CPE412 Pattern Recognition

Week 5

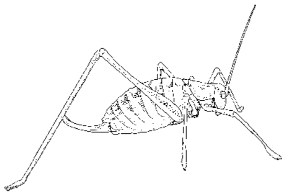
Bayesian Decision Theory

Dr. Nehad Ramaha,
Computer Engineering Department
Karabük Universities

The class notes are a compilation and edition from many sources. The instructor does not claim intellectual property or ownership of the lecture notes.

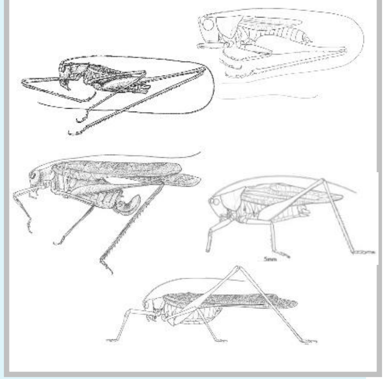
The Classification Problem
(informal definition)

Given a collection of annotated data. In this case 5 instances **Katydids** of and five of **Grasshoppers**, decide what type of insect the unlabeled example is.

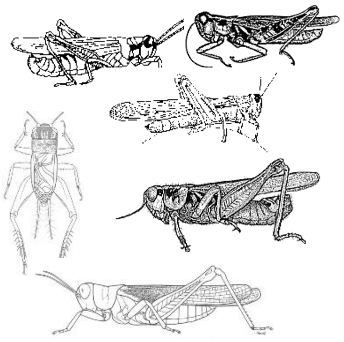


Katydid or **Grasshopper**?

Katydids



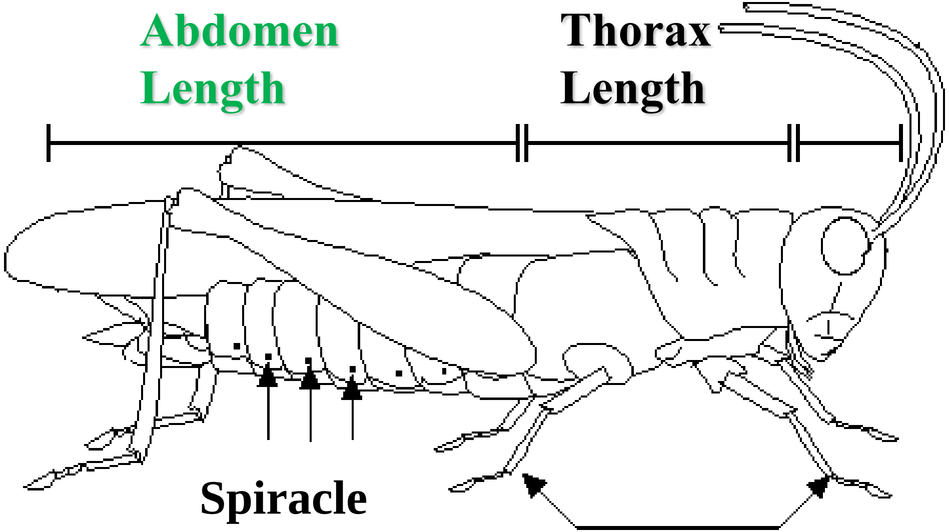
Grasshoppers



For any domain of interest, we can measure *features*

Color {Green, Brown, Gray, Other}

Has Wings?



