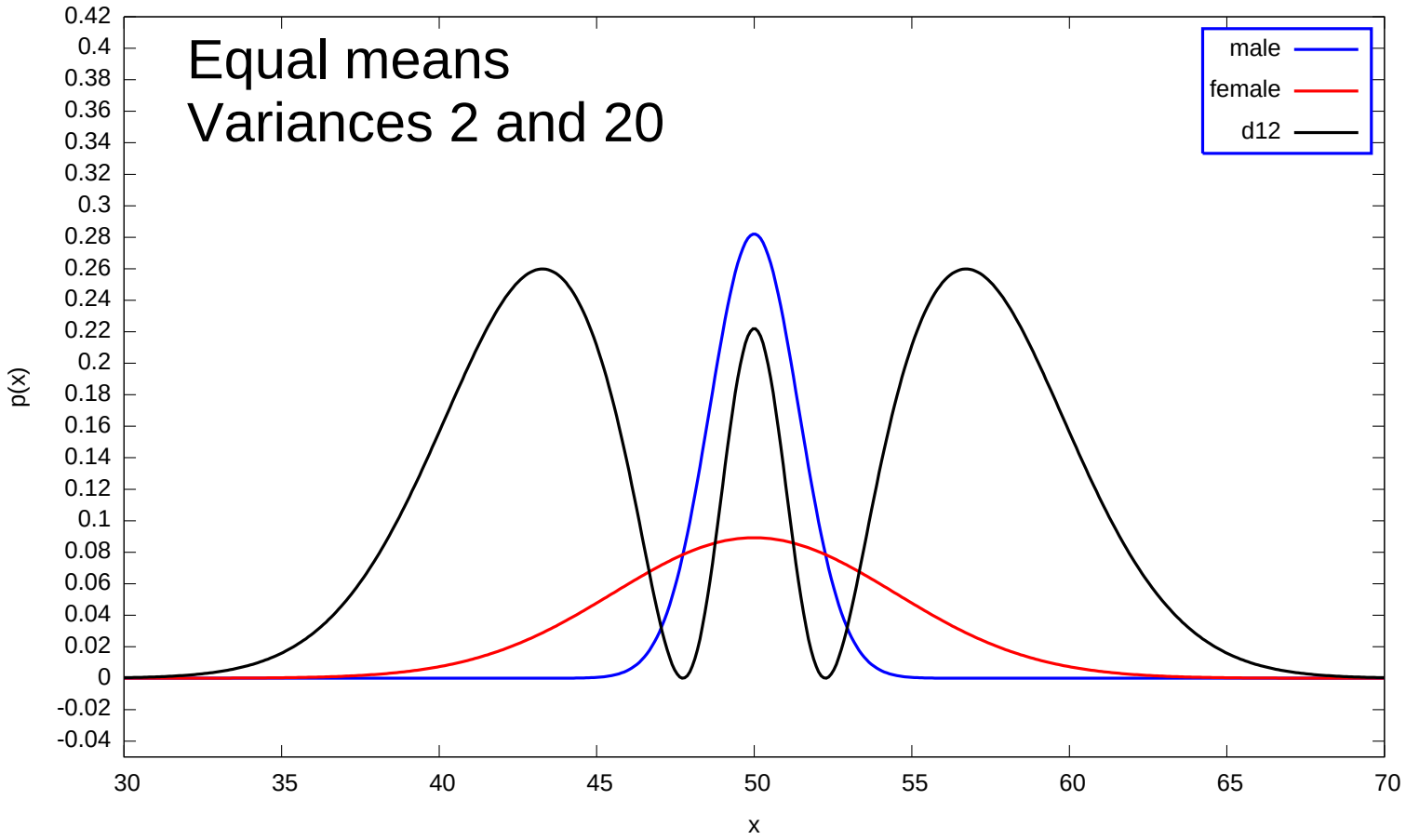
Divergence: The equal means case



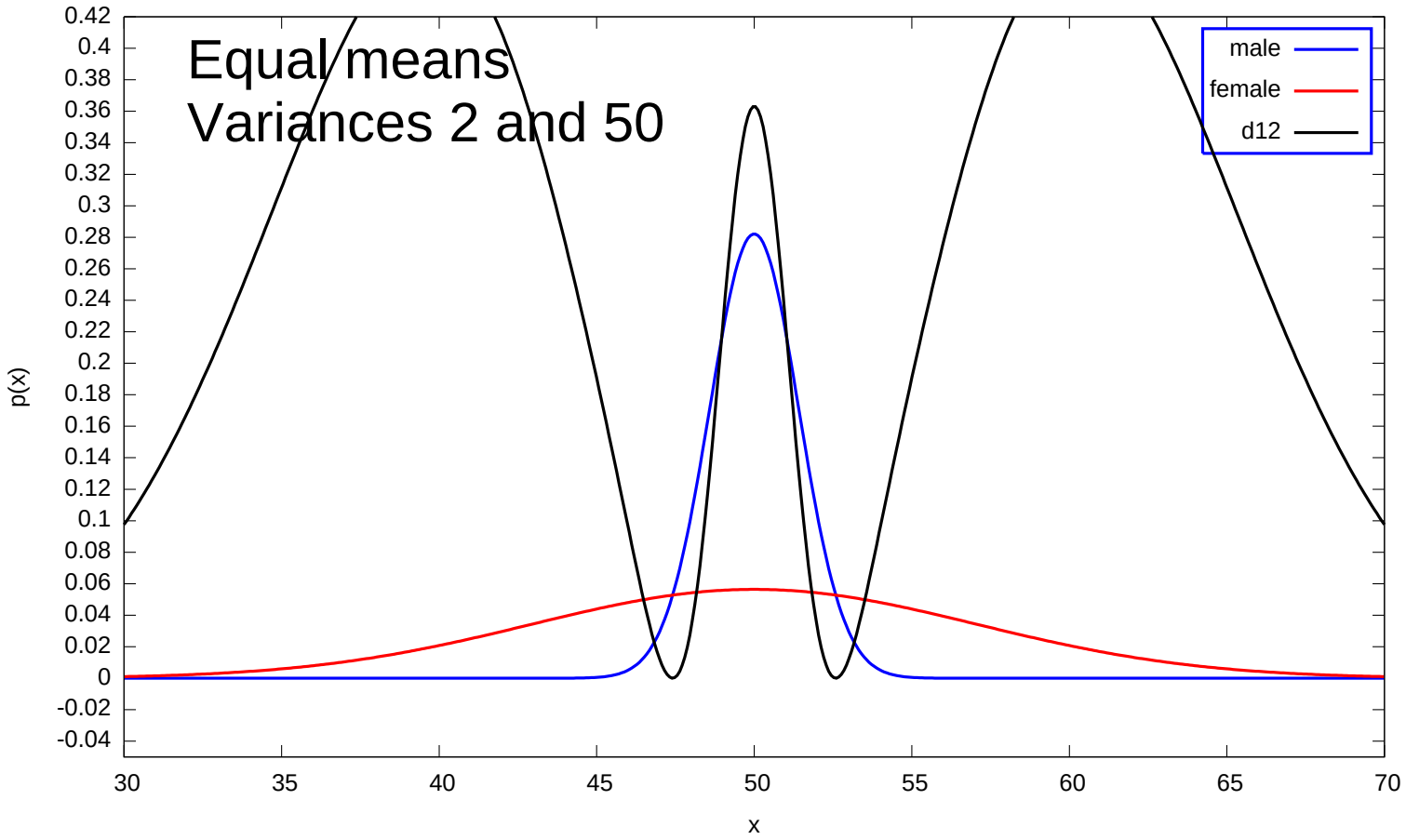
Divergence: The equal means case



Divergence: The equal means case



Divergence: The equal means case



Divergence for one-dimensional Gaussians

For 1-d Gaussians the equation for divergence becomes,

$$d_{12} = \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} - 2\right) + \frac{1}{2}(\mu_1 - \mu_2)^2\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)$$

where μ_1 and μ_2 are the means of the distributions, and σ_1 and σ_2 are the standard deviations[1].

