Classification Methods part 1

Module #3

THE UNIVERSITY of ADELAIDE

Naïve Bayes for Classification

THE UNIVERSITY of ADELAIDE

Review from Online Learning:

probabilities

- Which of the following can be considered a probability distribution? E.g. probability distribution of word ***w*** in 5 documents.
  - A: [11, 22, 3, 4, 8]
  - B: [0.1, 0.5, 0.8, 0.6, 0.7]
  - C: [0.05, 0.2, 0.15, 0.1, 0.5]

Naïve Bayes explained

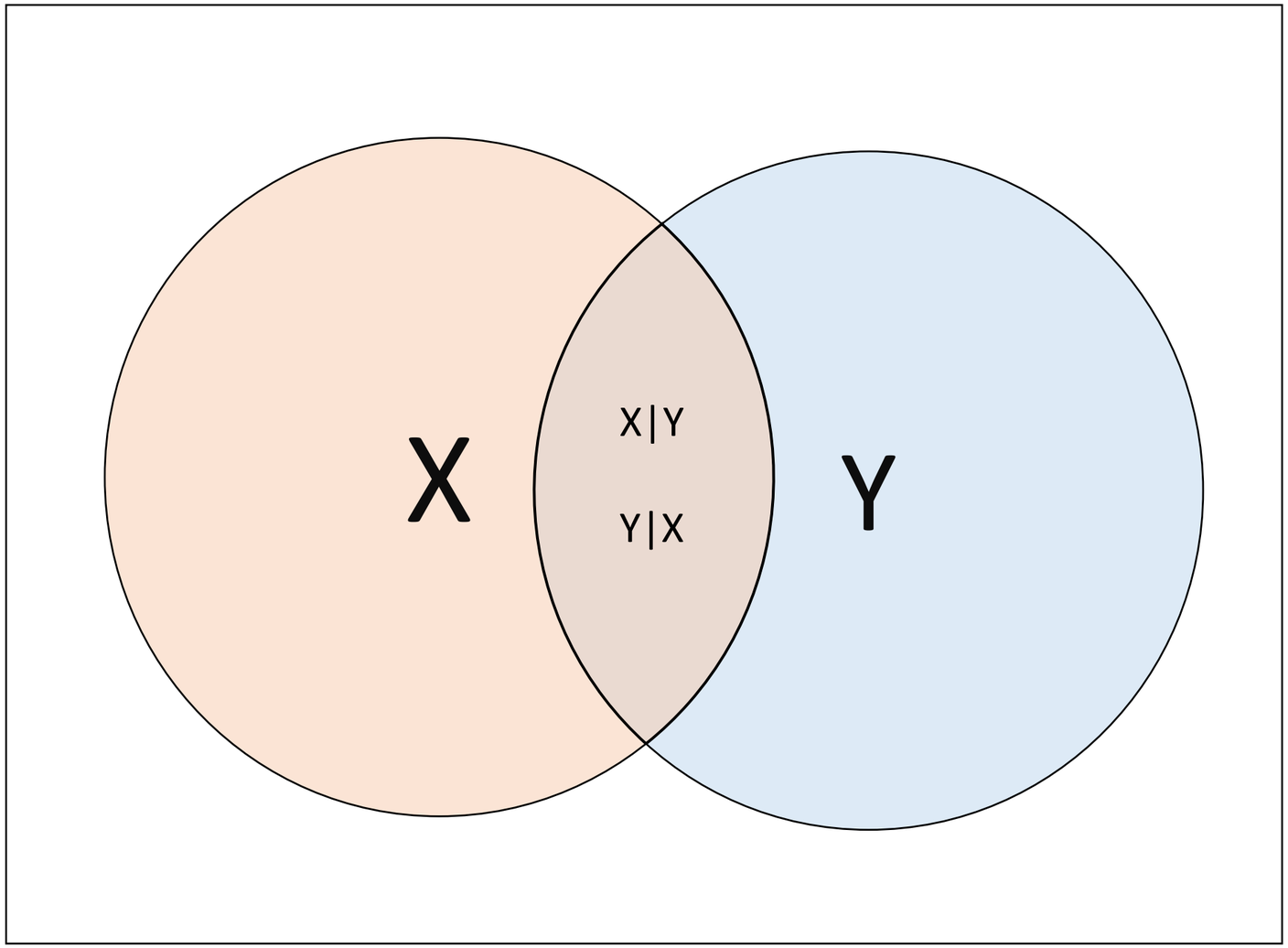
You can skip this part if you understand Naïve Bayes classification well, just go to the summary part and example.

