CS 484
Data Mining

Classification 6

Some slides are from Professor Eamonn Keogh at UC Riverside

Instance-Based Classifiers

Set of Stored Cases

Atr1	………	AtrN	Class
			A
			B
			B
			C
			A
			C
			B

- Store the training records
- Use training records to predict the class label of unseen cases

Unseen Case

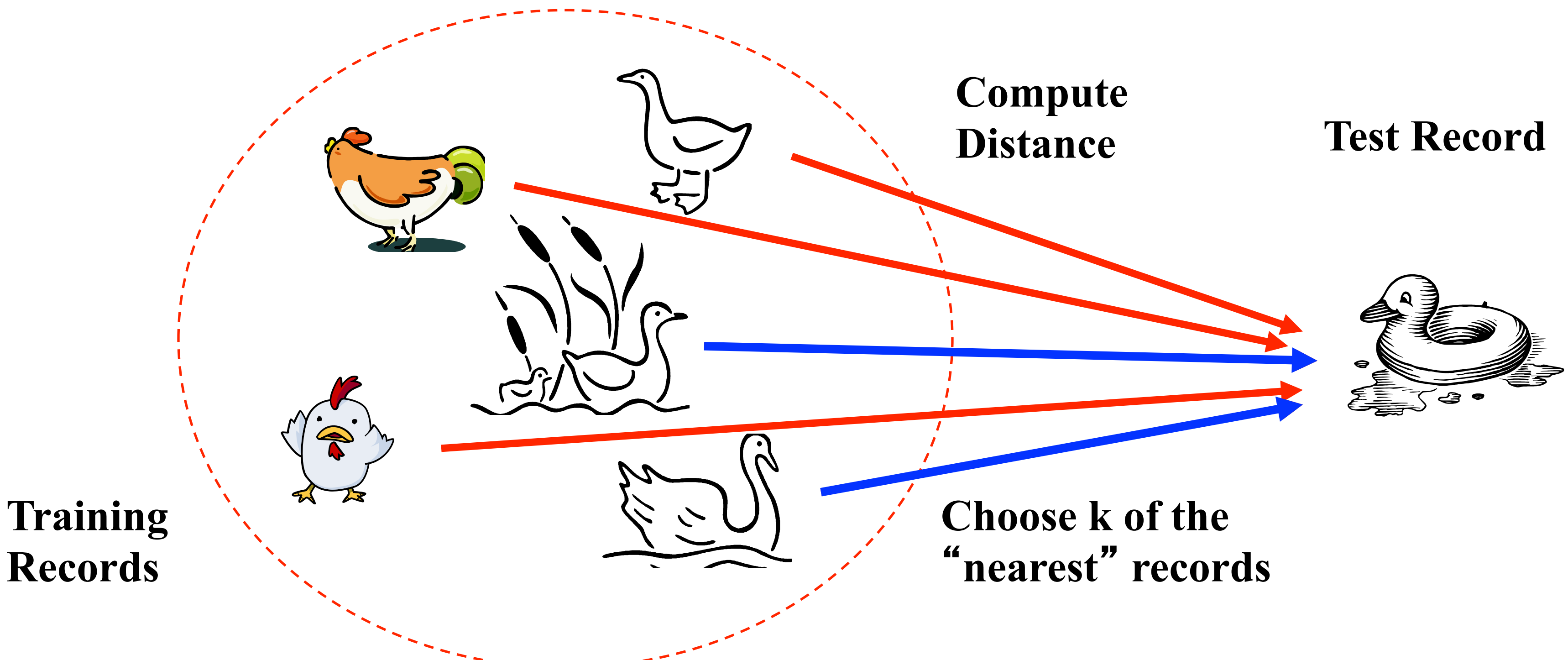
Atr1	………	AtrN

Instance Based Classifiers

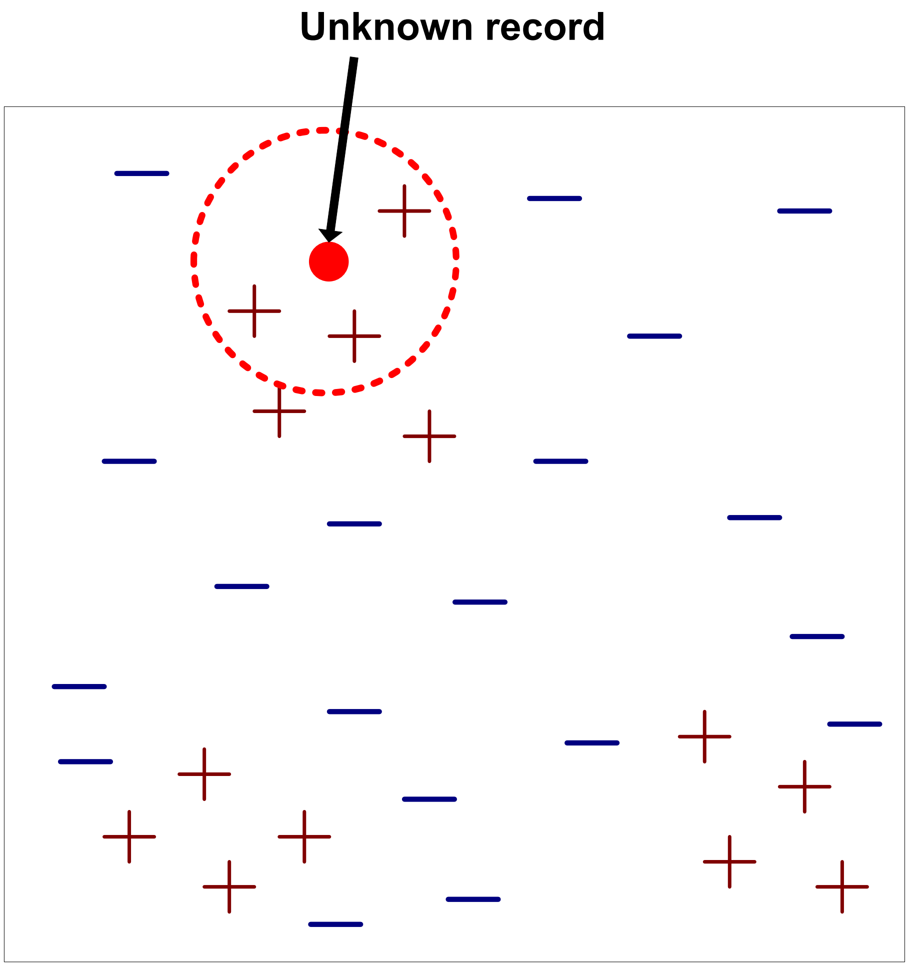
- Examples:
  - Rote-learner
    - Memorizes entire training data and performs classification only if attributes of record match one of the training examples exactly
  - Nearest neighbor
    - Uses the “closest” points (nearest neighbors) for performing classification

Nearest Neighbor Classifiers

- Basic idea:
  - If it walks like a duck, quacks like a duck, then it’s probably a duck