Abdomen Length

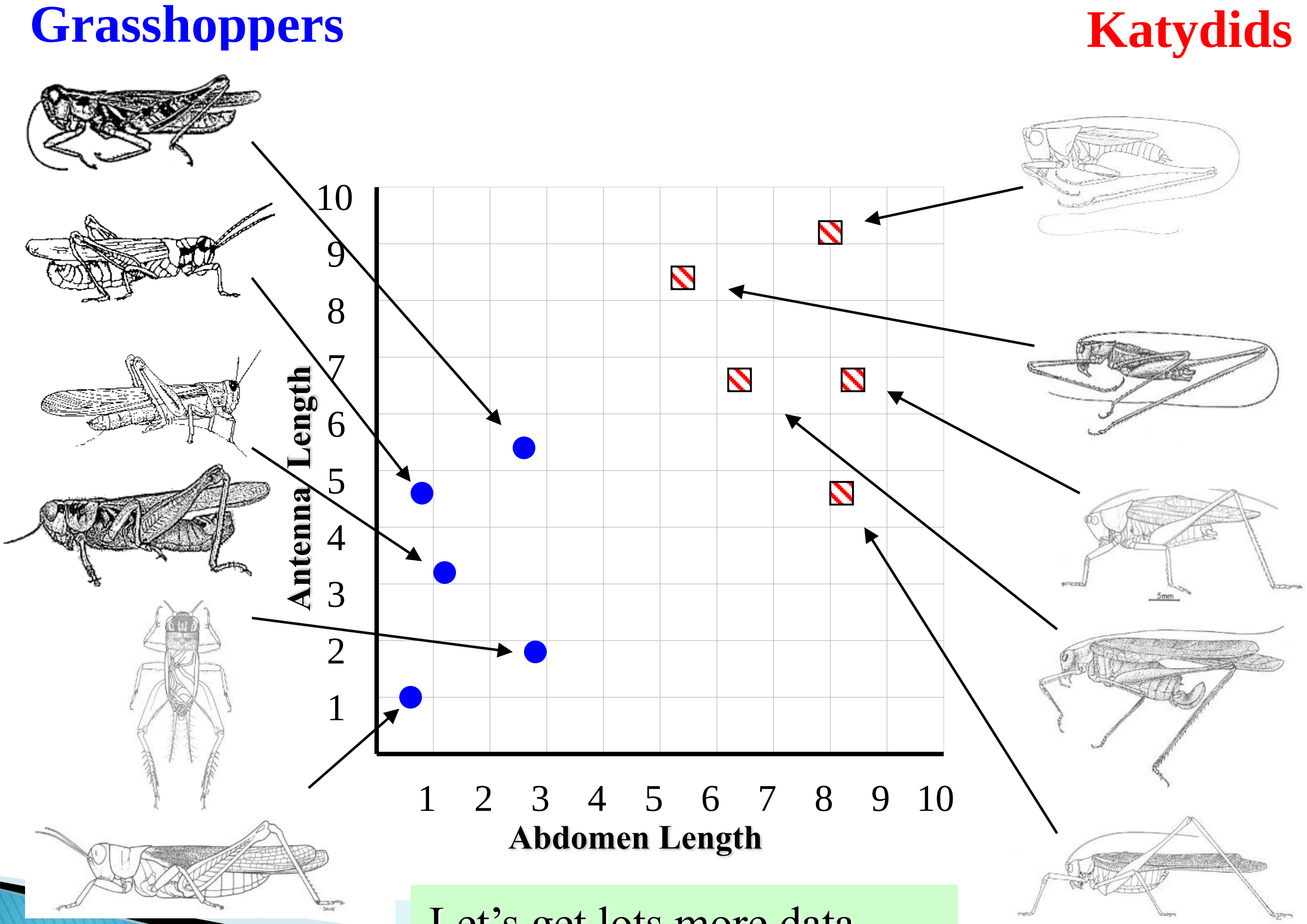
Thorax Length

Antennae Length

Mandible Size

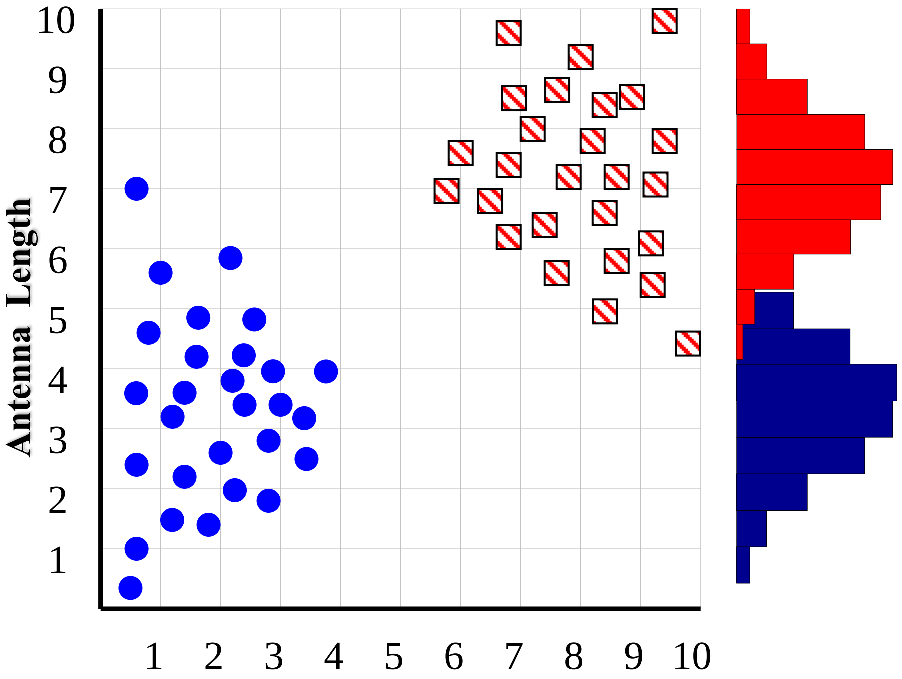
Spiracle Diameter

Leg Length



Let’s get lots more data…

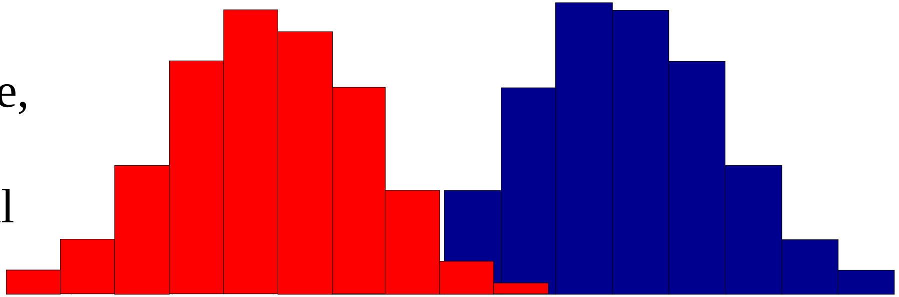
With a lot of data, we can build a histogram. Let us just build one for "Antenna Length" for now…