Probability… not statistics

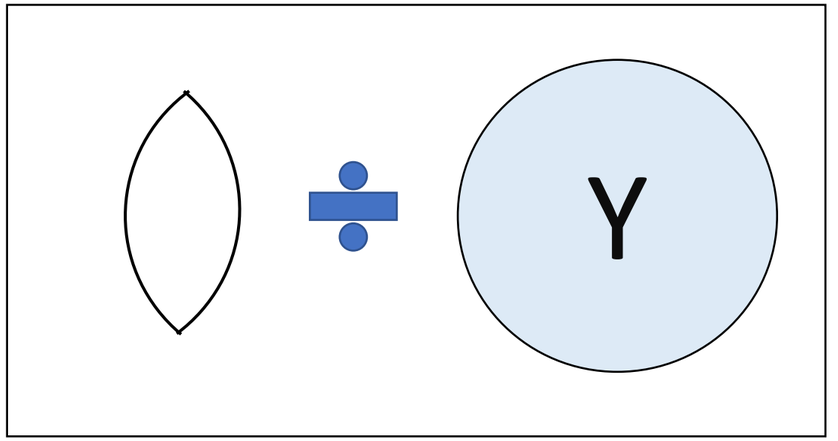
Bayes Theorem

$$P(Y|X) = \frac{P(Y) * P(X|Y)}{P(X)}$$

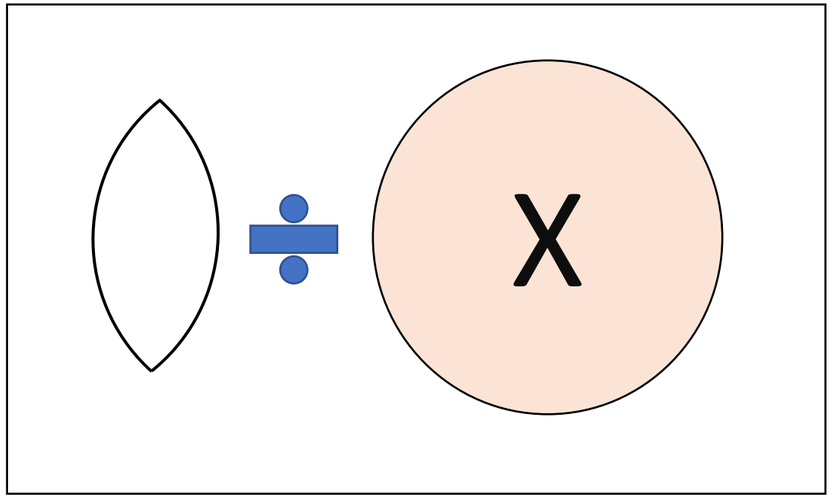
What does this mean?



X|Y



Y|X



Geometry would indicate

$$P(Y|X) * P(X) = P(Y) * P(X|Y)$$

With re-arrangement

$$P(Y|X) = \frac{P(Y) * P(X|Y)}{P(X)}$$

Linguistically

$$P(Y|X) = \frac{P(Y) * P(X|Y)}{P(X)}$$

We're trying to work out, what is the probability of the result (Y) occurring, given the indicator (X) occurred.

So we need to know...

- P(X). How likely is X to occur? (We can gather this easily enough)
- P(X|Y). If we know the outcome, how likely was the indicator?
- P(Y). How likely is the outcome (Y)… wait what?

A mental example

There is a person.

They have ‘the complete works of William Shakespeare’ on their shelf.

Are they more likely to be:

A) A farmer

B) A playwright

What do we know: The Prior

Well… there are like… at least 90 000 farmers in Australia.

And 92 playwrights on Wikipedia…

So if you take an average person from these two groups…

You have a 0.1 % chance that they are a playwright.

