Data Mining

Lecture # 7
Naïve Bayes Classifier

Naïve Bayes Classifier

Thomas Bayes
1702 - 1761

We will start off with a visual intuition, before looking at the math…

Background

- There are three methods to establish a classifier

a) Model a classification rule directly

Examples: k-NN, decision trees, perceptron, SVM

b) Model the probability of class memberships given input data

Example: multi-layered perceptron with the cross-entropy cost

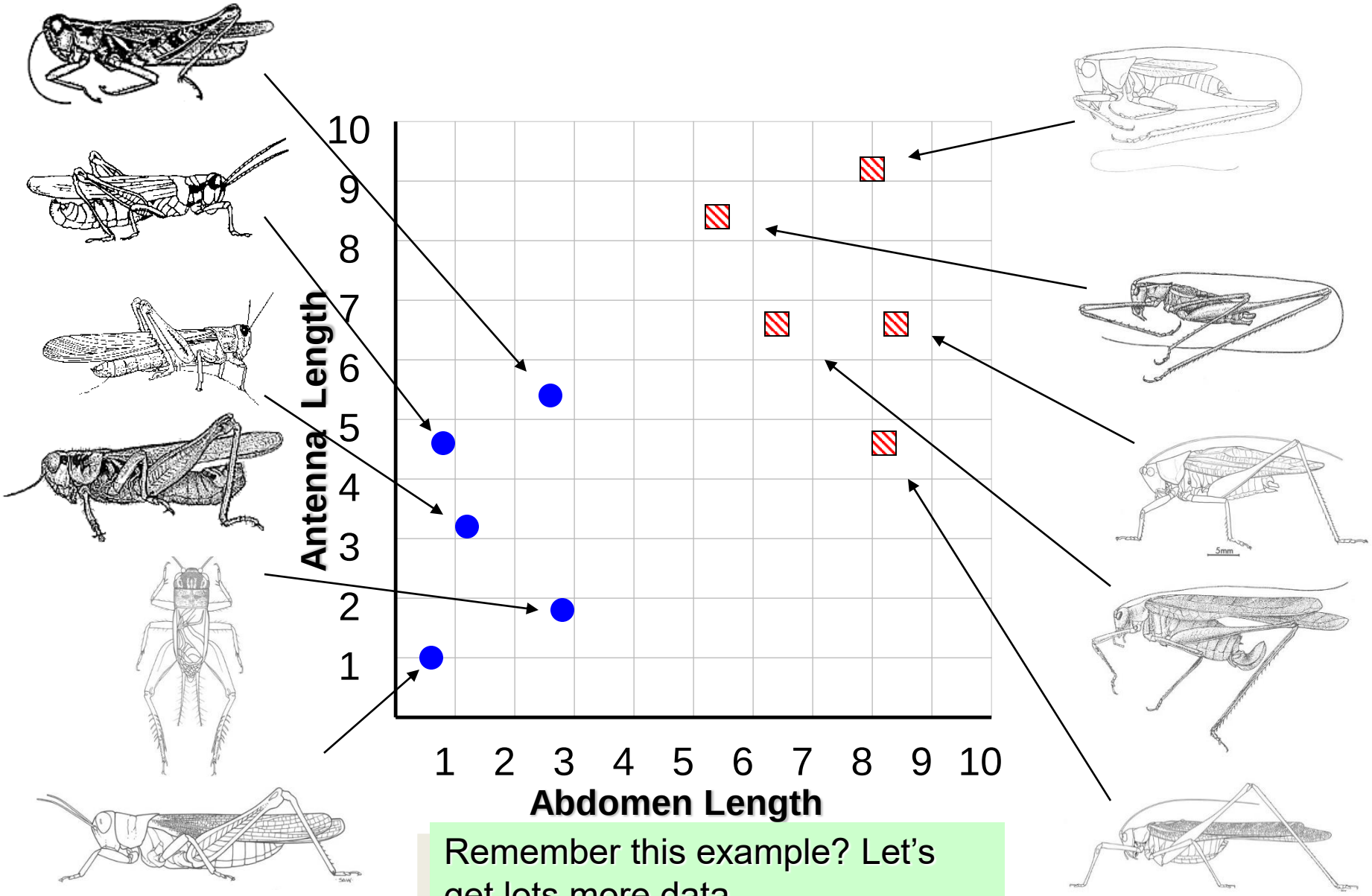
c) Make a probabilistic model of data within each class

Examples: naive Bayes, model based classifiers

- *a*) and *b*) are examples of discriminative classification
- *c*) is an example of generative classification
- *b*) and *c*) are both examples of probabilistic classification