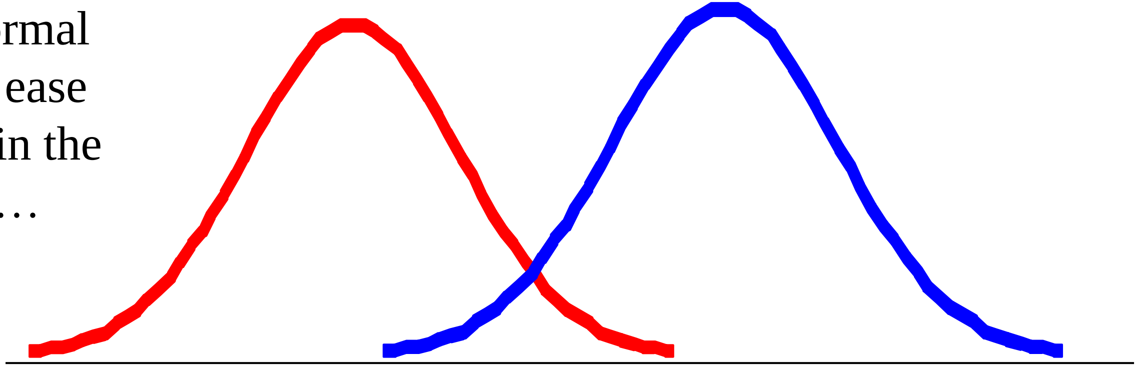
▨ Katydids

● Grasshoppers

We can leave the histograms as they are, or we can summarize them with two normal distributions.



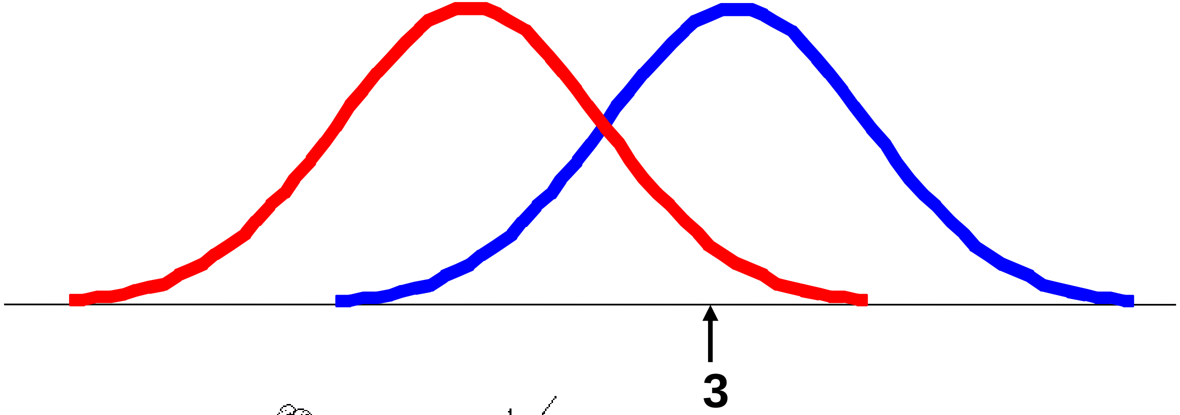
Let us us two normal distributions for ease of visualization in the following slides…

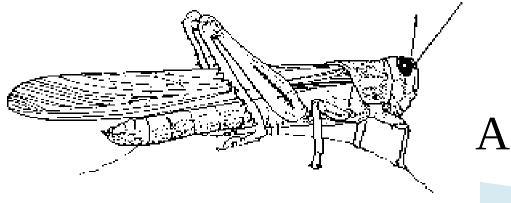


- We want to classify an insect we have found. Its antennae are 3 units long. How can we classify it?

- We can just ask ourselves, give the distributions of antennae lengths we have seen, is it more *probable* that our insect is a **Grasshopper** or a **Katydid**.
- There is a formal way to discuss the most *probable* classification…