P(Y=playwright) = 1/1000

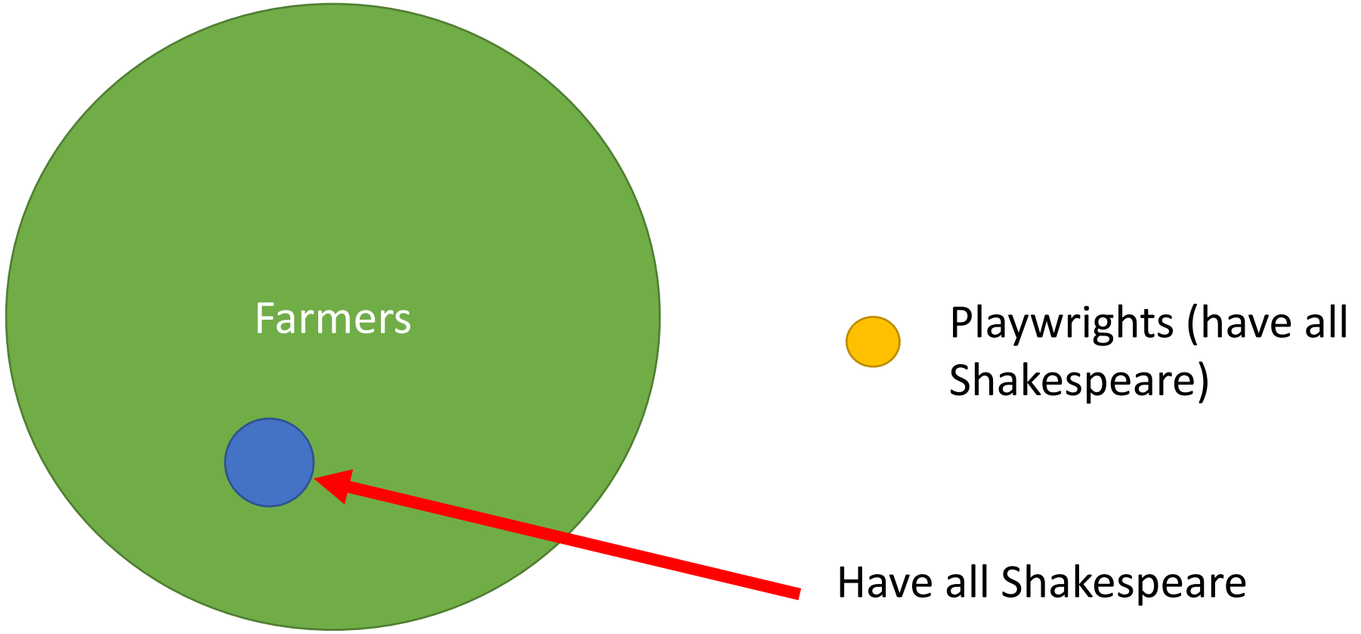
P(Y=farmer) = 999/1000

$$P(Y|X) = \frac{\boldsymbol{P(Y)} * P(X|Y)}{P(X)}$$

An intuition

Intuition

If 93 farmers (i.e. 1 in 1000 farmers) have the complete works of William Shakespeare it is more likely than not, that the ‘random person’, who has all Shakespeare, is actually a farmer, not a playwright. After all, 93 > 92.



What we also might know? The likelihood.

P(X|Y)… also known as, what is the probability that the person has ‘the complete works of William Shakespeare’ GIVEN they are a farmer. OR GIVEN they are a playwright.

Let’s assume that 100% of playwrights, have TCWOWS

P(TCWOWS|Playwright)= 100%

Given that P(Playwright) = 0.1% AND P(Farmer) = 99.9%

We can work out the relevant ‘switching point’.

Bayes in Action

Likelihood(F|TCWOWS) = P(F) * P(TCWOWS|F)

= 0.999 * 0.001

Likelihood(PW|TCWOWS) = P(PW) * P(TCWOWS|PW)

= 0.001 * 1.0

Where is the denominator?

Often we don't really need it…

It is always the same

Multiple Factors

How to calculate if there are multiple factors?

$$P(Pass|L, A) = P(P) * P(A|Pass) * P(L|Pass)/P(X)$$

Probability of passing based on doing all assignments and attending all lectures…

The easiest way to work out P(X) is just to sum all the top lines to 1 and never bother to calculate it.

Limitations of Bayes

No data…

(How is this a weakness, nothing works when you have no data…)

It does stuff up the maths though.

P = 0 isn't good for math equations where you are multiplying… say P_0 x P_1 x P_2.

So we have a correction factor to set a kind of minimum probability.

What about continuous ranges?

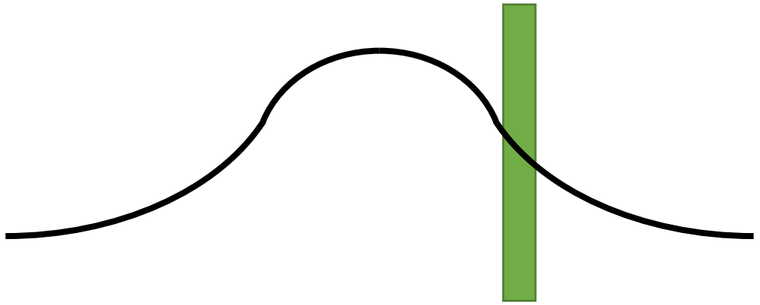
Probability doesn't work well. Need buckets.

$$P(x|y) = \frac{\gamma + n_{yx}}{\gamma \cdot |dom(X)| + n_y}$$

Overcoming Limitations: Continuous Variables

If you have a continuous variable, you can convert it to a normal distribution (or other distribution).

How?