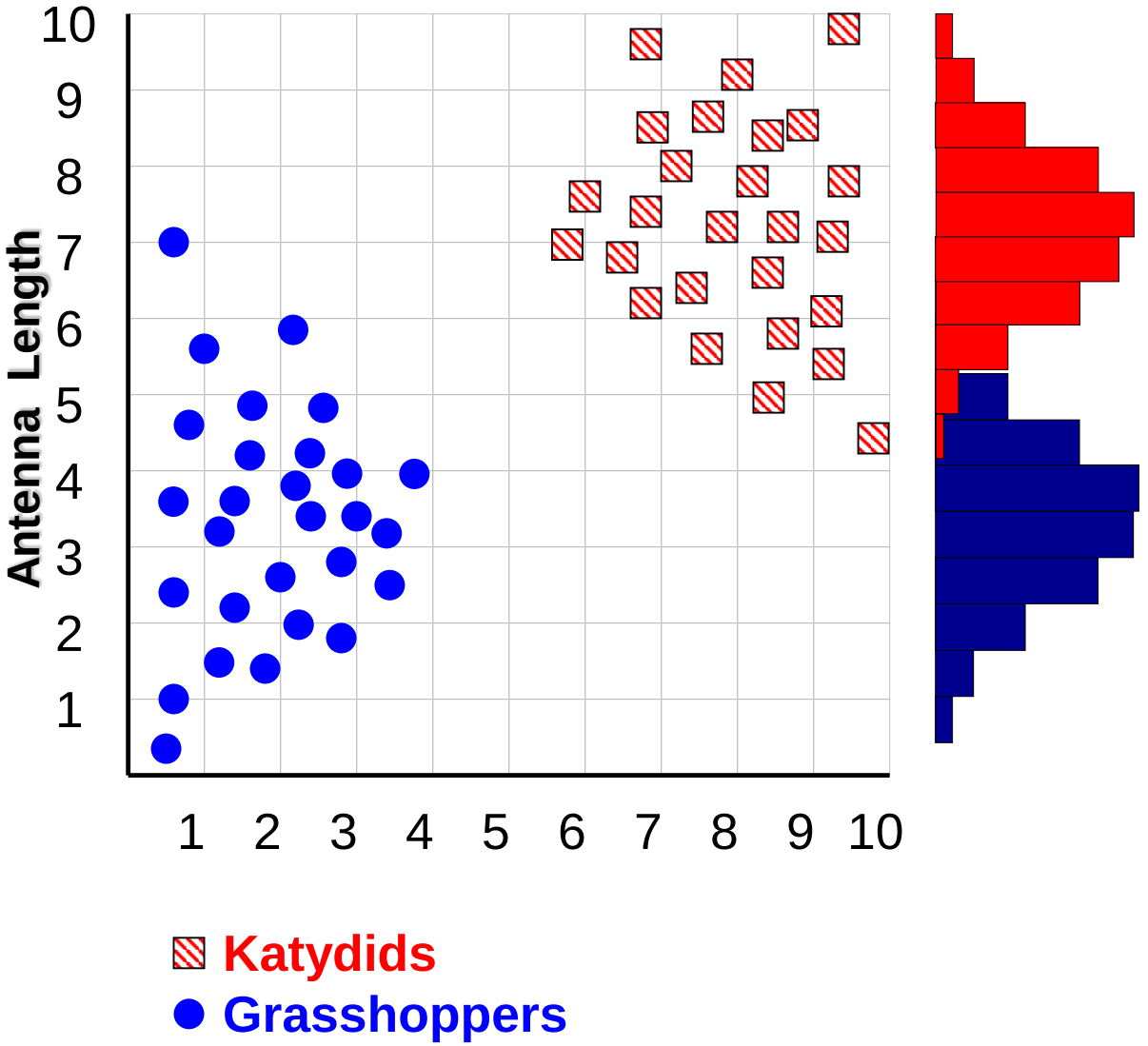
Grasshoppers

Katydids

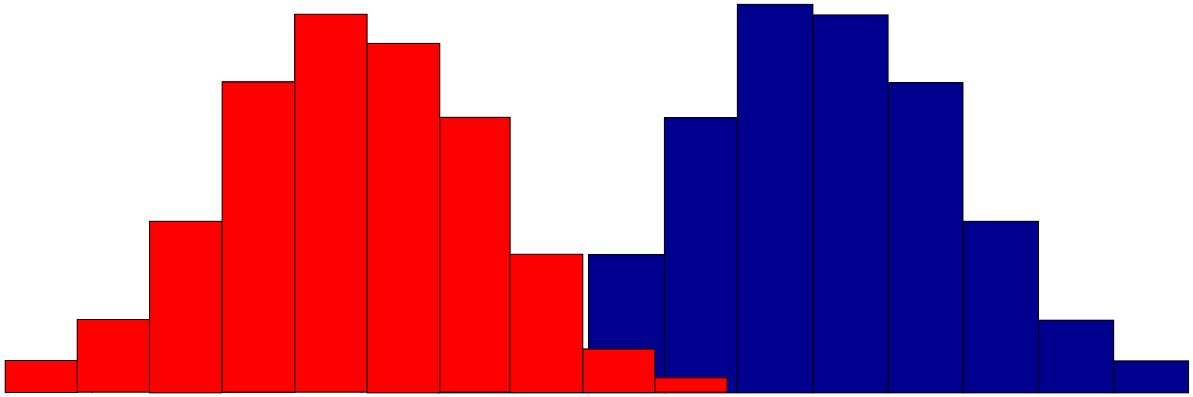


Remember this example? Let's get lots more data…

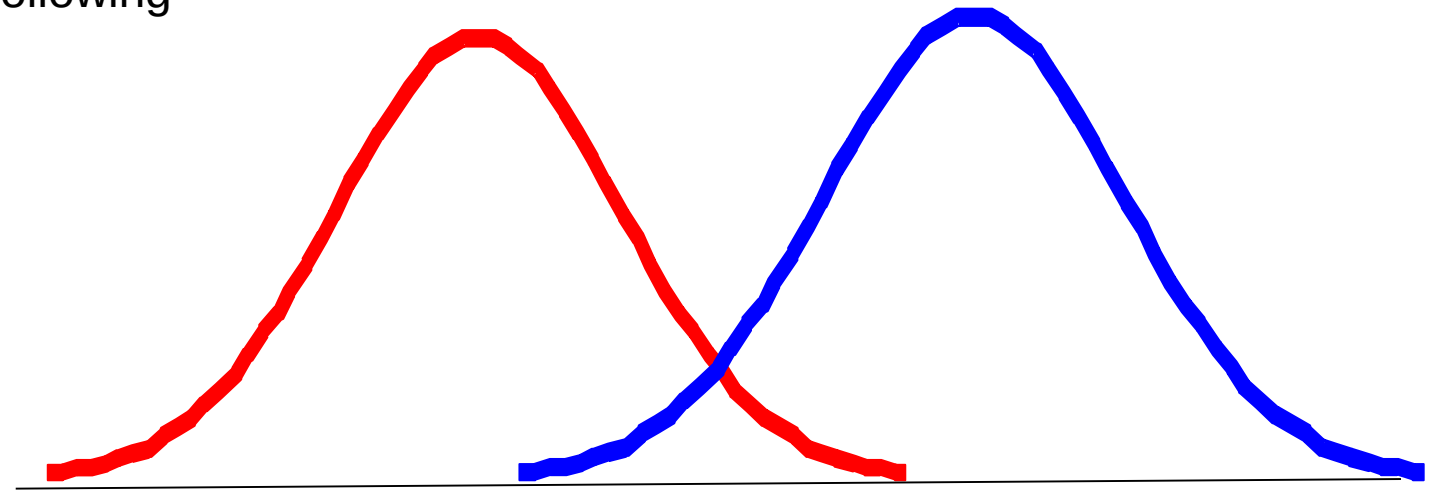
With a lot of data, we can build a histogram. Let us just build one for “Antenna Length” for now…



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



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

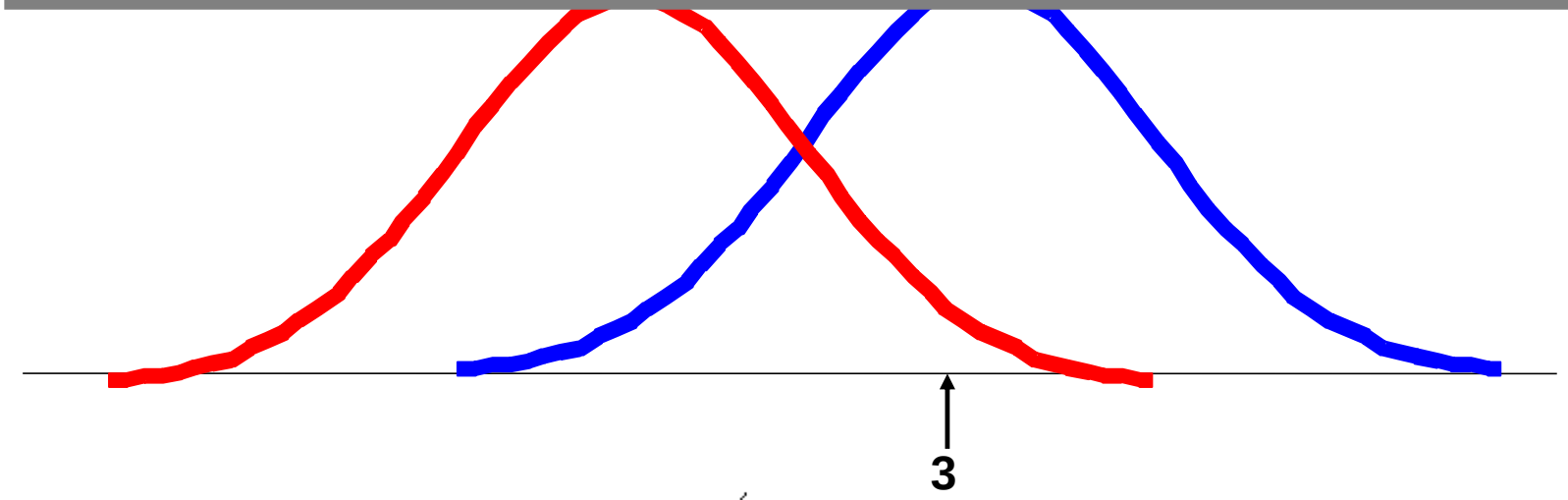


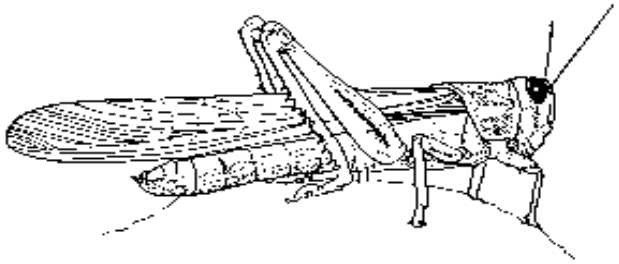
• We want to classify an insect we have found. Its antennae are 3 units long. How can we classify it?