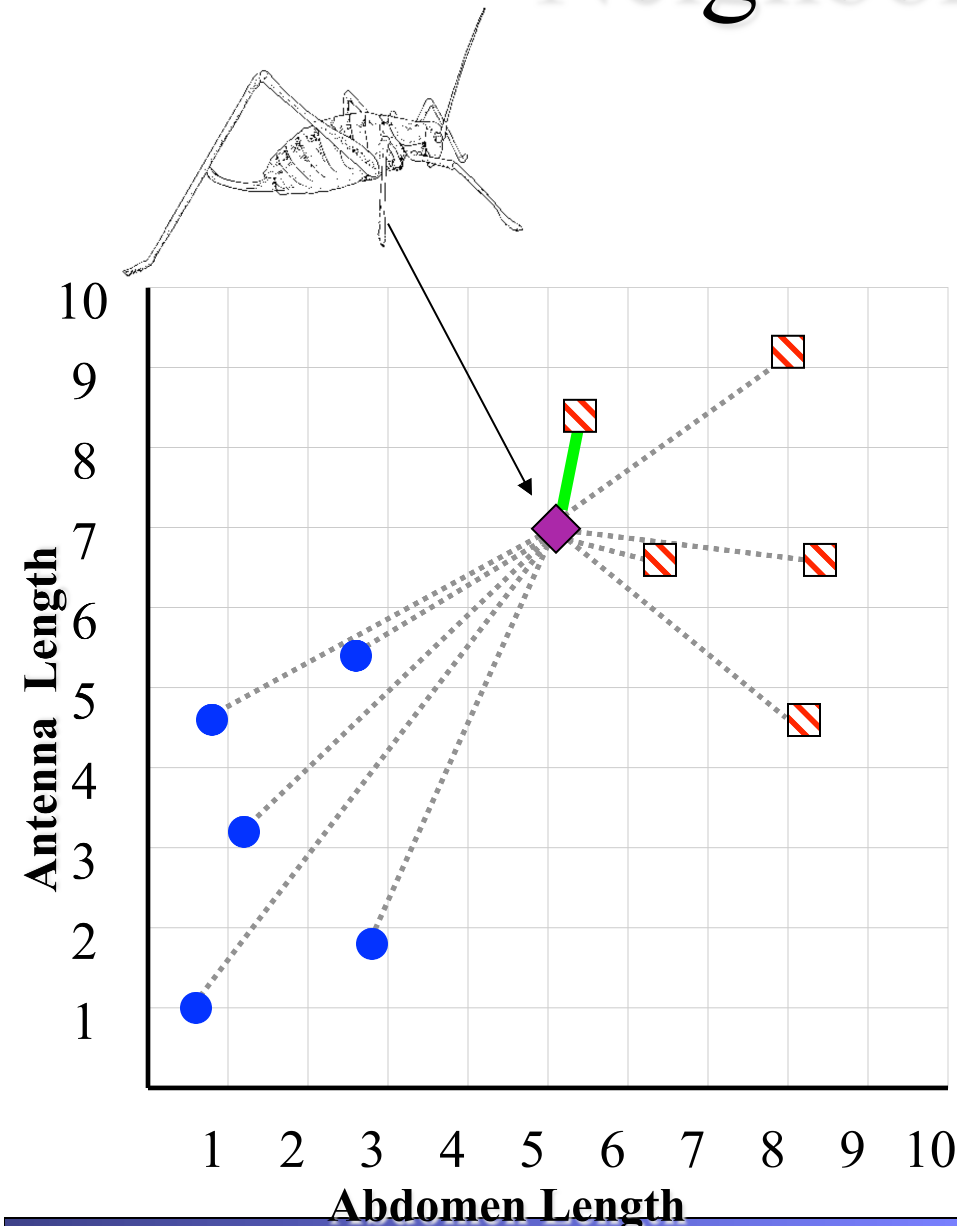
Nearest-Neighbor Classifiers



- Requires three things
  - The set of stored records
  - Distance Metric to compute distance between records
  - The value of *k*, the number of nearest neighbors to retrieve

- To classify an unknown record:
  - Compute distance to other training records
  - Identify *k* nearest neighbors
  - Use class labels of nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)