1. Convert the set of values into a mean and standard deviation, work out the distribution.
2. Work out where the new value lies in that distribution.
3. Profit! (I.e., a probability)

This is a lie… it is just another kind of bucket. The probability that a continuous value is some constant is always… zero (but between a range, that is a useful number)

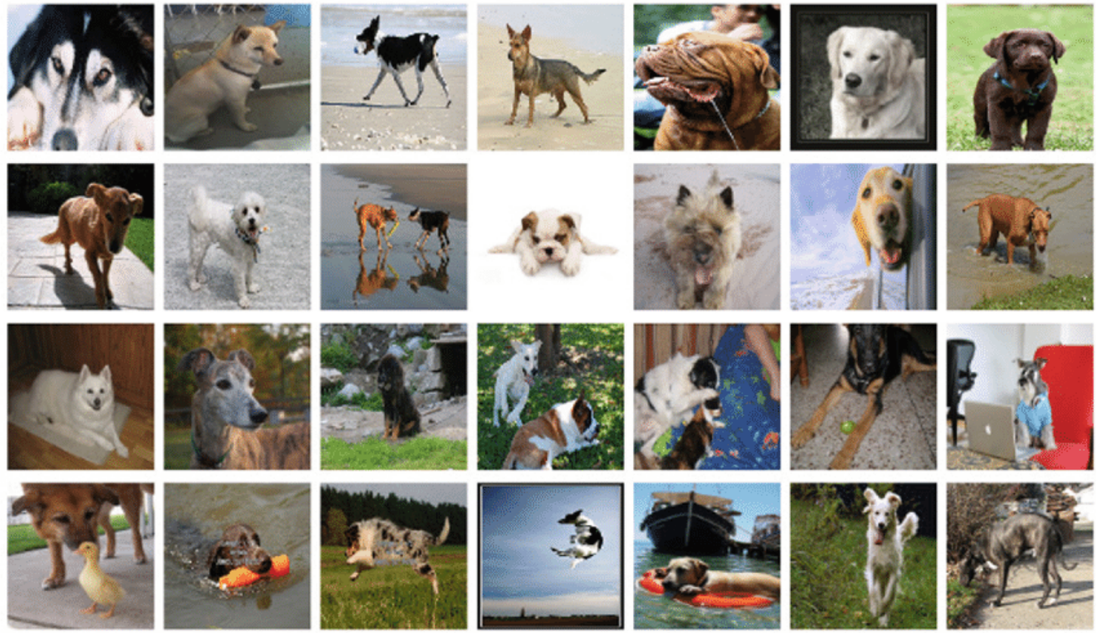
Maximum Likelihood

Suppose we are trying to determine which of a set of different classes, a single instance belongs to.

$$P(Y|X) = \frac{P(Y) * P(X|Y)}{P(X)}$$

Becomes:

$$y_{ML} = \underset{y \in Y}{\operatorname{argmax}} P(X|y)$$



Maximum Likelihood

Wait what?

How did we get there.

$$P(Y|X) = \frac{\cancel{P(Y)} * P(X|Y)}{\cancel{P(X)}}$$

We are comparing…

$P(X|y_1)$, $P(X|y_2)$….

#1: P(X) is always the same

#2: If the P(Y) is always the same (i.e., a balanced case where all the classes are equi-probable)…

Naïve Bayes

What are we doing?

- We are just applying Bayes theorem to a large dataset where we have ‘some prior’ as well as ‘some likelihood’ training data from which we can derive the useful values.
- We are also assuming independence. This is where the **Naïve** comes from. Lots of things are not actually independent.

After that it is just maths.

A good question here:

What would happen if we didn’t assume independence? How would that change our math? Would it break everything?

Naïve Bayes in summary and example

Just make sure you can make solve yourself a problem similar to the one given in the example.

Bayes Theorem for Classification

- Bayes theorem can be used to classify instances

$$\mathcal{D} = \{(\boldsymbol{X}_i, y_i) | i = 1,2, \dots, n\}$$

- Data of n instances, m classes *X* is vector of features, y_i is output class: $y_i \in (Y: \{y_1, \dots, y_j, \dots, y_m\})$

$$P(h|X) = \frac{P(X|h) \cdot P(h)}{P(X)} \quad h: y = y_j \text{ outcome to predict}$$

- We want *h* that maximises $P(h|X)$, maximum a posteriori hypothesis (MAP)

$$h_{MAP} = \arg\max_{h \in H} P(h|X) = \arg\max_{h \in H} \frac{P(X|h) \cdot P(h)}{P(X)} = \arg\max_{h \in H} P(X|h) \cdot P(h)$$

Bayes Theorem for Classification cont.

If we know or assume that all hypotheses *h* are equally probable, then we get the maximum likelihood hypothesis (ML)

$$h_{ML} = \arg\max_{h \in H} P(X|h)$$

In order to see how Bayes theorem can be used for classification, let's break down and simplify the formula. Notice that *P(X)* is independent of h, so it was omitted in MAP and ML formulas.