• 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



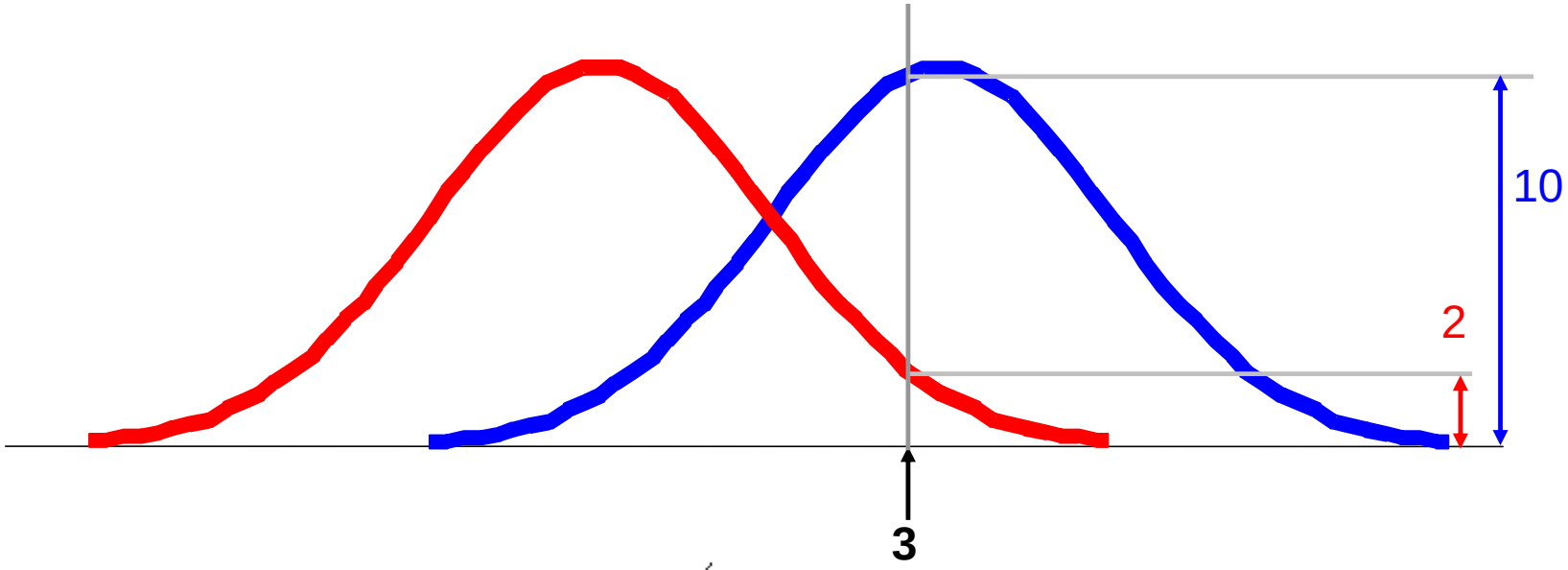


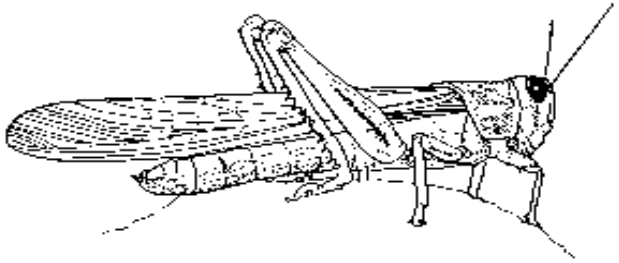
Antennae length is **3**

$p(c_j \mid d)$ = 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



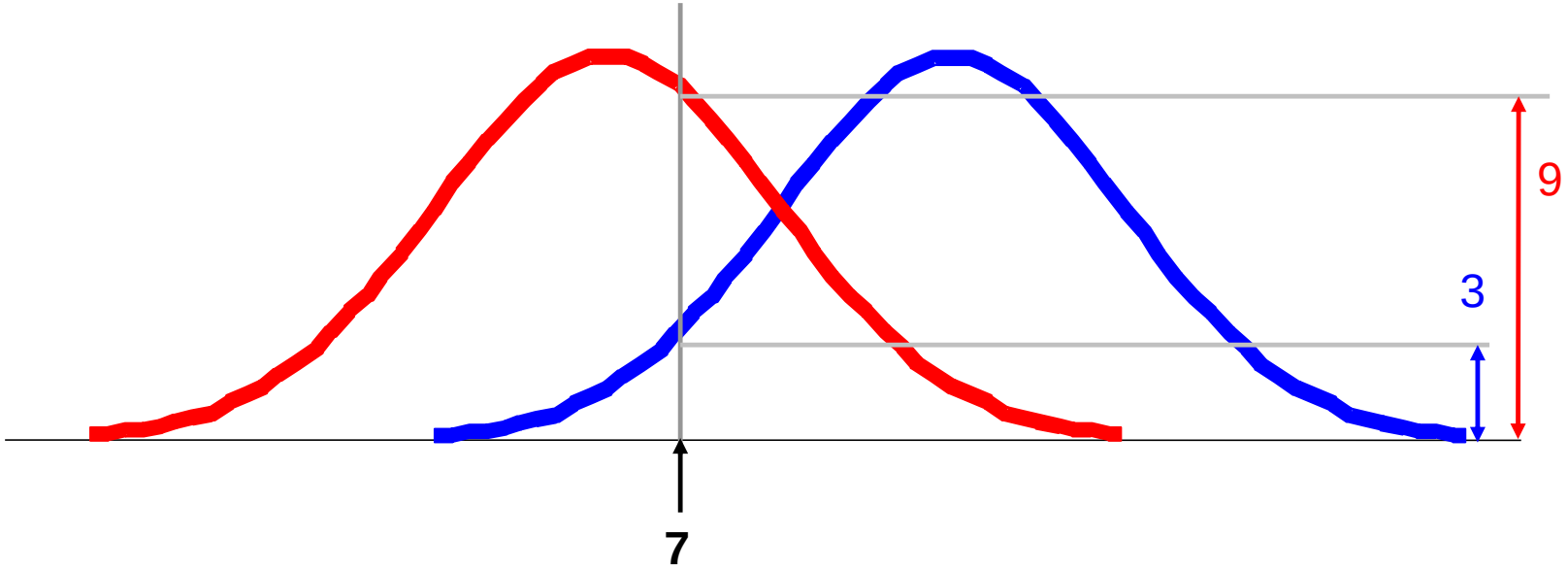


Antennae length is **3**

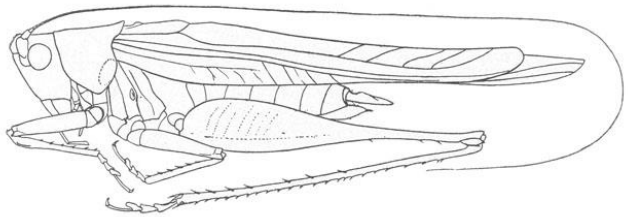
$p(c_j \mid d)$ = probability of class c_j, given that we have observed d

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

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



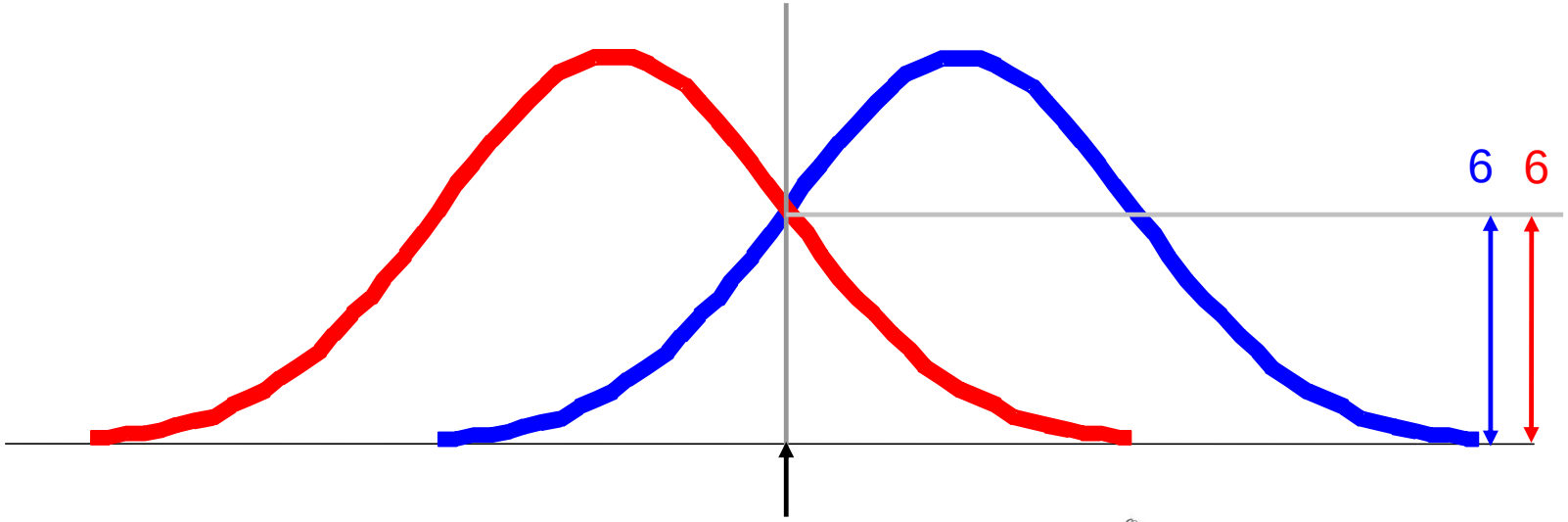
Antennae length is **7**



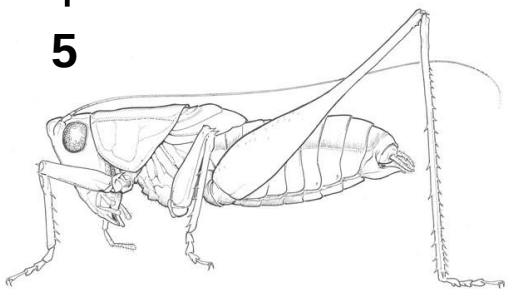
$p(c_j \mid d)$ = probability of class c_j, given that we have observed d

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

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



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.

Self Study
Concepts related to probability

Probability Basics

- Prior, conditional and joint probability
  - Prior probability: $P(X)$
  - Conditional probability: $P(X_1 \mid X_2), P(X_2 \mid X_1)$
  - Joint probability: $\mathbf{X} = (X_1, X_2), P(\mathbf{X}) = P(X_1, X_2)$
  - Relationship: $P(X_1, X_2) = P(X_2 \mid X_1)P(X_1) = P(X_1 \mid X_2)P(X_2)$
  - Independence: $P(X_2 \mid X_1) = P(X_2), P(X_1 \mid X_2) = P(X_1), P(X_1, X_2) = P(X_1)P(X_2)$
- Bayesian Rule

$$P(C \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid C)P(C)}{P(\mathbf{X})} \qquad Posterior = \frac{Likelihood \times Prior}{Evidence}$$

Meaning of Probabilities

- $\mathbb{P}(A)$ denotes our belief that event A will happen
  - ⋆ $\mathbb{P}(A) = 0$ means that the event never occurs
  - ⋆ $\mathbb{P}(A) = 1$ meaning that the event always occurs
  - ⋆ For a fair coin we expect $\mathbb{P}(\{\text{head}\}) = \mathbb{P}(\{\text{tail}\}) = 1/2$

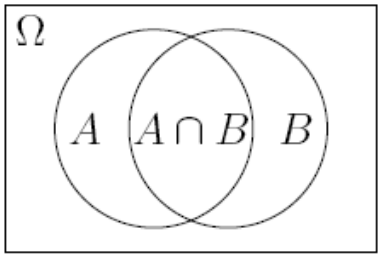
- Denoting the event "A does not occurring" by $\neg A$ then

$$\mathbb{P}(A) + \mathbb{P}(\neg A) = 1$$

- If we consider a set of *exhaustive* and *mutually exclusive* events, $\{A_i | i \in I\}$, where $I \in \mathbb{N}$ is an index set

$$\sum_{i \in I} \mathbb{P}(A_i) = 1$$

Joint Probabilities



- If we have two events A and B we can define the *joint* probability of them both occurring $\mathbb{P}(A \cap B)$ or $\mathbb{P}(A, B)$
- The probability of just A occurring is

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap \neg B)$$

- If $\{B_i | i \in I\}$ forms an exhaustive and mutually exclusive set of events then

$$\mathbb{P}(A) = \sum_{i \in \mathcal{I}} \mathbb{P}(A \cap B_i)$$

- This provides one of the fundamental rules for manipulating probabilities

Conditional Probabilities

- The *conditional probability* of event A occurring *given* that event B has occurred is denoted $\mathbb{P}(A|B)$
- Conditional probabilities are connected to joint probabilities by the relationship

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\,\mathbb{P}(B) = \mathbb{P}(B|A)\,\mathbb{P}(A)$$

- Note that conditional probabilities doesn't say anything about causality
- This provides the second fundamental rule for manipulating probabilities

Independence

- Two events, A and B, are said to be $independent$ if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\ \mathbb{P}(B)$$

- Since $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\,\mathbb{P}(B)$ an equivalent condition for independence is

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

- Independence makes life much simpler as we can ignore events that are independent to those we are interested in

A Calculus for Probabilities

- To manipulate probabilities we can use

$$\mathbb{P}(X, Y) = \mathbb{P}(X|Y)\,\mathbb{P}(Y) = \mathbb{P}(Y|X)\,\mathbb{P}(X)$$

- We can rewrite this as

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

- This is a trivial identity, but is really all we need

Example

A patient comes to a doctor with a symptom S. The doctor knows that 30% of the people with this symptom have a disease D from which they will die if they are not operated on immediately. However, the operation is dangerous with 50% fatality irrespective of whether the patient has the disease or not. The doctor knows that 60% of people with D have blood type A while only 30% of the normal population has this blood type. On performing a blood test it is found that the patient has blood type A. Should the doctor operate?

