UNIVERSITY OF ABERDEEN

CELEBRATING 525 YEARS 1495 – 2020

ABERDEEN 2040

Naïve Bayes

Data Mining & Visualisation

Lecture 13

2025

Today...

- Generative vs Discriminative Classifiers
- Bayes' Theorem
- Naïve Bayes

Today...

More broadly, we're going to continue our discussion on classification, focusing on two popular approaches:

- Naïve Bayes Classifiers
- Support Vector Machines

Generative vs Discriminative Classifiers

Generative vs Discriminative Classifiers

When we're looking to classify some categorical variable (Y), using some observable variable(s) (X), we look to estimate:

$$P(Y \mid X)$$

Generative vs Discriminative Classifiers

When we're looking to classify some categorical variable (Y), using some observable variable(s) (X), we look to estimate:

$$P(Y \mid X)$$

And when dealing with classification tasks, we sometimes differentiate between two main approaches:

- **Discriminative classifiers**
- **Generative classifiers**

Discriminative Classifiers

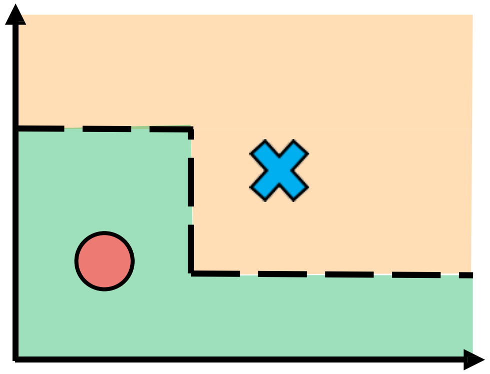
Discriminative Classifiers look to model the conditional probability of the target Y, given an observable x:

$$P(Y \mid X = x)$$

They do this through estimating the parameters of $P(Y \mid X)$ directly through the training data.

Discriminative Classifiers

Logistic regression and decision trees are both examples of discriminative classifiers, given that they use the training data to distinguish decision boundaries between classes.



Generative Classifiers

In contrast, **Generative Classifiers** look to model the conditional probability of the observable X, given a target y:

$$P(X \mid Y = y)$$

Generative Classifiers

In contrast, **Generative Classifiers** look to model the conditional probability of the observable X, given a target y:

$$P(X \mid Y = y)$$

They do this through instead estimating the parameters of

$$P(X \mid Y), P(Y)$$

Generative Classifiers

In contrast, **Generative Classifiers** look to model the conditional probability of the observable X, given a target y:

$$P(X \mid Y = y)$$

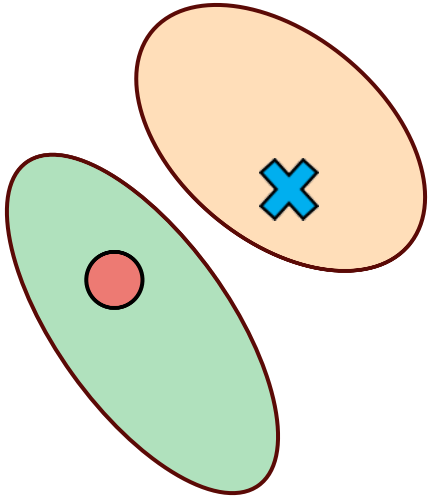
They do this through instead estimating the parameters of

$$P(X \mid Y), P(Y)$$

and then use **Bayes' Theorem** to calculate $P(Y \mid X = x)$

Generative Classifiers

Two of the most well-known generative classifiers are the Naïve Bayes Classifier (the subject of this lecture), and Linear Discriminant Analysis (LDA) – which we'll get to another day…



Bayes' Theorem

Bayes' Theorem

Thomas Bayes was a statistician and philosopher who lived in the 1700s.

Today, he is most well-known for his theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' Theorem

Bayes' theorem is used to determine the probability of a hypothesis, based on prior knowledge of conditions that are assumed to be relevant to that hypothesis.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' Theorem: Example

Let's say you go to the doctor with some unusual symptoms, and the doctor runs a few tests.

You test positive for a rare condition that affects 1/100 people.

99% of people with the disease test positive (*sensitivity*).

94% of people without the disease test negative (*specificity*).

How worried should you be?

Bayes' Theorem: Example

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' Theorem: Example

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(Have\ disease|Positive\, test) = \frac{P(Positive\ test|Have\ disease)P(Have\ disease)}{P(Positive\ test)}$$

Bayes' Theorem: Example

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(Have\ disease|Positive\ test) = \frac{P(Positive\ test|Have\ disease)P(Have\ disease)}{P(Positive\ test)}$$

$$P(Have\ disease|Positive\ test) = \frac{0.99 * 0.01}{(0.99 * 0.01) + (.06 * 0.99)}$$

Bayes' Theorem: Example

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(Have\ disease|Positive\ test) = \frac{P(Positive\ test|Have\ disease)P(Have\ disease)}{P(Positive\ test)}$$

$$P(Have\ disease|Positive\ test) = \frac{0.99 * 0.01}{(0.99 * 0.01) + (.06 * 0.99)} = 0.143$$

Bayes' Theorem

So even with a positive test, there's still an 85.7% chance that you don't actually have the disease.

Bayes' Theorem

So even with a positive test, there's still an 85.7% chance that you don't actually have the disease.

In short, Bayes' Theorem allows us to calculate conditional probabilities based on prior knowledge.

Bayes' Theorem

So even with a positive test, there's still an 85.7% chance that you don't actually have the disease.

In short, Bayes' Theorem allows us to calculate conditional probabilities based on prior knowledge.