P(h) (prior)for each hypothesis (each class) can be calculated as $\frac{n_h}{n}$, where n_h is number of instances of training set classified into class y_j, and n is the number of instances in training set (which we use to estimate the probabilities)

From Bayes Theorem to Naïve Bayes

Now let's consider $P(X|h)$. In general, $P(X|h)$ is calculated using chain rule:

$P(x_1 \wedge x_2 \wedge \cdots \wedge x_m) = \prod_{i=1}^{m} P(x_i | \bigcap_{j=1}^{i-1} x_j)$.

In order to calculate this, we would need all possible combinations of X, which we do not have. In Naïve Bayes, we assume that x_i occur independently, so the formula is simplified:

$$P(x_1 \wedge x_2 \wedge \cdots \wedge x_m) = P(x_1) * P(x_2) * \cdots * P(x_m)$$

Therefore:

$$P(D = (x_1, \ldots, x_m|h) = P(x_1|h) \cdot \cdots \cdot P(x_m|h) = \prod_{x_i \in D} P(x_i|h)$$

Although the independence assumption is often wrong, the Bayes classifier yields good results.

How to Classify Using Naïve Bayes

1. **Using training dataset:**
   a) For each class, calculate *P(h)* as $\frac{n_h}{n}$, where *n* is number of instances, n_h is number of instances in class *y* for hypothesis *h*
   b) For each feature in training set, calculate $P(x_j|h)$ as $\frac{n_j}{n_h}$, where n_h is number of instances in class *h*, n_j is number of instances in class *h* with feature *j*
   c) Store the above calculations for later
2. **Given test instance with features :** $\{x_1, \dots, x_j, \dots\}$ **and class** y_k:
   a) Retrieve all *P(h)* and all $P(x_j|h)$ for these features
   b) Calculate $P(X|h) \cdot P(h) = P(h) * P(x_1|h) * \cdots * P(x_j|h) * \cdots$ for all features *x* in the test instance, and all hypotheses *h*, corresponding to all classes. So we will have a vector of *m* values, one for each class
   c) Take arg max of that vector, as the predicted class

Using Naïve Bayes: Example

outlook	temperature	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

We have Play Tennis data on the left

And we have new instance: $\langle Outlk = sun, Temp = cool, Humid = high, Wind = true \rangle$

Notice that we only need P(play), and counts for 4 attributes, so:

P(play=yes) = 9/14, P(play=no) = 5/14

P(sun | y=2/9, P(cool | y)=3/9, P(high | y)=3/9, P(true | y)=3/9

P(sun | n=3/5, P(cool | n)=1/5, P(high | n)=4/5, P(true | n)=3/5

Therefore

arg max(9/14*2/9*3/9*3/9*3/9,
5/14*3/5*1/5*4/5*3/5) =

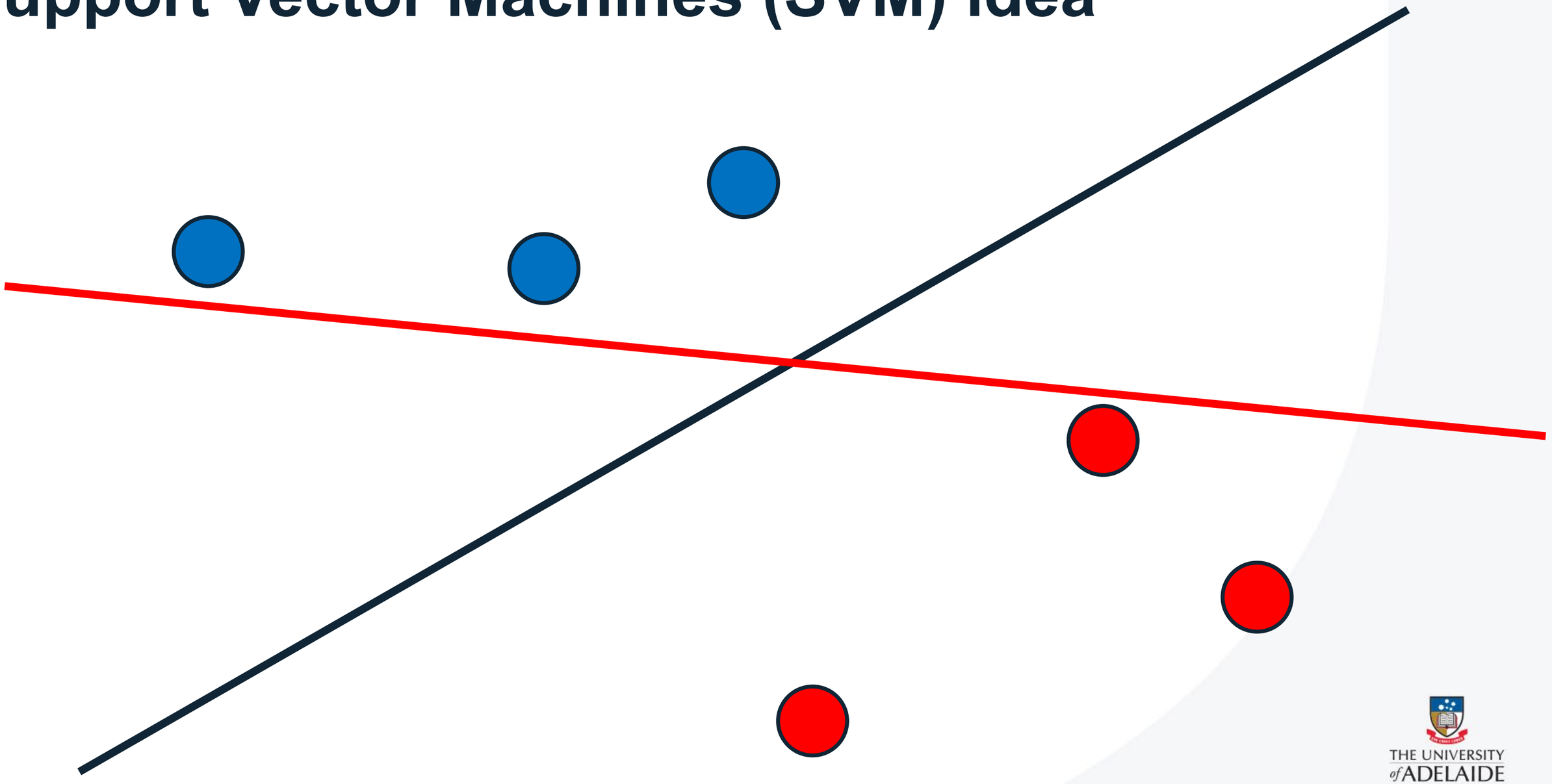
arg max(y:0.0053,n:0,0206)=no

THE UNIVERSITY of ADELAIDE

Support Vector Machines

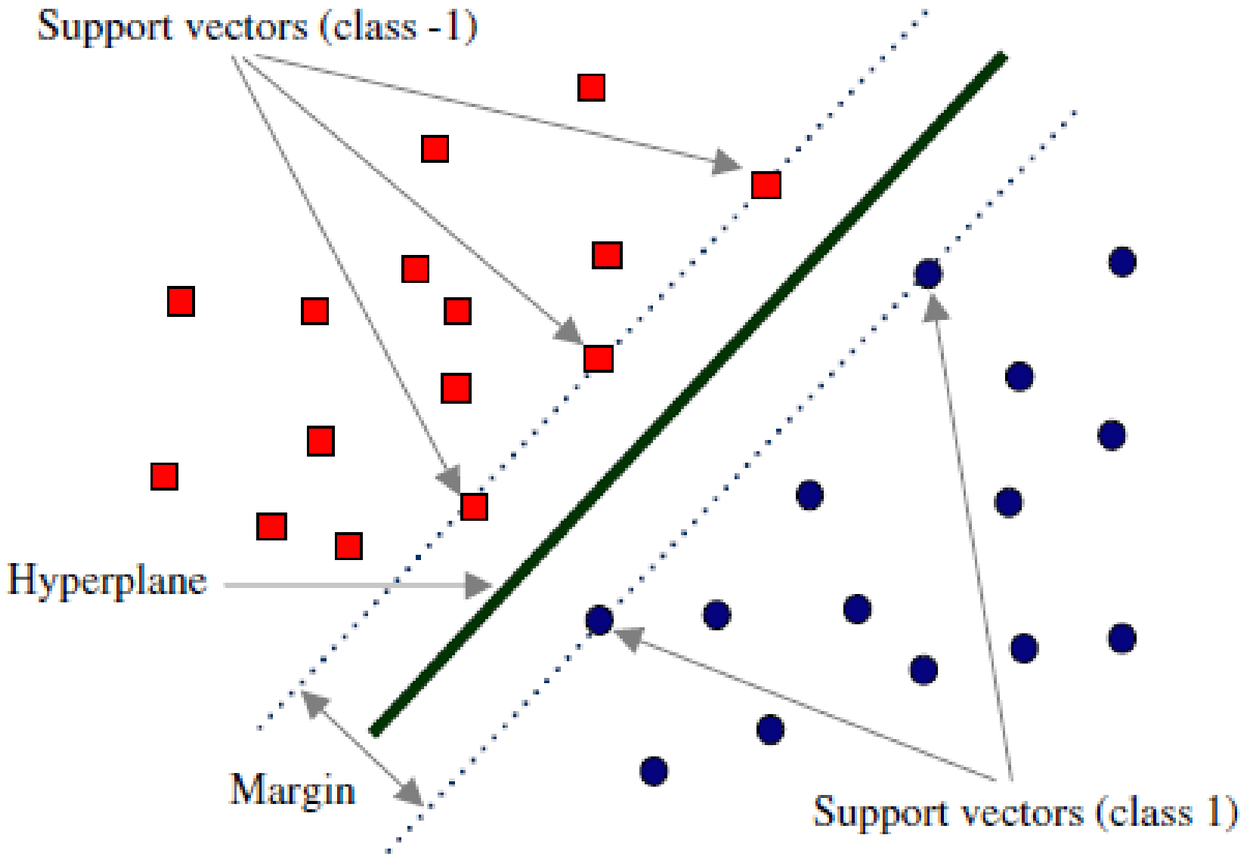