Solution

$$\mathbb{P}(D|A,S) = \frac{\mathbb{P}(A|D,S)\,\mathbb{P}(D|S)}{\mathbb{P}(A|S)}$$

Prior: $\mathbb{P}(D|S) = 0.3$

Likelihood: $\mathbb{P}(A|D,S) = \mathbb{P}(A|D) = 0.6$

Evidence: $\mathbb{P}(A|S)$ which we can calculate as

$$\begin{aligned}
\mathbb{P}(A|S) &= \mathbb{P}(A,D|S) + \mathbb{P}(A,\neg D|S) \\
&= \mathbb{P}(A|D,S)\,\mathbb{P}(D|S) + \mathbb{P}(A|\neg D,S)\,\mathbb{P}(\neg D|S) \\
&= 0.6 * 0.3 + 0.3 * 0.7 = 0.39
\end{aligned}$$

Posterior: $\mathbb{P}(D|A,S) = 0.6 * 0.3/0.39 = 0.46$

Bayes' Rule

- We want to decide between different hypotheses, $\mathcal{H}_1$, $\mathcal{H}_2$, $\cdots$, $\mathcal{H}_n$
- To make the decision we generate some data $\mathcal{D}$
- Bayes' rule

$$\mathbb{P}(\mathcal{H}_i|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\mathcal{H}_i)\,\mathbb{P}(\mathcal{H}_i)}{\mathbb{P}(\mathcal{D})}$$

- This is a very important rule that you should learn
- It has a long and controversial history

Bayes Rule

- Bayes' rule

$$\mathbb{P}(\mathcal{H}_i|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\mathcal{H}_i)\,\mathbb{P}(\mathcal{H}_i)}{\mathbb{P}(\mathcal{D})}$$

  - $\mathbb{P}(\mathcal{H}_i|\mathcal{D})$ is the **posterior** probability (i.e. the probability of $\mathcal{H}_i$ after we know the data)
  - $\mathbb{P}(\mathcal{D}|\mathcal{H}_i)$ is the **likelihood** of the data given the hypothesis. Note, that we calculated this from the forward problem
  - $\mathbb{P}(\mathcal{H}_i)$ is the **prior** probability (i.e. the probability of $\mathcal{H}_i$ before we know the data)
  - $\mathbb{P}(\mathcal{D})$ is the **evidence**. It is a normalising constant given by

$$\mathbb{P}(\mathcal{D}) = \sum_{i=1}^{n} \mathbb{P}(\mathcal{H}_i, \mathcal{D})$$

Solving Inverse Problems

- Inference (the objective of machine learning) is an inverse problem
- We want the posterior $\mathbb{P}(\mathcal{H}_i|\mathcal{D})$ (i.e. the probability of what happened given some evidence)
- The Bayesian formalism converts this into the forward problem

$$\mathbb{P}(\mathcal{H}_i|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\mathcal{H}_i)\,\mathbb{P}(\mathcal{H}_i)}{\mathbb{P}(\mathcal{D})}$$

- We calculate the likelihood $\mathbb{P}(\mathcal{D}|\mathcal{H}_i)$ (i.e. assuming the hypothesis what is the chance of obtaining the data)
- But we also need to know the prior $\mathbb{P}(\mathcal{H}_i)$

Probabilistic Classification

- Establishing a probabilistic model for classification
  - Discriminative model

$$P(C \mid \mathbf{X}) \quad C = c_1, \cdots, c_L, \mathbf{X} = (X_1, \cdots, X_n)$$

  - Generative model

$$P(\mathbf{X} \mid C) \quad C = c_1, \cdots, c_L, \mathbf{X} = (X_1, \cdots, X_n)$$

- MAP classification rule
  - MAP: Maximum A Posterior
  - Assign x to c^* if $P(C = c^* \mid \mathbf{X} = \mathbf{x}) > P(C = c \mid \mathbf{X} = \mathbf{x}) \quad c \neq c^*, \; c = c_1, \cdots, c_L$
- Generative classification with the MAP rule
  - Apply Bayesian rule to convert $P(C \mid \mathbf{X}) = \dfrac{P(\mathbf{X} \mid C)P(C)}{P(\mathbf{X})} \propto P(\mathbf{X} \mid C)P(C)$

Naïve Bayes

- Bayes classification

$$P(C \mid \mathbf{X}) \propto P(\mathbf{X} \mid C)P(C) = P(X_1, \cdots, X_n \mid C)P(C)$$

Difficulty: learning the joint probability $P(X_1, \cdots, X_n \mid C)$

- Naïve Bayes classification
  - Making the assumption that all input attributes are independent

$$\begin{aligned} P(X_1, X_2, \cdots, X_n \mid C) &= P(X_1 \mid X_2, \cdots, X_n; C)P(X_2, \cdots, X_n \mid C) \\ &= P(X_1 \mid C)P(X_2, \cdots, X_n \mid C) \\ &= P(X_1 \mid C)P(X_2 \mid C) \cdots P(X_n \mid C) \end{aligned}$$

  - MAP classification rule

$$[P(x_1 \mid c^*) \cdots P(x_n \mid c^*)]P(c^*) > [P(x_1 \mid c) \cdots P(x_n \mid c)]P(c), \quad c \neq c^*, c = c_1, \cdots, c_L$$

Naïve Bayes

- Naïve Bayes Algorithm (for discrete input attributes)
  - Learning Phase: Given a training set **S**,

For each target value of c_i $(c_i = c_1, \cdots, c_L)$

$\hat{P}(C = c_i) \leftarrow$ estimate $P(C = c_i)$ with examples in **S**;

For every attribute value a_{jk} of each attribute x_j $(j = 1, \cdots, n; k = 1, \cdots, N_j)$

$\hat{P}(X_j = a_{jk} \mid C = c_i) \leftarrow$ estimate $P(X_j = a_{jk} \mid C = c_i)$ with examples in **S**;

Output: conditional probability tables; for x_j, $N_j \times L$ elements

  - Test Phase: Given an unknown instance $\mathbf{X}' = (a_1', \cdots, a_n')$ ′

Look up tables to assign the label c^* to **X′** if

$$[\hat{P}(a_1' \mid c^*) \cdots \hat{P}(a_n' \mid c^*)]\hat{P}(c^*) > [\hat{P}(a_1' \mid c) \cdots \hat{P}(a_n' \mid c)]\hat{P}(c), \quad c \neq c^*, c = c_1, \cdots, c_L$$

Example

- Example: Play Tennis

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Example

- Learning Phase

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=*Yes*	Play=*No*
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

$P(\text{Play}=Yes) = 9/14$ $P(\text{Play}=No) = 5/14$

Example

- Test Phase
  - Given a new instance,

    $\mathbf{x}'$=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)
  - Look up tables

P(Outlook=*Sunny*|Play=*Yes*) = 2/9

P(Temperature=*Cool*|Play=*Yes*) = 3/9

P(Huminity=*High*|Play=*Yes*) = 3/9

P(Wind=*Strong*|Play=*Yes*) = 3/9

P(Play=*Yes*) = 9/14

P(Outlook=S*unny*|Play=*No*) = 3/5

P(Temperature=*Cool*|Play==*No*) = 1/5

P(Huminity=*High*|Play=*No*) = 4/5

P(Wind=*Strong*|Play=*No*) = 3/5

P(Play=*No*) = 5/14

  - MAP rule

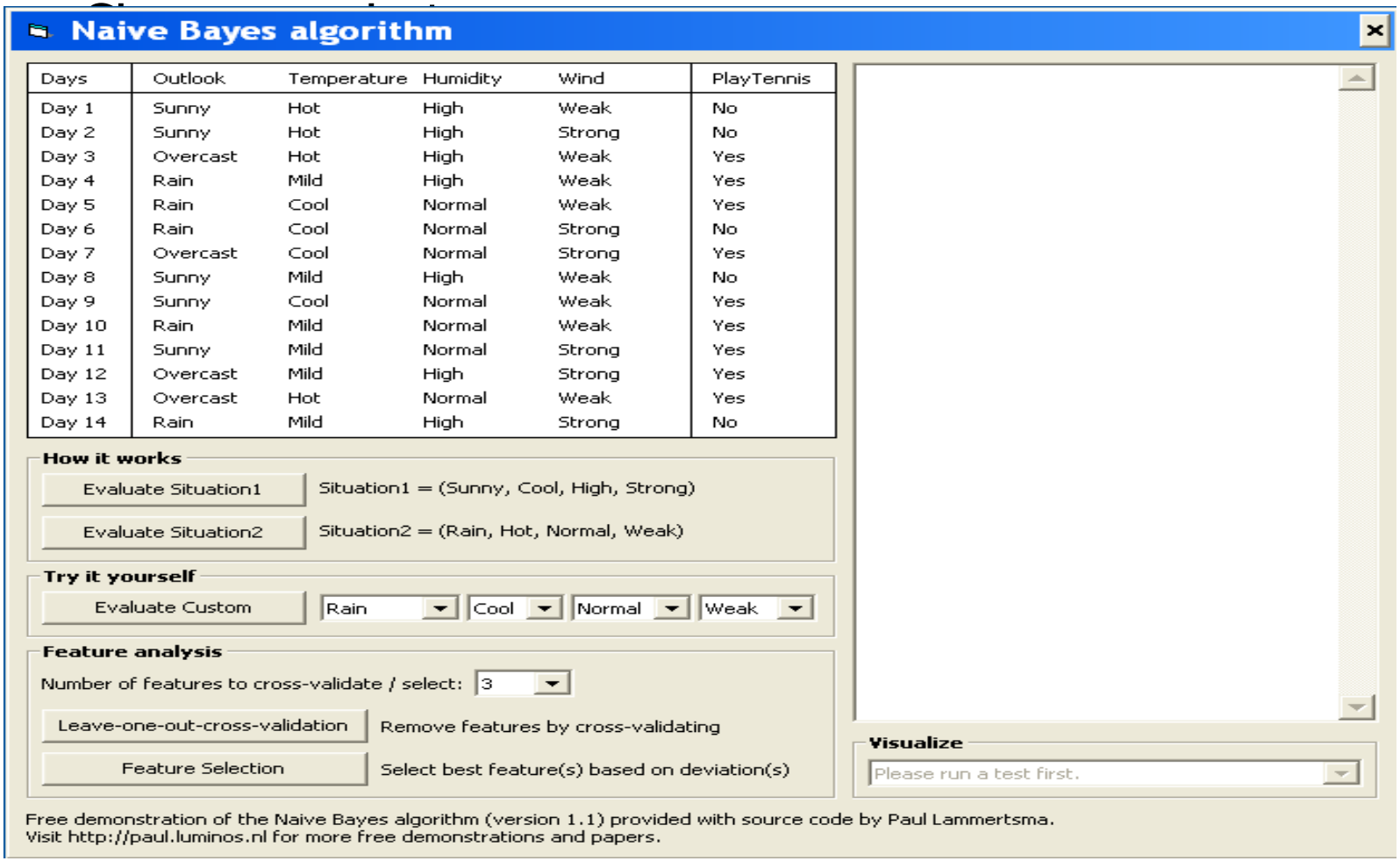
P(*Yes*|$\mathbf{x}'$): [P(*Sunny*|Yes)P(*Cool*|*Yes*)P(*High*|Yes)P(*Strong*|*Yes*)]P(Play=*Yes*) = 0.0053

P(*No*|$\mathbf{x}'$): [P(*Sunny*|No) P(*Cool*|No)P(*High*|*No*)P(*Strong*|*No*)]P(Play=*No*) = 0.0206

Given the fact P(*Yes*|$\mathbf{x}'$) < P(*No*|$\mathbf{x}'$), we label $\mathbf{x}'$ to be "*No*".