$p(c_j \mid d)$ = probability of class c_j, *given* that we have observed d



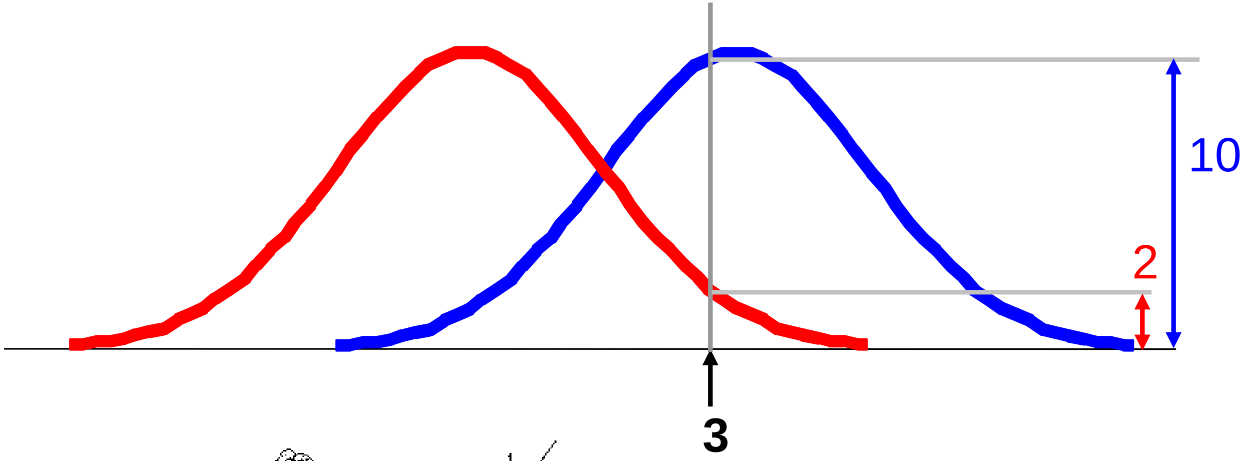


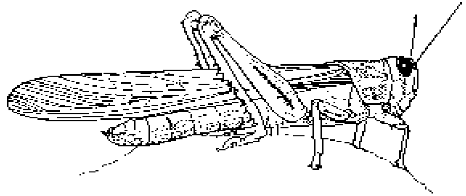
Antennae length is **3**

$p(c_j | d)$ = probability of class c_j, given that we have observed d

P(**Grasshopper** | **3**) = 10 / (10 + 2) = 0.833

P(**Katydid** | **3**) = 2 / (10 + 2) = 0.166



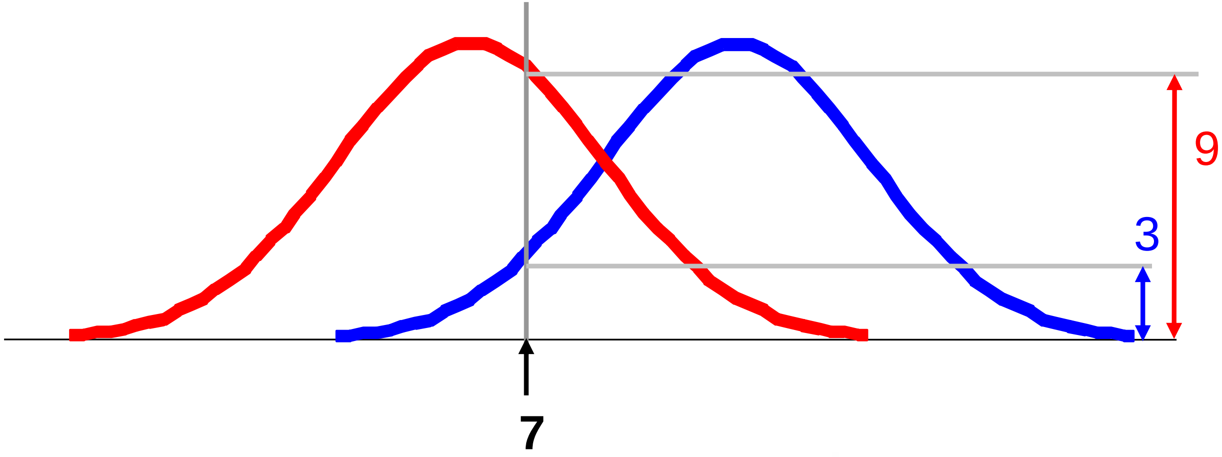


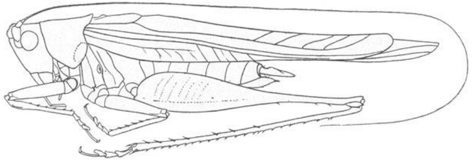
Antennae length is **3**

$p(c_j \mid d)$ = probability of class c_j, given that we have observed d

P(**Grasshopper** | **7**) = 3 / (3 + 9) = 0.250

P(**Katydid** | **7**) = 9 / (3 + 9) = 0.750



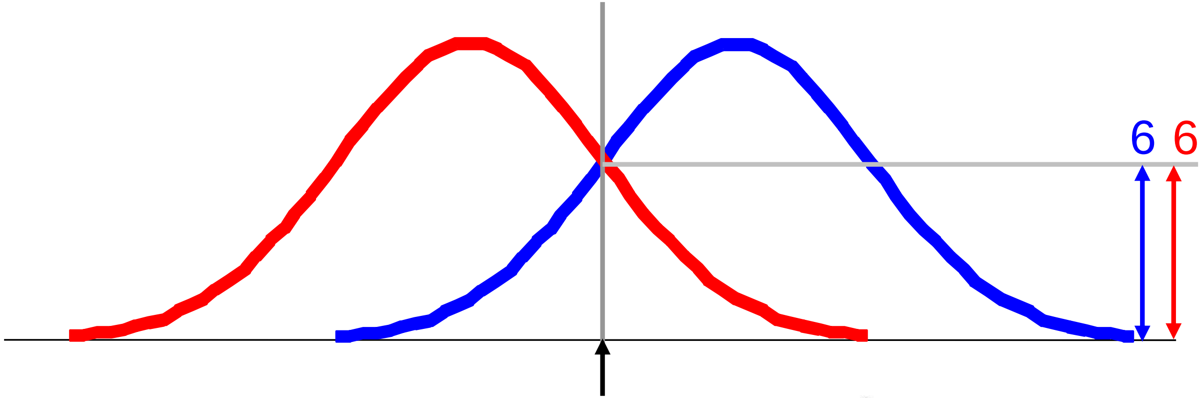


Antennae length is **7**

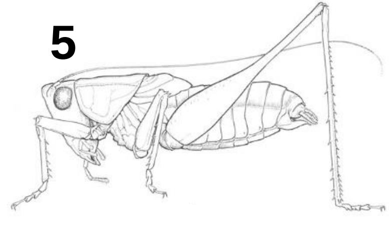
$p(c_j \mid d)$ = probability of class c_j, given that we have observed d