THE UNIVERSITY of ADELAIDE

Support Vector Machines (SVM) idea



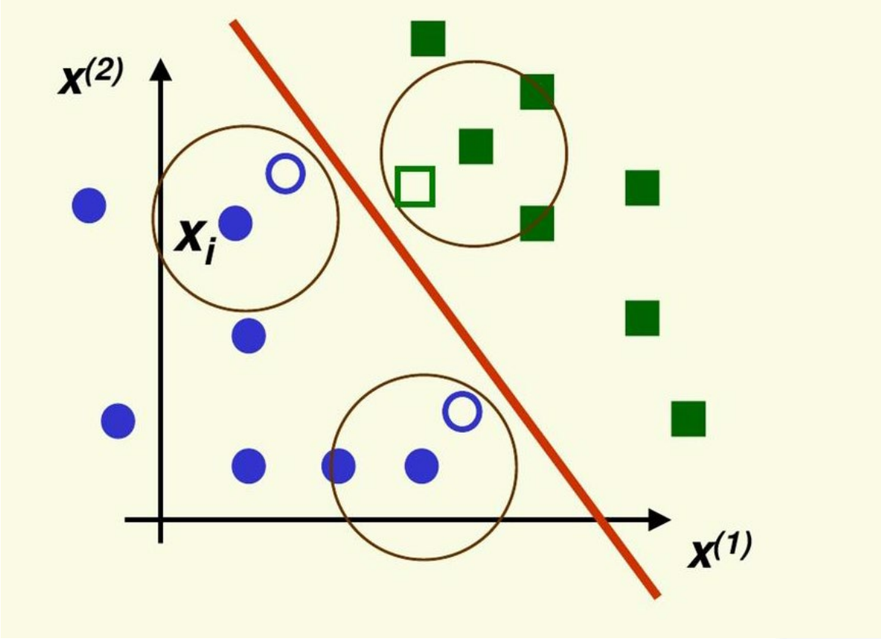
THE UNIVERSITY of ADELAIDE

Support Vector Machines (SVM) idea



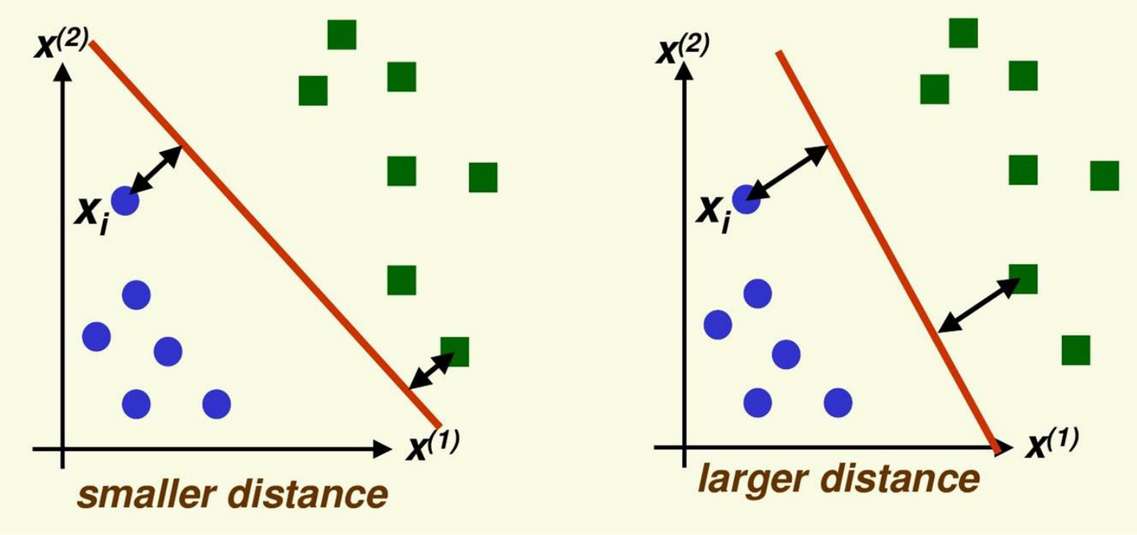
What makes a good separating hyperplane

- Hyperplane should be as far as possible from every sample
- In this way, new samples close to old samples will be classified correctly
- Good generalization