Example

- Test Phase



Given the fact $P(Yes|\mathbf{x}') < P(No|\mathbf{x}')$, we label $\mathbf{x}'$ to be "*No*".

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 . 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

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

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

Officer Drew is more likely to be a Female.

Officer Drew

Officer Drew IS a female

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

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

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

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

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

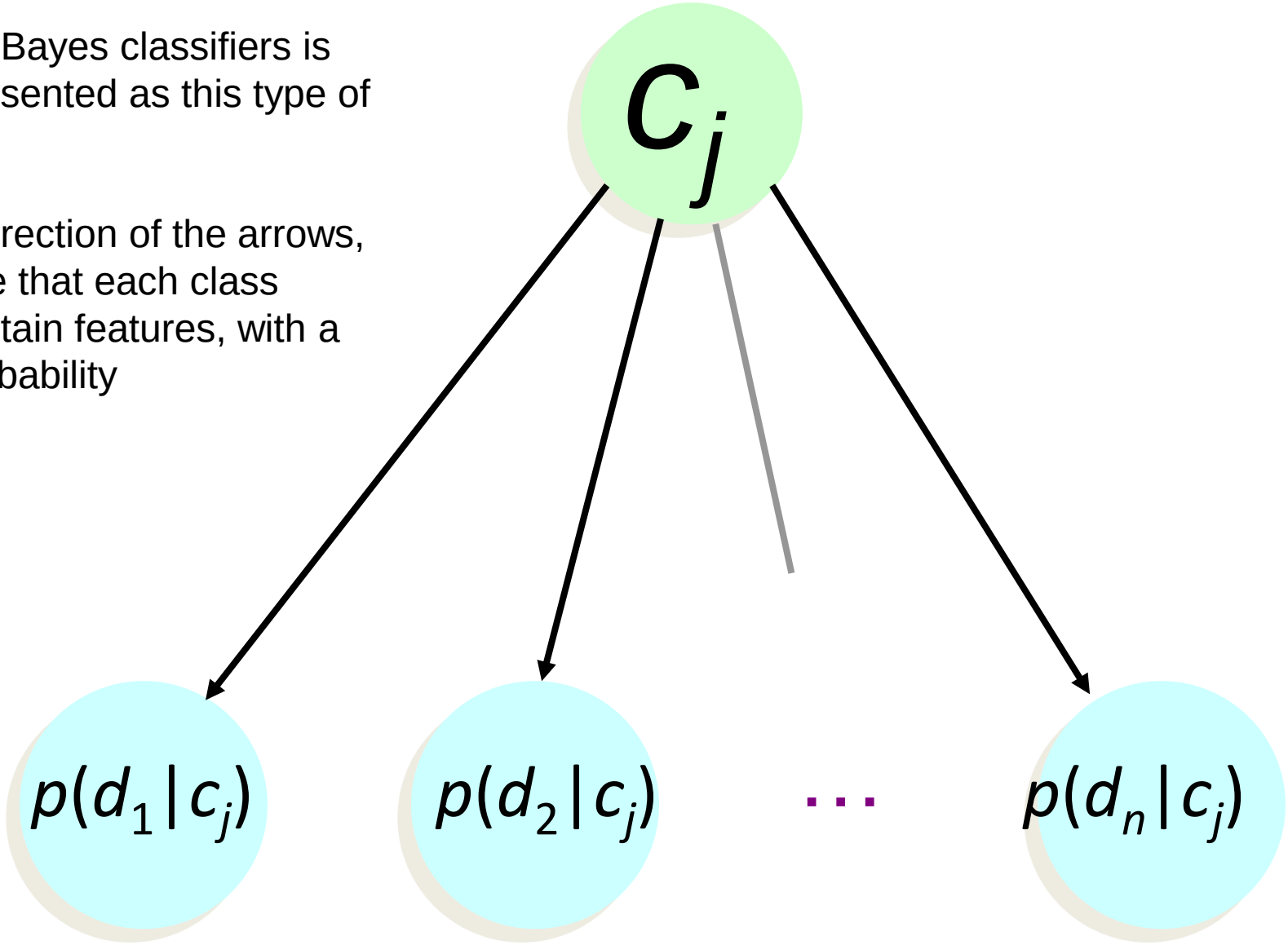
Officer Drew is blue-eyed, over 170_{cm} tall, and has long hair

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

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

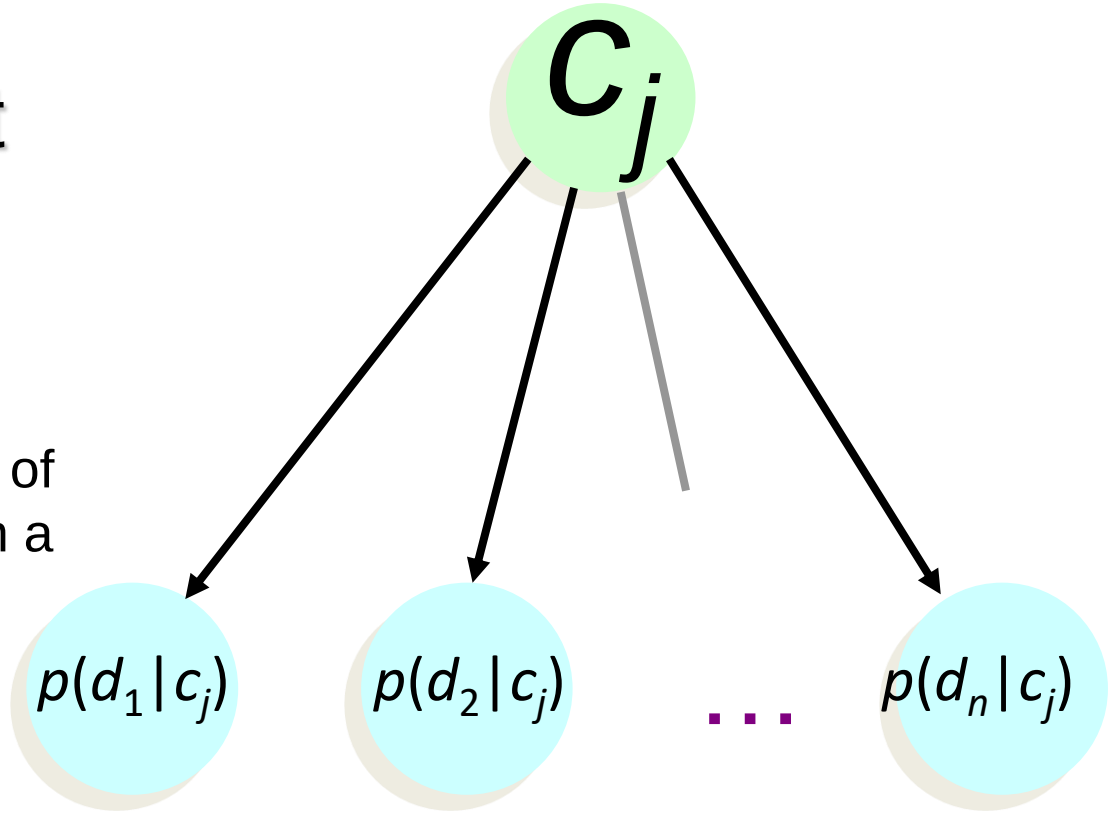
The Naive Bayes classifiers is often represented as this type of graph…

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	Over190$_{cm}$	
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	

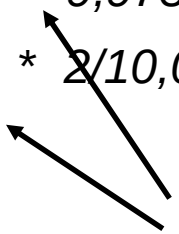
Naïve Bayes is NOT sensitive to irrelevant features...

Suppose we are trying to classify a persons sex based on several features, including eye color. (Of course, eye color is completely irrelevant to a persons gender)

$p(\text{Jessica} \mid c_j) = p(\text{eye} = \text{brown} \mid c_j) * p(\text{wears\_dress} = \text{yes} \mid c_j) * \ldots$

$p(\text{Jessica} \mid \text{Female}) = 9{,}000/10{,}000 * 9{,}975/10{,}000 * \ldots$

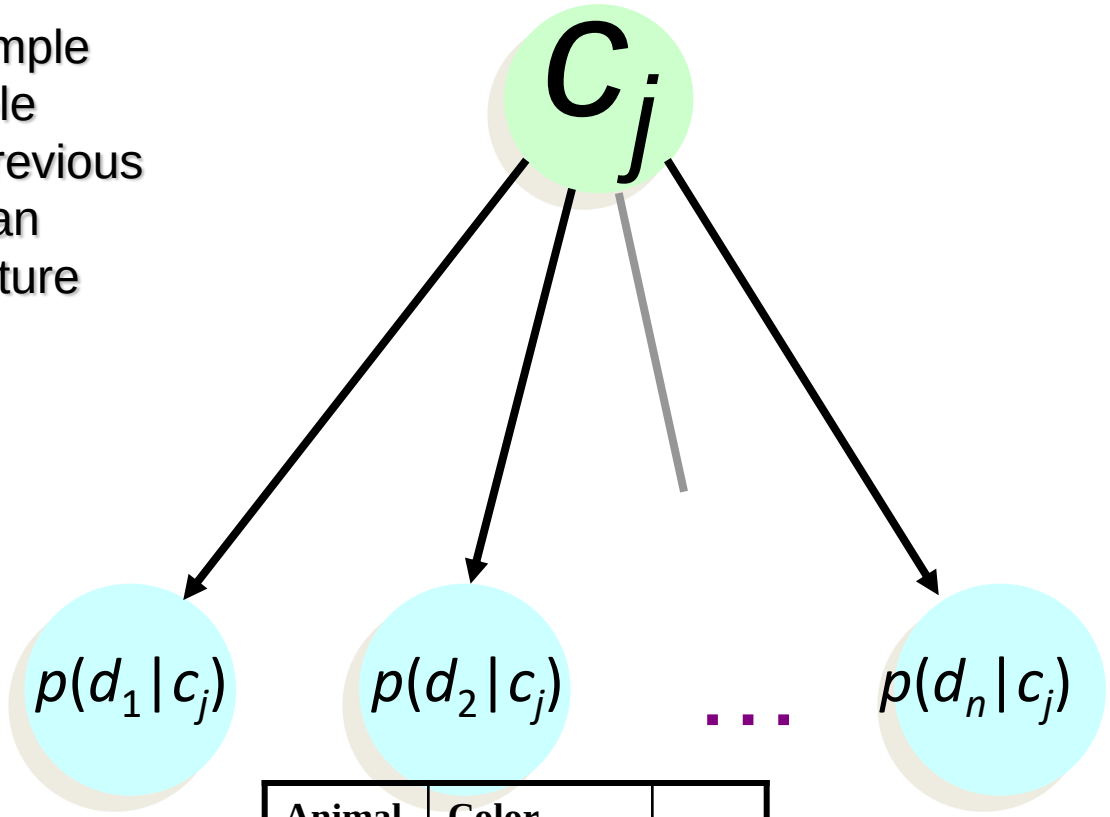
$p(\text{Jessica} \mid \text{Male}) = 9{,}001/10{,}000 * 2/10{,}000 * \ldots$



Almost the same!