Nearest Neighbor Classifiers



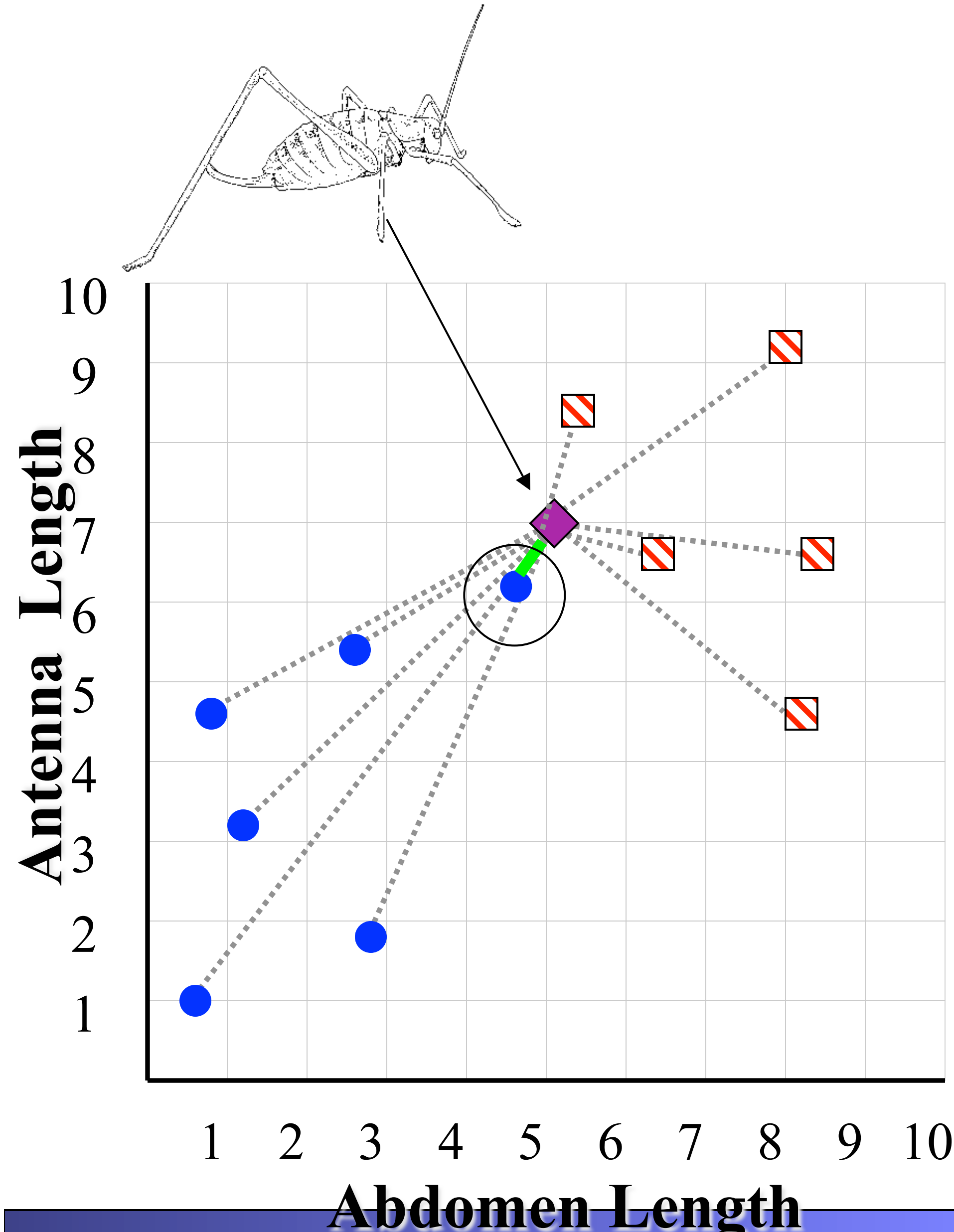
Evelyn Fix
1904-1965

Joe Hodges
1922-2000

If the **nearest** instance to the previously unseen instance **is a Katydid**
class is **Katydid**
else
class is **Grasshopper**

Katydids
Grasshoppers

Nearest Neighbor Classifier is sensitive to outliers