Support Vector Machine

Idea: maximize the distance to the closest example = margin



This leads to an optimization problem

SVM is a Supervised Algorithm

Supervised:

- We need labelled data
- We hope our labelled data is representative of ‘the real data’
- We hope our prediction factors separate our classes in a generalised way

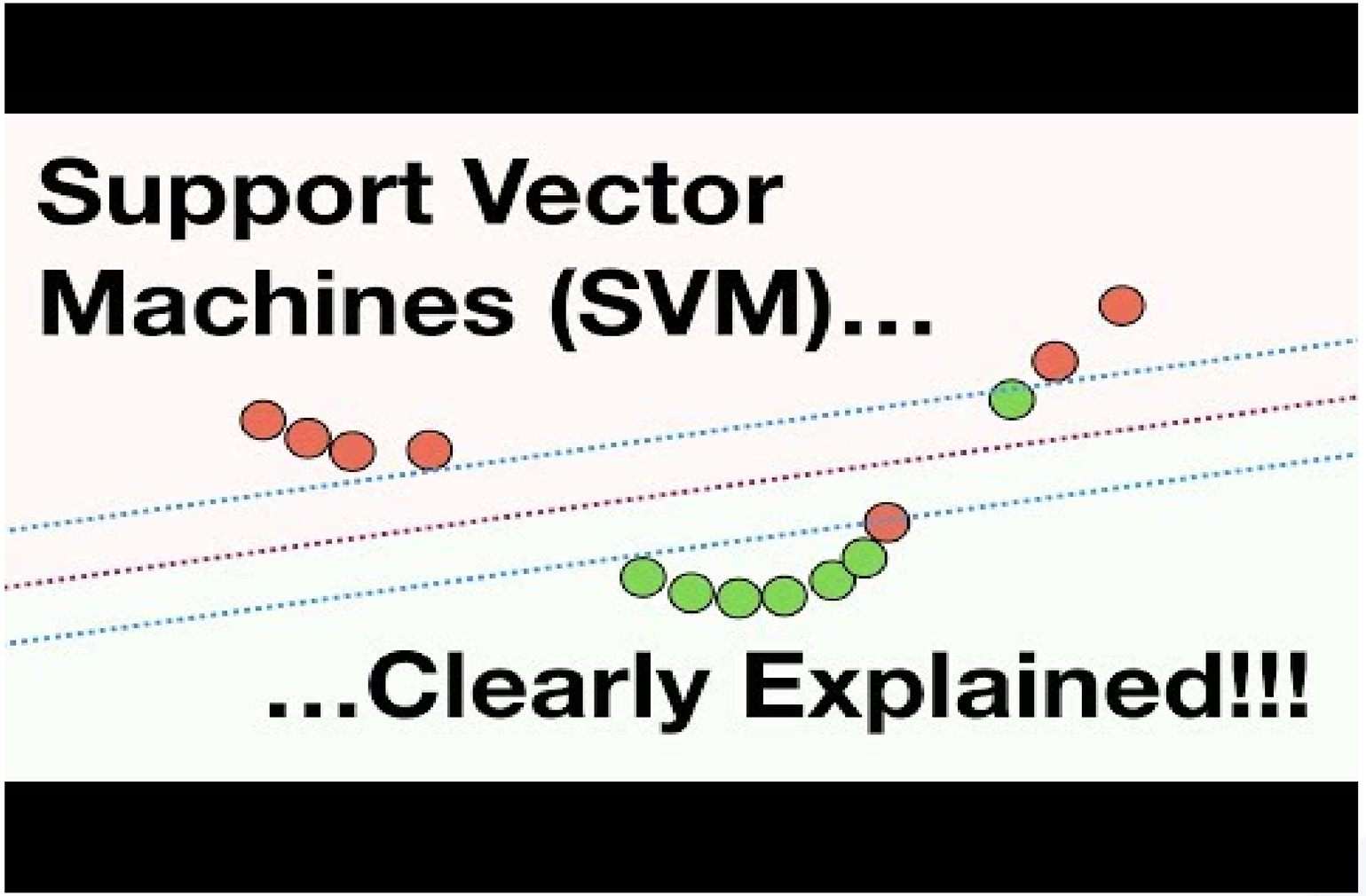
Heart Disease?	Height	Weight	Blood Pressure	…
Yes	1.8	25	120/80	…
No	1.7	100	130/70	
No	1.75	80	110/70	
Yes	1.6	70	110/60	
Yes	1.9	110	150/70	
No	1.7	160	180/50	
Yes	1.5	110	120/80	

Class

Prediction Factors

THE UNIVERSITY of ADELAIDE

This video explains soft margin and kernel trick



https://www.youtube.com/watch?v=efR1C6CvhmE

THE UNIVERSITY
of ADELAIDE