However, this assumes that we have good enough estimates of the probabilities, so the more data the better.

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

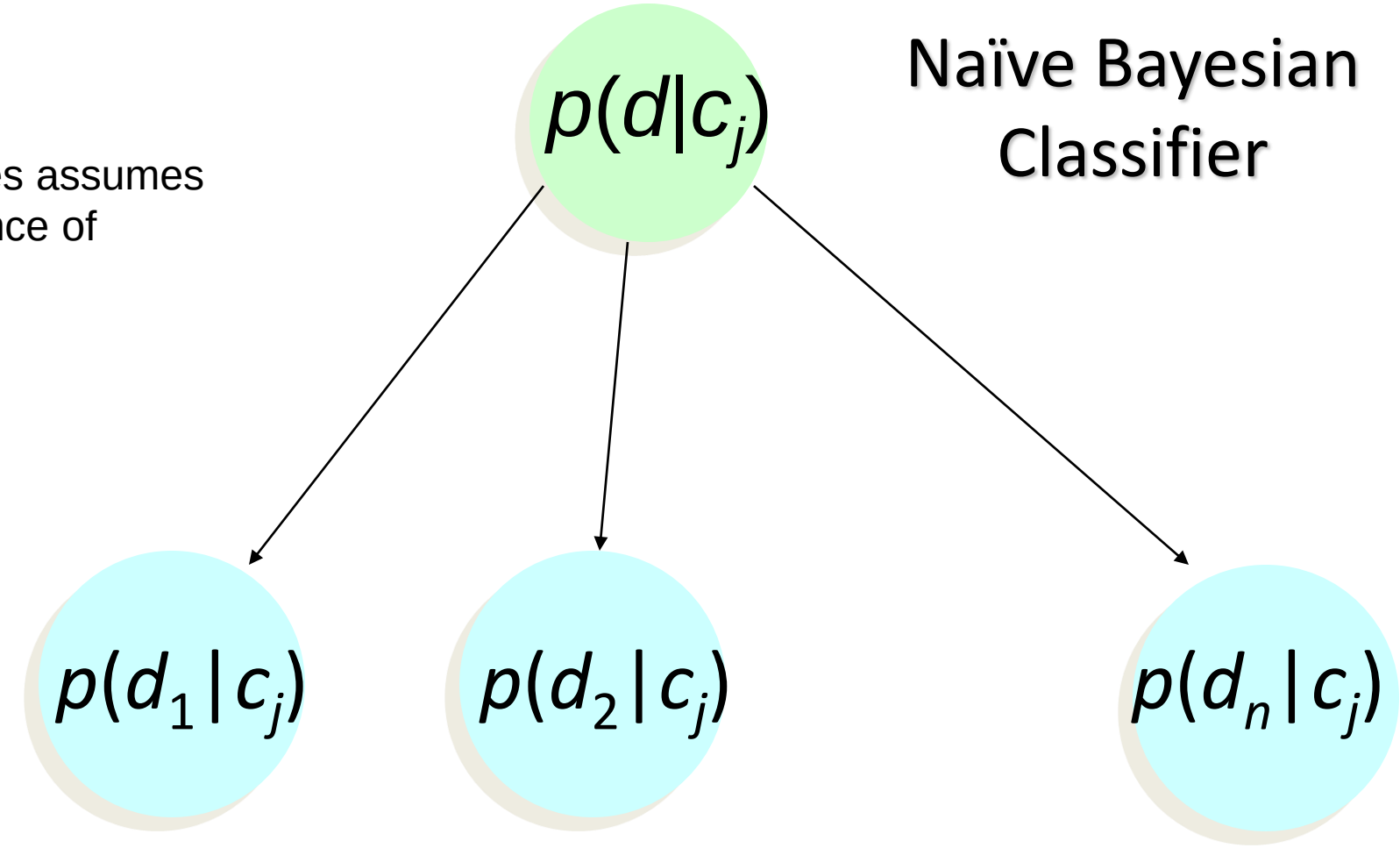
Animal	Color	
Cat	Black	0.33
	White	0.23
	Brown	0.44
Dog	Black	0.97
	White	0.03
	Brown	0.90
Pig	Black	0.04
	White	0.01
	Brown	0.95

Animal
Cat
Dog
Pig

Naïve Bayesian Classifier

Problem!

Naïve Bayes assumes independence of features…



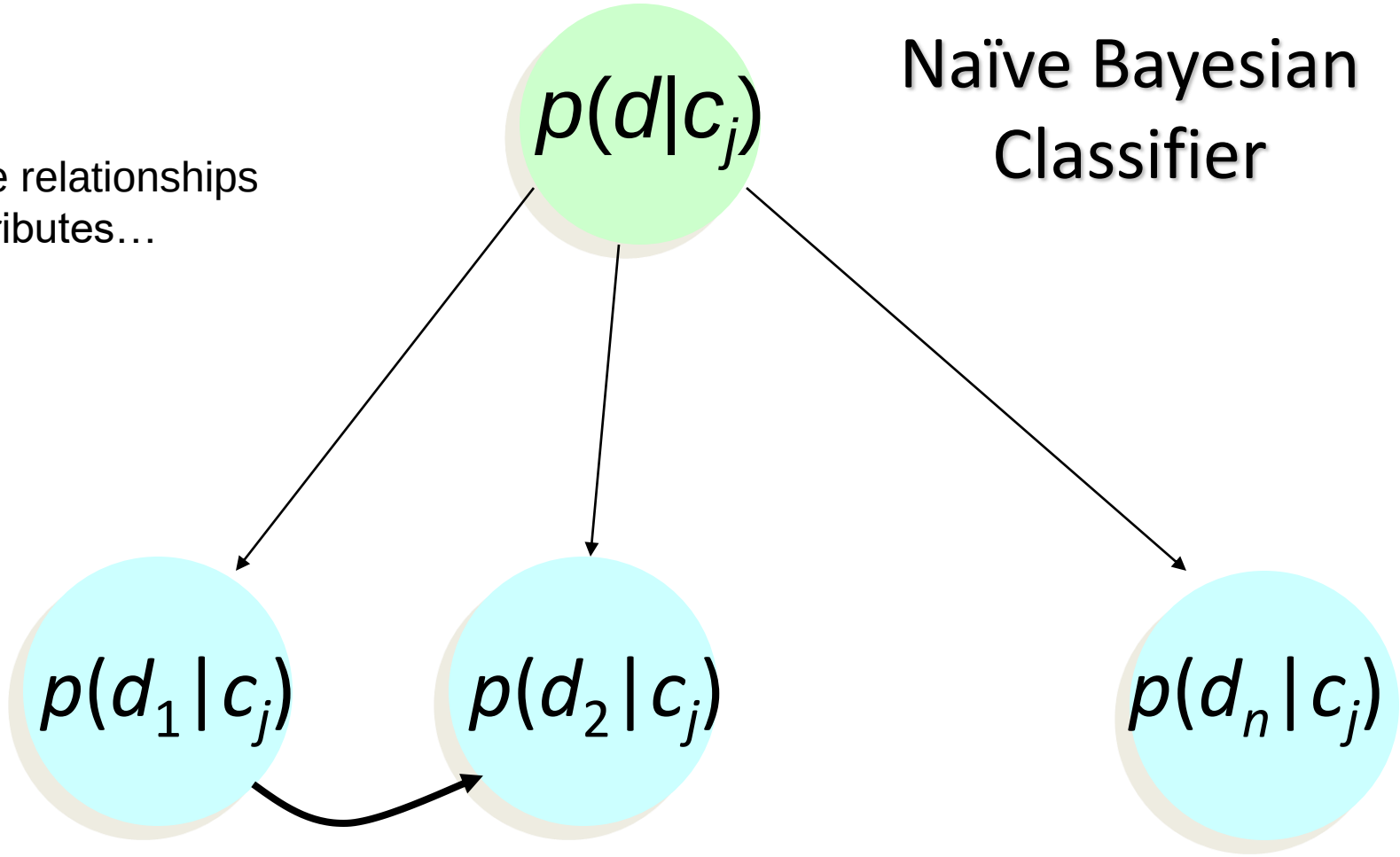
Sex	Over 6 foot	
Male	Yes	0.15
	No	0.85
Female	Yes	0.01
	No	0.99