P(**Grasshopper** | **5**) = 6 / (6 + 6) = 0.500

P(**Katydid** | **5**) = 6 / (6 + 6) = 0.500



5



Antennae length is
5

Bayes Classifiers

That was a visual intuition for a simple case of the Bayes classifier, also called:

- Idiot Bayes
- Naïve Bayes
- Simple Bayes

We are about to see some of the mathematical formalisms, and more examples, but keep in mind the basic idea.

Find out the probability of the previously unseen instance belonging to each class, then simply pick the most probable class.

Bayes Classifiers

Assume that we have two classes

c_1 = male, and c_2 = female.

We have a person whose sex we do not know, say "*drew*" or *d*.

Classifying *drew* as male or female is equivalent to asking is it more probable that *drew* is male or female, i.e which is greater *p*(male | *drew*) or *p*(female | *drew*)

(Note: "Drew can be a male or female name")

Drew Barrymore

Drew Carey

What is the probability of being called "*drew*" given that you are a male?

$$p(\text{male} \mid drew) = \frac{p(drew \mid \text{male})\, p(\text{male})}{p(drew)}$$

What is the probability of being a male?

What is the probability of being named "*drew*"? (actually irrelevant, since it is that same for all classes)

Officer Drew

This is Officer Drew (who arrested me in 1997). Is Officer Drew a Male or Female?

Luckily, we have a small database with names and sex.

We can use it to apply Bayes rule…

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

Officer Drew

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(\text{male} \mid drew) = \frac{1/3 \;*\; 3/8}{3/8} \qquad = \frac{0.125}{3/8}$$

$$p(\text{female} \mid drew) = \frac{2/5 \;*\; 5/8}{3/8} \qquad = \frac{0.250}{3/8}$$

Officer Drew is more likely to be a Female.

Officer Drew IS a female!

Officer Drew

$$p(\text{male} \mid drew) = \frac{1/3 \; * \; 3/8}{3/8} \qquad = \frac{0.125}{3/8}$$

$$p(\text{female} \mid drew) = \frac{2/5 \; * \; 5/8}{3/8} \qquad = \frac{0.250}{3/8}$$

So far, we have only considered Bayes Classification when we have one attribute (the "*antennae length*", or the "*name*"). But we may have many features.

How do we use all the features?

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

Name	**Over 170CM**	**Eye**	**Hair length**	**Sex**
Drew	No	Blue	Short	Male
Claudia	Yes	Brown	Long	Female
Drew	No	Blue	Long	Female
Drew	No	Blue	Long	Female
Alberto	Yes	Brown	Short	Male
Karin	No	Blue	Long	Female
Nina	Yes	Brown	Short	Female
Sergio	Yes	Blue	Long	Male

- To simplify the task, **naïve Bayesian classifiers** assume attributes have independent distributions, and thereby estimate

$$p(d|c_j) = p(d_1|c_j) * p(d_2|c_j) * \ldots.* p(d_n|c_j)$$

The probability of class c_j generating instance d, equals….

The probability of class c_j generating the observed value for feature 1, multiplied by..

The probability of class c_j generating the observed value for feature 2, multiplied by..

- To simplify the task, **naïve Bayesian classifiers** assume attributes have independent distributions, and thereby estimate

$$p(d|c_j) = p(d_1|c_j) * p(d_2|c_j) * \ldots * p(d_n|c_j)$$

$$p(\text{officer drew}|c_j) = p(\text{over\_170}_{\text{cm}} = \text{yes}|c_j) * p(\text{eye} = blue|c_j) * \ldots$$