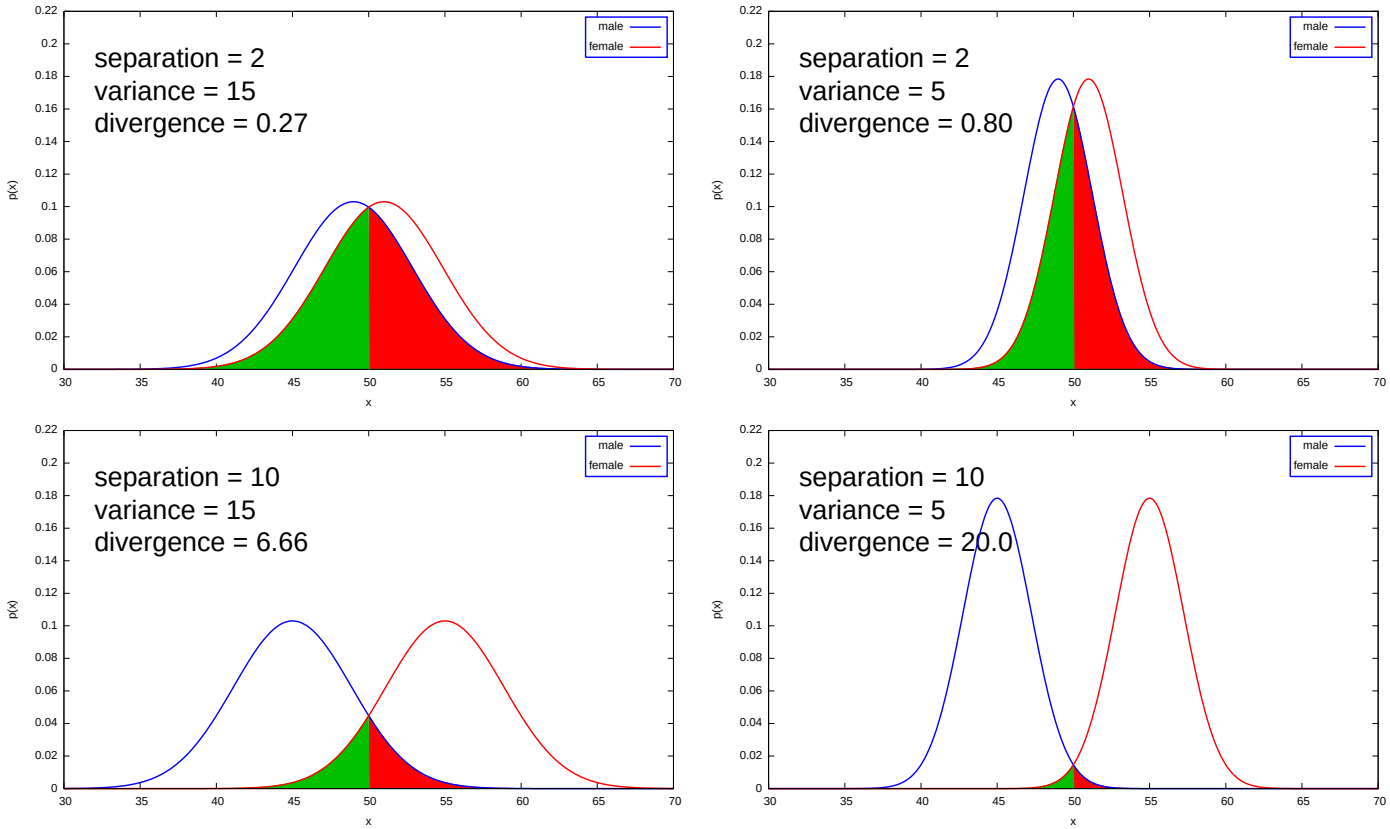
Note the following behaviours of the equation,

- d_{12} increases as means become more separated.
- d_{12} increases as variance become smaller.
- d_{12} increases as variances become more unequal.

[1] Remember, the standard deviation is just the square root of the variance

Divergence

Divergence for pairs of Gaussian distributions:



Note, divergence increases as the distributions become more greatly separated.

Selecting features

- Consider again our gender features.
- Say we have computed the divergence between male and female distributions for each feature:

left arm length	2.5
right arm length	2.4
hair length	1.8
weight	1.3
shoe size	1.2
eye colour	0.01[1]

- If using one feature, best feature is '**left arm length**'
- If using two features, best features would be '**left arm length**' and '**right arm length**'... Right?
- **No!**, not necessarily... we need to consider their joint distribution.

[1] These figures are made up for sake of illustration and are not intended to be realistic!

Selecting features

- Consider again our gender features.
- Say we have computed the divergence between male and female distributions for each feature:

left arm length	2.5
right arm length	2.4
hair length	1.8
weight	1.3
shoe size	1.2
eye colour	0.01[1]

- If using one feature, best feature is '**left arm length**'
- If using two features, best features would be '**left arm length**' and '**right arm length**'... Right?
- **No!**, not necessarily... we need to consider their joint distribution.

[1] These figures are made up for sake of illustration and are not intended to be realistic!

Selecting features

- Consider again our gender features.
- Say we have computed the divergence between male and female distributions for each feature:

left arm length	2.5
right arm length	2.4
hair length	1.8
weight	1.3
shoe size	1.2
eye colour	0.01[1]

- If using one feature, best feature is '**left arm length**'
- If using two features, best features would be '**left arm length**' and '**right arm length**'. . . Right?
- **No!**, not necessarily. . . we need to consider their joint distribution.

[1] These figures are made up for sake of illustration and are not intended to be realistic!

Selecting features

- Consider again our gender features.
- Say we have computed the divergence between male and female distributions for each feature:

left arm length	2.5
right arm length	2.4
hair length	1.8
weight	1.3
shoe size	1.2
eye colour	0.01[1]

- If using one feature, best feature is '**left arm length**'
- If using two features, best features would be '**left arm length**' and '**right arm length**'. . . Right?
- **No!**, not necessarily. . . we need to consider their joint distribution.

[1] These figures are made up for sake of illustration and are not intended to be realistic!

Summary

- Classification lies at the heart of many computer applications (character recognition, speech recognition, face detection, fingerprint matching, etc, etc, etc).
- Hardest task is finding suitable features.
- We want to find features that ‘separate’ the classes.
- Divergence is a useful measure of separatedness of two distributions.
- We can’t generally select features by considering each feature separately.
- Next lecture we’ll look at how to select multiple features.