Sex	Over 200 pounds	
Male	Yes	0.11
	No	0.80
Female	Yes	0.05
	No	0.95

Solution

Consider the relationships between attributes…

Naïve Bayesian Classifier



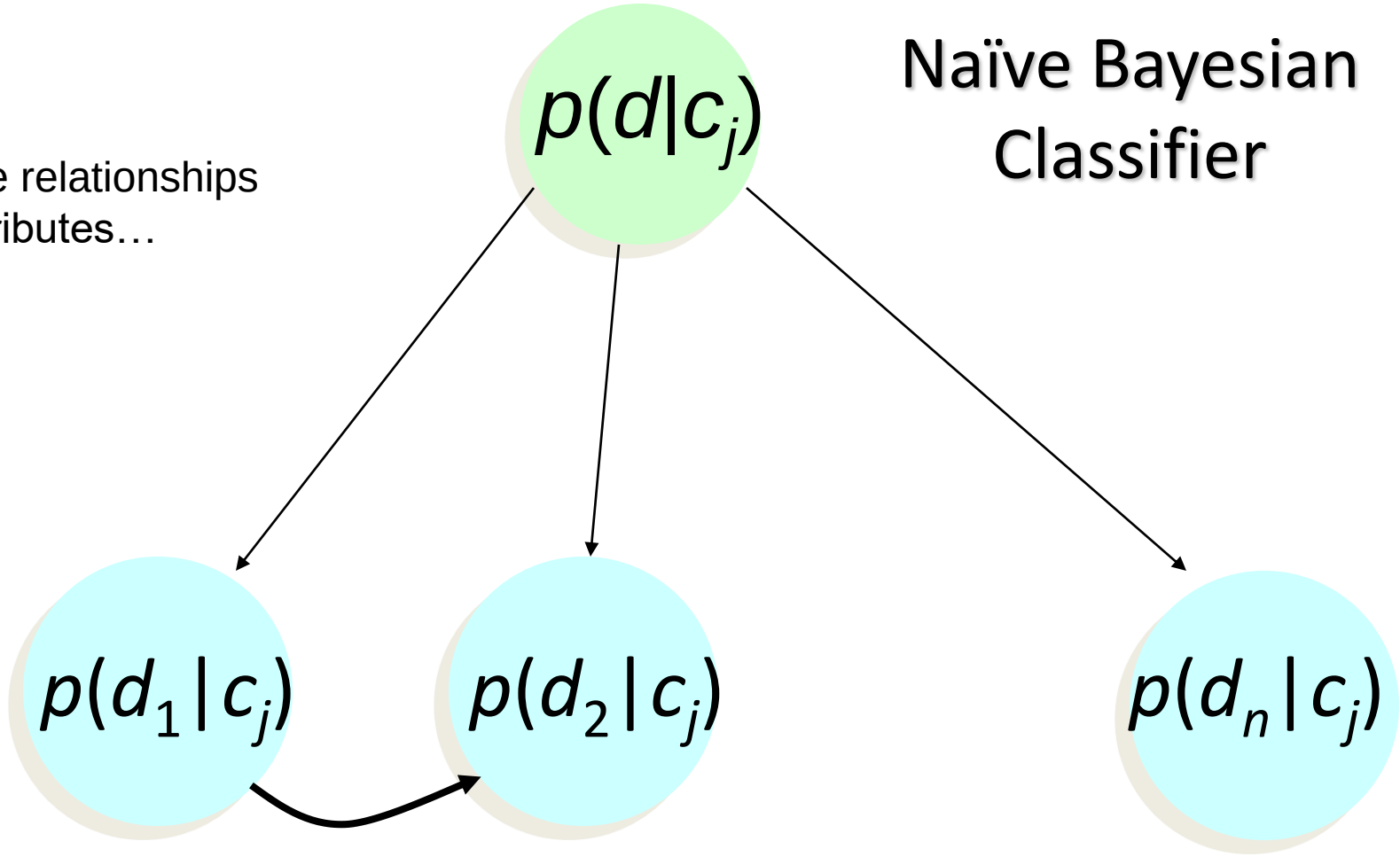
Sex	**Over 6 foot**	
Male	Yes	0.15
	No	0.85
Female	Yes	0.01
	No	0.99

Sex	**Over 200 pounds**	
Male	Yes and **Over 6 foot**	0.11
	No and **Over 6 foot**	0.59
	Yes and NOT **Over 6 foot**	0.05
	No and NOT **Over 6 foot**	0.35
Female	Yes and **Over 6 foot**	0.01

Solution

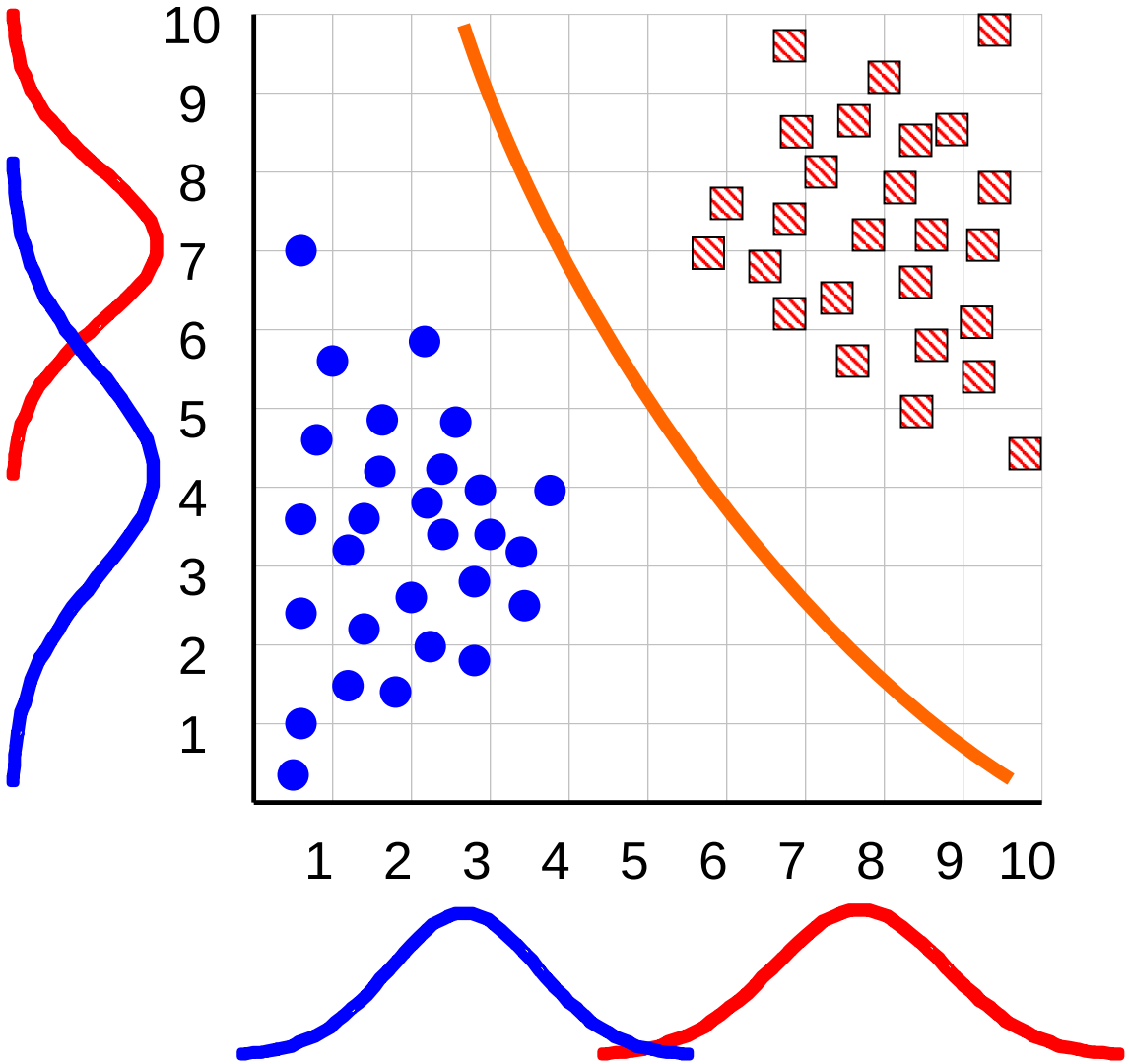
Consider the relationships between attributes…

Naïve Bayesian Classifier



But how do we find the set of connecting arcs??

The Naïve Bayesian Classifier has a quadratic decision boundary



Relevant Issues

- Violation of Independence Assumption
  - For many real world tasks, $P(X_1,\cdots,X_n \mid C) \neq P(X_1 \mid C)\cdots P(X_n \mid C)$
  - Nevertheless, naïve Bayes works surprisingly well anyway!
- Zero conditional probability Problem
  - If no example contains the attribute value $X_j = a_{jk}$, $\hat{P}(X_j = a_{jk} \mid C = c_i) = 0$
  - In this circumstance, $\hat{P}(x_1 \mid c_i)\cdots\hat{P}(a_{jk} \mid c_i)\cdots\hat{P}(x_n \mid c_i) = 0$ during test
  - For a remedy, conditional probabilities estimated with

$$\hat{P}(X_j = a_{jk} \mid C = c_i) = \frac{n_c + mp}{n + m}$$

n_c : number of training examples for which $X_j = a_{jk}$ and $C = c_i$

n : number of training examples for which $C = c_i$

p : prior estimate (usually, $p = 1/t$ for t possible values of X_j)

m : weight to prior (number of "virtual" examples, $m \geq 1$)

Estimating Probabilities

- Normally, probabilities are estimated based on observed frequencies in the training data.
- If D contains n_k examples in category y_k, and n_{ijk} of these n_k examples have the jth value for feature X_i, x_{ij}, then:

$$P(X_i = x_{ij} \mid Y = y_k) = \frac{n_{ijk}}{n_k}$$

- However, estimating such probabilities from small training sets is error-prone.
  - If due only to chance, a rare feature, X_i, is always false in the training data, $\forall y_k : P(X_i = \text{true} \mid Y = y_k) = 0$.
  - If X_i=true then occurs in a test example, X, the result is that $\forall y_k$: $P(X \mid Y = y_k) = 0$ and $\forall y_k$: $P(Y = y_k \mid X) = 0$

Probability Estimation Example

Ex	Size	Color	Shape	Category
1	small	red	circle	positive
2	large	red	circle	positive
3	small	red	triangle	negitive
4	large	blue	circle	negitive

Probability	positive	negative
P(*Y*)	0.5	0.5
P(small | *Y*)	0.5	0.5
P(medium | *Y*)	0.0	0.0
P(large | *Y*)	0.5	0.5
P(red | *Y*)	1.0	0.5
P(blue | *Y*)	0.0	0.5
P(green | *Y*)	0.0	0.0
P(square | *Y*)	0.0	0.0
P(triangle | *Y*)	0.0	0.5
P(circle | *Y*)	1.0	0.5

Test Instance *X*:
<medium, red, circle>

P(positive | *X*) = 0.5 * 0.0 * 1.0 * 1.0 / P(X) = 0

P(negative | *X*) = 0.5 * 0.0 * 0.5 * 0.5 / P(X) = 0

Naïve Bayes Example

Probability	positive	negative
P(*Y*)	0.5	0.5
P(small | *Y*)	0.4	0.4
P(medium | *Y*)	0.1	0.2
P(large | *Y*)	0.5	0.4
P(red | *Y*)	0.9	0.3
P(blue | *Y*)	0.05	0.3
P(green | *Y*)	0.05	0.4
P(square | *Y*)	0.05	0.4
P(triangle | *Y*)	0.05	0.3
P(circle | *Y*)	0.9	0.3

Test Instance:
<medium ,red, circle>

49

Naïve Bayes Example

Probability	positive	negative
P(Y)	0.5	0.5
P(medium | Y)	0.1	0.2
P(red | Y)	0.9	0.3
P(circle | Y)	0.9	0.3

Test Instance:
<medium ,red, circle>

P(positive | X) = P(positive)*P(medium | positive)*P(red | positive)*P(circle | positive) / P(X)

0.5 * 0.1 * 0.9 * 0.9

= 0.0405 / P(X) = 0.0405 / 0.0495 = 0.8181

P(negative | X) = P(negative)*P(medium | negative)*P(red | negative)*P(circle | negative) / P(X)

0.5 * 0.2 * 0.3 * 0.3

= 0.009 / P(X) = 0.009 / 0.0495 = 0.1818

P(positive | X) + P(negative | X) = 0.0405 / P(X) + 0.009 / P(X) = 1

P(X) = (0.0405 + 0.009) = 0.0495

Smoothing

- To account for estimation from small samples, probability estimates are adjusted or *smoothed*.
- Laplace smoothing using an *m*-estimate assumes that each feature is given a prior probability, *p*, that is assumed to have been previously observed in a "virtual" sample of size *m*.

$$P(X_i = x_{ij} \mid Y = y_k) = \frac{n_{ijk} + mp}{n_k + m}$$

- For binary features, *p* is simply assumed to be 0.5.