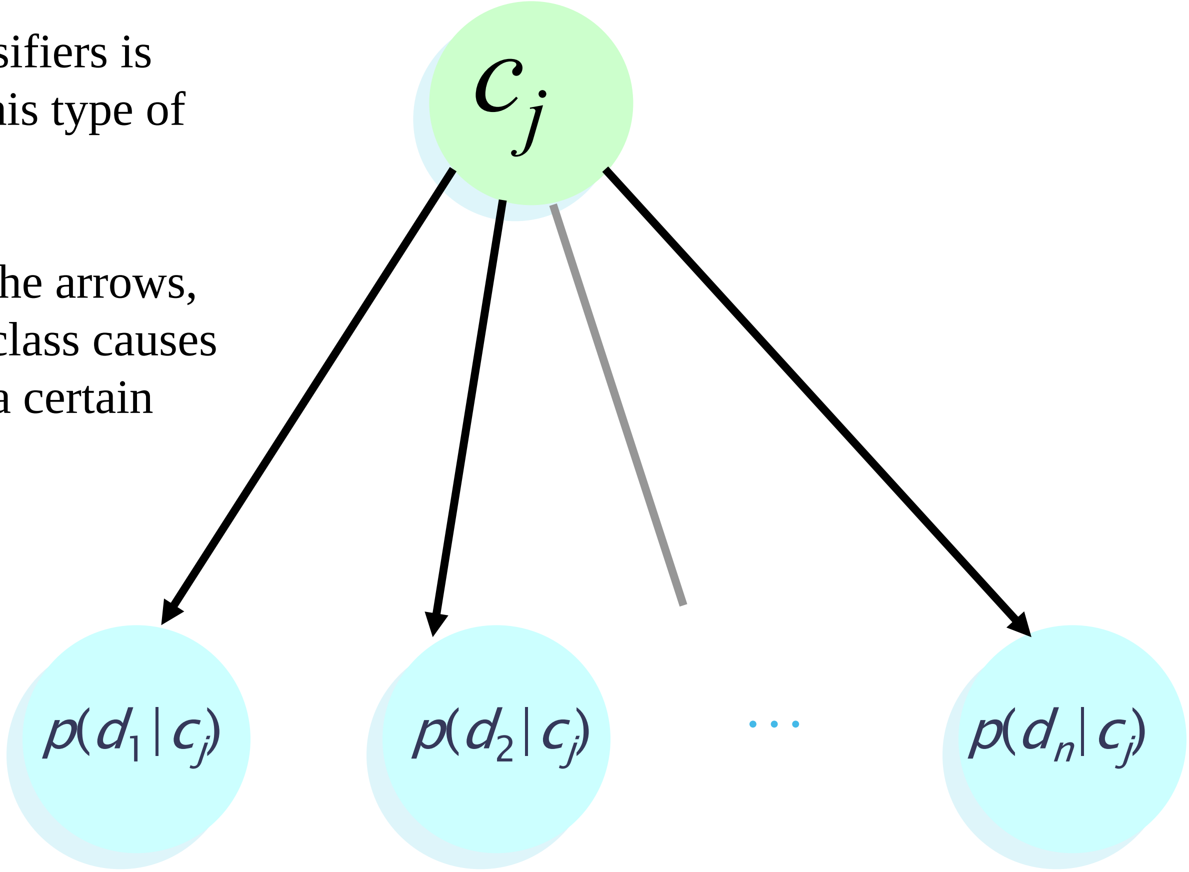
Officer Drew is blue-eyed, over 170$_{\text{cm}}$ tall, and has long hair

$$p(\text{officer drew}|\ \text{Female}) = 2/5 \quad * \quad 3/5 \quad * \quad \ldots$$

$$p(\text{officer drew}|\ \text{Male}) = 2/3 \quad * \quad 2/3 \quad * \quad \ldots$$

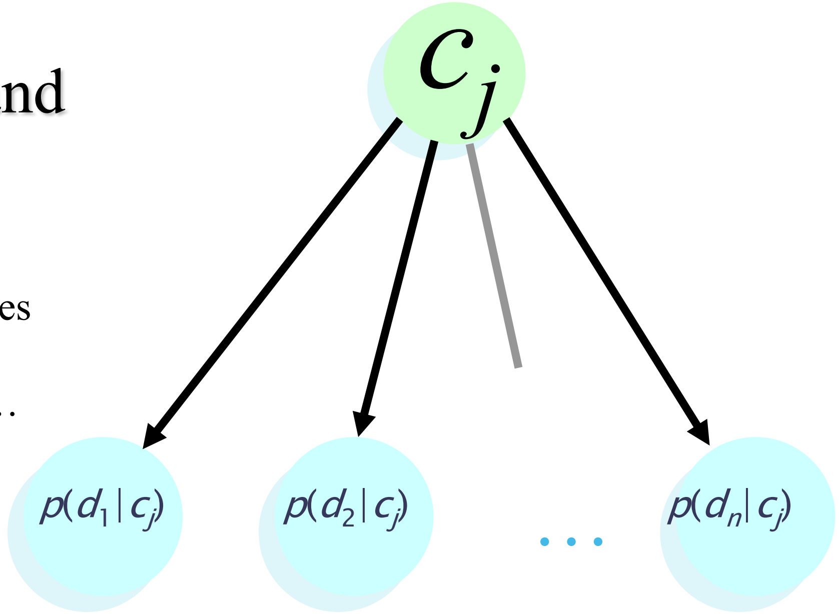
The Naive Bayes classifiers is often represented as this type of graph…

Note the direction of the arrows, which state that each class causes certain features, with a certain probability



Naïve Bayes is fast and space efficient

We can look up all the probabilities with a single scan of the database and store them in a (small) table…