So let's see how this can apply to classification!

Naïve Bayes Classification

Naïve Bayes

Naïve Bayes is the name given to a family of supervised learning algorithms for classification.

It is based on **Bayes' Theorem**, and is frequently used with high-dimensional datasets, such as within text classification.

Common examples include email (spam) classification, document classification, sentiment analysis, etc.

Class Conditional Independence Assumption

Naïve Bayes classifiers are naïve in the sense that they assume **all features are conditionally independent, given the class**.

In other words, we assume that all features are completely independent of each other.

The ‘Bag of Words’ Model

When talking about the **Class Conditional Independence Assumption** within the context of text analysis, we often use the phrase ‘bag of words’.

Imagine we take every word in a document, put it in a ‘bag’, and shake that bag up. We discard all ordering, context, etc.

This is essentially what it means when we’re assuming that all features are completely independent of each other.

How Good are Naïve Bayes Classifiers?

So despite their naivety, how good are Naïve Bayes Classifiers?

In practice, they actually work fairly well in many contexts!

Let’s walk through an example to see how they work.

Example: Spam Email Classification

Let’s say we want to build an email spam filter (a frequent example when discussing Naïve Bayes classifiers).

We have a dataset of emails:

- 75% are ‘Normal’ emails
- 25% are ‘Spam’ emails

Example: Spam Email Classification

We can start by building a histogram of all the words in both the Normal subset and the Spam subset.

Example: Spam Email Classification

We can start by building a histogram of all the words in both the Normal subset and the Spam subset.

Normal (75%)
Hey: 8
Uni: 6
Bank: 2

Spam (25%)
Hey: 2
Uni: 1
Bank: 6

Example: Spam Email Classification

We then calculate the probabilities for each word in each class

P(**N**) = .75
P(Hey|**N**) = 8 / 16 = .5
P(Uni|**N**) = 6 / 16 = .375
P(Bank|**N**) = 2 / 16 = .125

P(**S**) = .25
P(Hey|**S**) = 2 / 9 = .11
P(Uni|**S**) = 1 / 9 = .22
P(Bank|**S**) = 6 / 9 = .66

Example: Spam Email Classification

P(**N**) = .75
P(Hey|**N**) = 8 / 16 = .5
P(Uni|**N**) = 6 / 16 = .375
P(Bank|**N**) = 2 / 16 = .125

P(**S**) = .25
P(Hey|**S**) = 2 / 9 = .11
P(Uni|**S**) = 1 / 9 = .22
P(Bank|**S**) = 6 / 9 = .66

We receive a new email which contains the words 'hey' and 'uni'. What is the probability that it's spam?

For each class, we multiply each word's probability by *the prior probability*

Example: Spam Email Classification

P(**N**) = .75
P(Hey|**N**) = 8 / 16 = .5
P(Uni|**N**) = 6 / 16 = .375
P(Bank|**N**) = 2 / 16 = .125

P(**S**) = .25
P(Hey|**S**) = 2 / 9 = .11
P(Uni|**S**) = 1 / 9 = .22
P(Bank|**S**) = 6 / 9 = .66

Normal: P(**N**) * P(Hey|**N**) * P(Uni|**N**)

Spam: P(**S**) * P(Hey|**S**) * P(Uni|**S**)

Example: Spam Email Classification

P(**N**) = .75
P(Hey|**N**) = 8 / 16 = .5
P(Uni|**N**) = 6 / 16 = .375
P(Bank|**N**) = 2 / 16 = .125

P(**S**) = .25
P(Hey|**S**) = 2 / 9 = .11
P(Uni|**S**) = 1 / 9 = .22
P(Bank|**S**) = 6 / 9 = .66

Normal: P(**N**) * P(Hey|**N**) * P(Uni|**N**) = .75 * .5 * .375

Spam: P(**S**) * P(Hey|**S**) * P(Uni|**S**) = .25 * .11 * .22

Example: Spam Email Classification

P(**N**) = .75
P(Hey|**N**) = 8 / 16 = .5
P(Uni|**N**) = 6 / 16 = .375
P(Bank|**N**) = 2 / 16 = .125

P(**S**) = .25
P(Hey|**S**) = 2 / 9 = .11
P(Uni|**S**) = 1 / 9 = .22
P(Bank|**S**) = 6 / 9 = .66

Normal: P(**N**) * P(Hey|**N**) * P(Uni|**N**) = .75 * .5 * .375 = **.14**

Spam: P(**S**) * P(Hey|**S**) * P(Uni|**S**) = .25 * .11 * .22 = **.01**

Example: Spam Email Classification

P(**N**) = .75
P(Hey|**N**) = 8 / 16 = .5
P(Uni|**N**) = 6 / 16 = .375
P(Bank|**N**) = 2 / 16 = .125

P(**S**) = .25
P(Hey|**S**) = 2 / 9 = .11
P(Uni|**S**) = 1 / 9 = .22
P(Bank|**S**) = 6 / 9 = .66

P(**N**|Hey, Uni) ∝ **.14**

P(**S**|Hey, Uni) ∝ **.01**

Example: Spam Email Classification

P(**N**) = .75
P(Hey|**N**) = 8 / 16 = .5
P(Uni|**N**) = 6 / 16 = .375
P(Bank|**N**) = 2 / 16 = .125

P(**S**) = .25
P(Hey|**S**) = 2 / 9 = .11
P(Uni|**S**) = 1 / 9 = .22
P(Bank|**S**) = 6 / 9 = .66

P(**N**|Hey, Uni) ∝ **.14** **So we assume it's a normal email!**

P(**S**|Hey, Uni) ∝ **.01**