Laplace Smothing Example

- Assume training set contains 10 positive examples:
  - 4: small
  - 0: medium
  - 6: large
- Estimate parameters as follows (if m=1, p=1/3)
  - P(small | positive) = (4 + 1/3) / (10 + 1) = 0.394
  - P(medium | positive) = (0 + 1/3) / (10 + 1) = 0.03
  - P(large | positive) = (6 + 1/3) / (10 + 1) = 0.576
  - P(small or medium or large | positive) = 1.0

52

Numerical Stability

- It is often the case that machine learning algorithms need to work with very small numbers
  - Imagine computing the probability of 2000 independent coin flips
  - MATLAB thinks that $(.5)^{2000}=0$

Stochastic Language Models

- Models *probability* of generating strings (each word in turn) in the language (commonly all strings over ∑). E.g., unigram model

Model M

0.2	the
0.1	a
0.01	guy
0.01	fruit
0.03	said
0.02	likes

the	guy	likes	the	fruit
0.2	0.01	0.02	0.2	0.01

multiply

P(s | M) = 0.00000008

Numerical Stability

- Instead of comparing $P(Y=5|X_1,\ldots,X_n)$ with $P(Y=6|X_1,\ldots,X_n)$,
  - Compare their logarithms

$$\begin{aligned}
\log\left(P(Y|X_1,\ldots,X_n)\right) &= \log\left(\frac{P(X_1,\ldots,X_n|Y)\cdot P(Y)}{P(X_1,\ldots,X_n)}\right) \\
&= \text{constant} + \log\left(\prod_{i=1}^{n} P(X_i|Y)\right) + \log P(Y) \\
&= \text{constant} + \sum_{i=1}^{n} \log P(X_i|Y) + \log P(Y)
\end{aligned}$$

Underflow Prevention

- Multiplying lots of probabilities, which are between 0 and 1 by definition, can result in floating-point underflow.
- Since $\log(xy) = \log(x) + \log(y)$, it is better to perform all computations by summing logs of probabilities rather than multiplying probabilities.
- Class with highest final un-normalized log probability score is still the most probable.

Relevant Issues

- Continuous-valued Input Attributes
  - Numberless values for an attribute
  - Conditional probability modeled with the normal distribution

$$\hat{P}(X_j \mid C = c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp\left(-\frac{(X_j - \mu_{ji})^2}{2\sigma_{ji}^2} \right)$$

μ_{ji} : mean (avearage) of attribute values X_j of examples for which $C = c_i$

σ_{ji} : standard deviation of attribute values X_j of examples for which $C = c_i$

  - Learning Phase: for $\mathbf{X} = (X_1, \cdots, X_n)$, $C = c_1, \cdots, c_L$
    Output: $n \times L$ normal distributions and $P(C = c_i)$ $i = 1, \cdots, L$
  - Test Phase: for $\mathbf{X}' = (X_1', \cdots, X_n')$
    - Calculate conditional probabilities with all the normal distributions
    - Apply the MAP rule to make a decision

Data with Numeric Attributes

sex	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

sex	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033e-02	176.25	1.2292e+02	11.25	9.1667e-01
female	5.4175	9.7225e-02	132.5	5.5833e+02	7.5	1.6667e+00

Data with Numeric Attributes

sex	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

$$posterior(male) = \frac{P(male)\, p(height|male)\, p(weight|male)\, p(footsize|male)}{evidence}$$

$$posterior(female) = \frac{P(female)\, p(height|female)\, p(weight|female)\, p(footsize|female)}{evidence}$$

$$evidence = P(male)\, p(height|male)\, p(weight|male)\, p(footsize|male)$$
$$+P(female)\, p(height|female)\, p(weight|female)\, p(footsize|female)$$

$$P(male) = 0.5$$

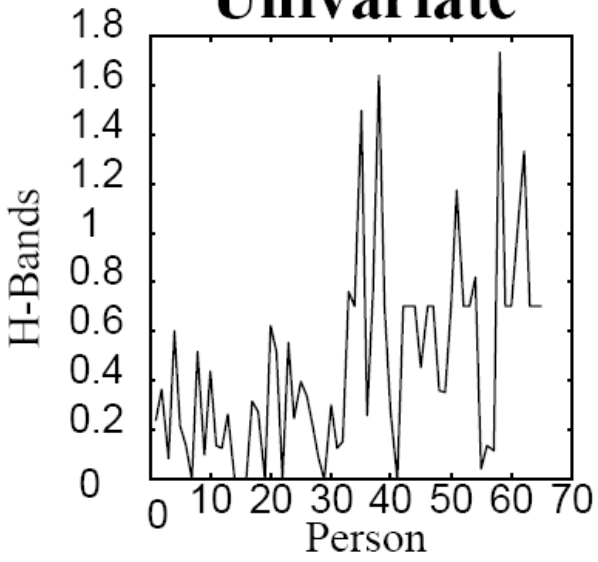
$$p(\text{height}|\text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) \approx 1.5789,$$

$$p(\text{weight}|\text{male}) = 5.9881 \cdot 10^{-6}$$
$$p(\text{foot size}|\text{male}) = 1.3112 \cdot 10^{-3}$$
$$\text{posterior numerator (male)} = \text{their product} = 6.1984 \cdot 10^{-9}$$

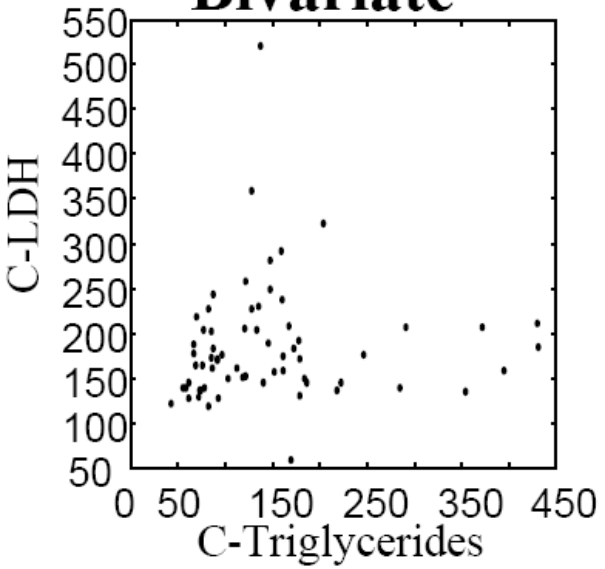
$$P(\text{female}) = 0.5$$
$$p(\text{height}|\text{female}) = 2.2346 \cdot 10^{-1}$$
$$p(\text{weight}|\text{female}) = 1.6789 \cdot 10^{-2}$$
$$p(\text{foot size}|\text{female}) = 2.8669 \cdot 10^{-1}$$
$$\text{posterior numerator (female)} = \text{their product} = 5.3778 \cdot 10^{-4}$$

Since posterior numerator is greater in the female case, we predict the sample is female.

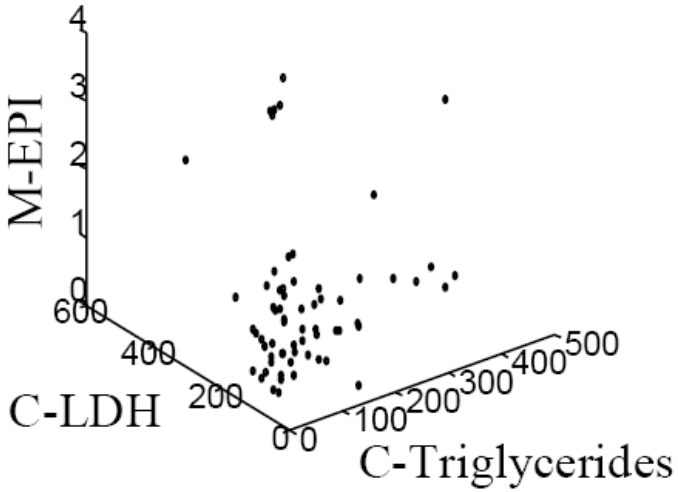
Univariate



Bivariate



Trivariate



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

Conclusions

- Naïve Bayes based on the independence assumption
  - Training is very easy and fast; just requiring considering each attribute in each class separately
  - Test is straightforward; just looking up tables or calculating conditional probabilities with normal distributions
- A popular generative model
  - Performance competitive to most of state-of-the-art classifiers even in presence of violating independence assumption
  - Many successful applications, e.g., spam mail filtering
  - Apart from classification, naïve Bayes can do more…

Conclusions

- Naïve Bayes is:
  - Really easy to implement and often works well
  - Often a good first thing to try
  - Commonly used as a “punching bag” for smarter algorithms

Acknowledgements

- Introduction to Machine Learning, Alphaydin
- Statistical Pattern Recognition: A Review – A.K Jain et al., PAMI (22) 2000
- Pattern Recognition and Analysis Course – A.K. Jain, MSU
- *Pattern Classification"* by Duda et al., John Wiley & Sons.
- http://www.doc.ic.ac.uk/~sgc/teaching/pre2012/v231/lecture13.html
- http://en.wikipedia.org/wiki/Naive_Bayes_classifier
- Some Material adopted from Dr. Adam Prugel-Bennett

Material in these slides has been taken from, the following resources