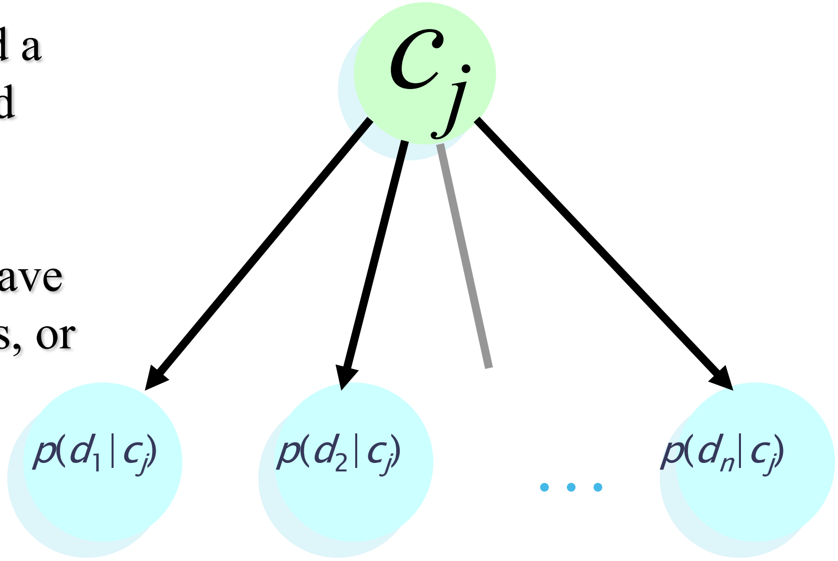


Sex	**Over190cm**	
Male	Yes	0.15
	No	0.85
Female	Yes	0.01
	No	0.99

Sex	**Long Hair**	
Male	Yes	0.05
	No	0.95
Female	Yes	0.70
	No	0.30

Sex	
Male	
Female	

An obvious point. I have used a simple two class problem, and two possible values for each example, for my previous examples. However we can have an arbitrary number of classes, or feature values



Animal	Mass >10_{kg}	
Cat	Yes	0.15
	No	0.85
Dog	Yes	0.91
	No	0.09
Pig	Yes	0.99
	No	0.01

Animal	Color	
Cat	Black	0.33
	White	0.23
	Brown	0.44
Dog	Black	0.97
	White	0.03
	Brown	0.90
Pig	Black	0.04
	White	0.01
	Brown	0.95

Animal
Cat
Dog
Pig

Advantages/Disadvantages of Naïve Bayes

- Advantages:
  - Fast to train (single scan). Fast to classify
  - Not sensitive to irrelevant features
  - Handles real and discrete data
  - Handles streaming data well
- Disadvantages:
  - Assumes independence of features:

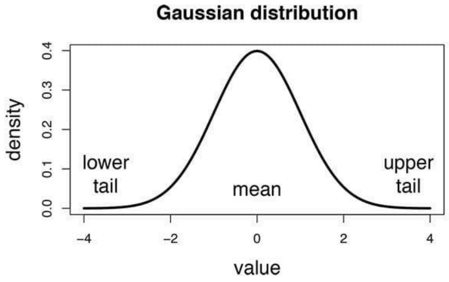
    Relationships between variables cannot be modeled because operations are performed assuming the features are independent of each other.

Naive Bayes Application Areas

- Real Time Systems
- Multiple Classification Problems (News / E-Commerce Categories)
- Text Classification (Spam Filtering / Sentiment Analysis)
- Disease Diagnosis
- Recommendation System

Types of Naive Bayes Classifier:

- Multinomial Naive Bayes:
  - This is mostly used for document classification problem, i.e whether a document belongs to the category of sports, politics, technology etc. The features/predictors used by the classifier are the frequency of the words present in the document.
- Bernoulli Naive Bayes:
  - This is similar to the multinomial naive bayes but the predictors are Boolean variables. The parameters that we use to predict the class variable take up only values yes or no, for example if a word occurs in the text or not.
- Gaussian Naive Bayes:
  - When the predictors take up a continuous value and are not discrete, we assume that these values are sampled from a gaussian distribution.



Naive Bayes Tips

- If your continuous features are not normally distributed, we must convert them to normal distribution using various methods or transformations.
- If there is a zero-frequency situation in our data set,
- If you have two categories that are very similar to each other and have a lot of relationship, it is recommended to remove one of them. This is because this feature will be counted as voted twice and will seem overly important.
- There are not many parameters in the Naive Bayes algorithm that you can play with and improve the model. If you are going to use Naive Bayes for this, you need to do data pre-processing, especially feature selection, very well.

Naïve Bayes' Classifier Example:

- Suppose we have a dataset of weather conditions and corresponding target variable "Play". So using this dataset we need to decide that whether we should play or not on a particular day according to the weather conditions.
- To solve this problem, we need to follow the below steps:
  - Convert the given dataset into frequency tables.
  - Generate Likelihood table by finding the probabilities of given features.
  - Now, use Bayes theorem to calculate the posterior probability.

Gaussian Naïve Bayes' Classifier Example:

- Problem: If the weather is sunny, then the Player should play or not?
- Solution: To solve this, first consider the below dataset:

	Outlook	Play
0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rainy	Yes
6	Sunny	Yes
7	Overcast	Yes
8	Rainy	No
9	Sunny	No
10	Sunny	Yes
11	Rainy	No
12	Overcast	Yes
13	Overcast	Yes

Gaussian Naïve Bayes' Classifier Example:

- Frequency table for the Weather Conditions:

Weather	Yes	No
Overcast	5	0
Rainy	2	2
Sunny	3	2
Total	10	5

Gaussian Naïve Bayes' Classifier Example:

- Likelihood table weather condition:

Weather	No	Yes	
Overcast	0	5	5/14= 0.35
Rainy	2	2	4/14=0.29
Sunny	2	3	5/14=0.35
All	4/14=0.29	10/14=0.71	

Gaussian Naïve Bayes' Classifier Example:

- Applying Bayes' theorem:
- **P(Yes|Sunny)= P(Sunny|Yes)*P(Yes)/P(Sunny)**
- P(Sunny|Yes)= 3/10= 0.3
- P(Sunny)= 0.35
- P(Yes)=0.71
- So P(Yes|Sunny) = 0.3*0.71/0.35= **0.60**

Gaussian Naïve Bayes' Classifier Example:

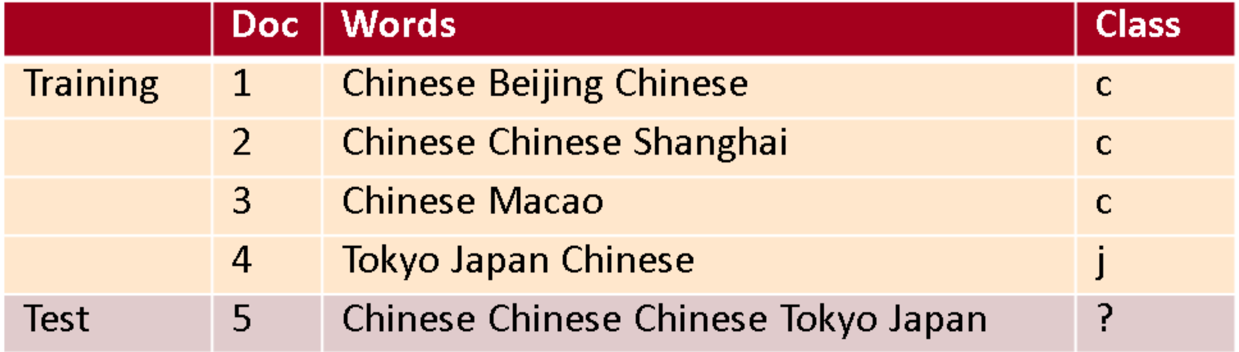
- Applying Bayes’ theorem:
- **P(No|Sunny)= P(Sunny|No)*P(No)/P(Sunny)**
- P(Sunny|NO)= 2/4=0.5
- P(No)= 0.29
- P(Sunny)= 0.35
- So P(No|Sunny)= 0.5*0.29/0.35 = **0.41**

Gaussian Naïve Bayes' Classifier Example:

- Applying Bayes' theorem:
- P(Yes|Sunny) = **0.60**
- P(No|Sunny)= **0.41**
- So as we can see from the above calculation that P(Yes|Sunny)>P(No|Sunny)
- Hence on a Sunny day, Player can play the game.

Multinomial Naive Bayes :

Whether a document/topic belongs to a particular category. The features/predictors used by the classifier are the frequency of words found in the document.

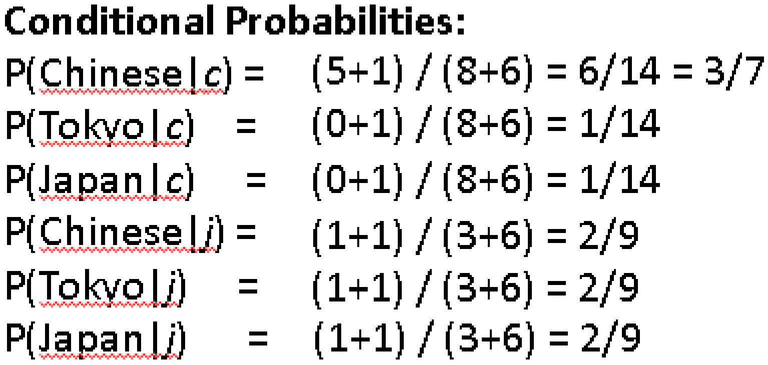


	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Multinomial Naive Bayes :

- P(C) = 3/4 = 0.75 (Ratio of rows in category C to all rows in the data to be taught)
- P(J) = 1/4 = 0.25 (Ratio of rows in the Japan category to all rows in the data to be taught)
- P(X| Y) =(Number of repetitions of the expression “X” in the lines in category Y +1) / (Number of all words in the lines in category Y + Number of data taught)

Multinomial Naive Bayes :



Conditional Probabilities:

P(Chinese|*c*) = (5+1) / (8+6) = 6/14 = 3/7

P(Tokyo|*c*) = (0+1) / (8+6) = 1/14

P(Japan|*c*) = (0+1) / (8+6) = 1/14

P(Chinese|*j*) = (1+1) / (3+6) = 2/9

P(Tokyo|*j*) = (1+1) / (3+6) = 2/9

P(Japan|*j*) = (1+1) / (3+6) = 2/9

P(C | Test) = P(C) * P(Chinese | C) * P(Chinese | C) * P(Chinese | C) * P(Tokyo| C) * P(Japan| C)

P(Ç | Test) = 0.75 * 0.428 * 0.428 * 1 * 0.428 * 0.071 * 0.071 = **0.0003**

P(Japan| Test) = P(J) * P(Chinese | J) * P(Chinese | J) * P(Chinese | J) * P(Tokyo| J) * P(Japan| J)

P(Japan| Test) = 0.25 * 0.222 * 0.222 * 1 * 0.222 * 0.022 * 0.022 = 0.0001

Thanks ☺