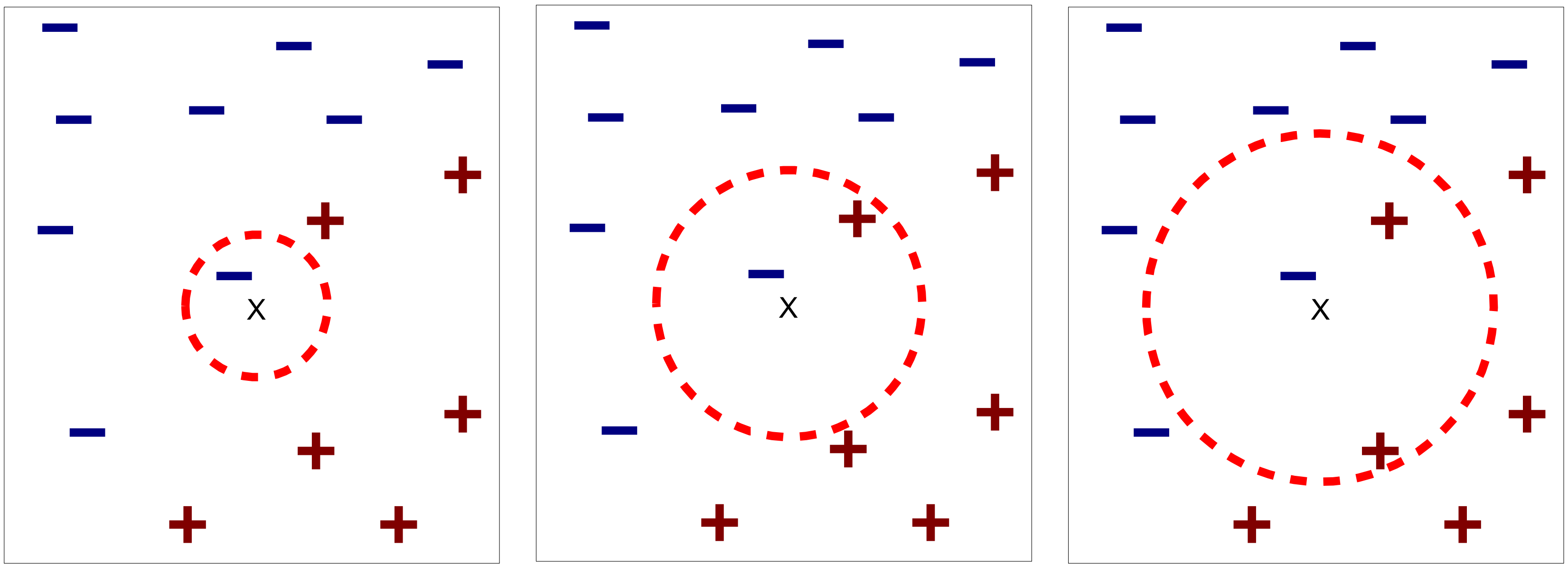


If the **nearest** instance to the previously unseen instance **is a Katydid**
    class is **Katydid**
else
    class is **Grasshopper**

Katydids
Grasshoppers

Solution: Use K nearest neighbors instead, and take majority vote!

Definition of Nearest Neighbor

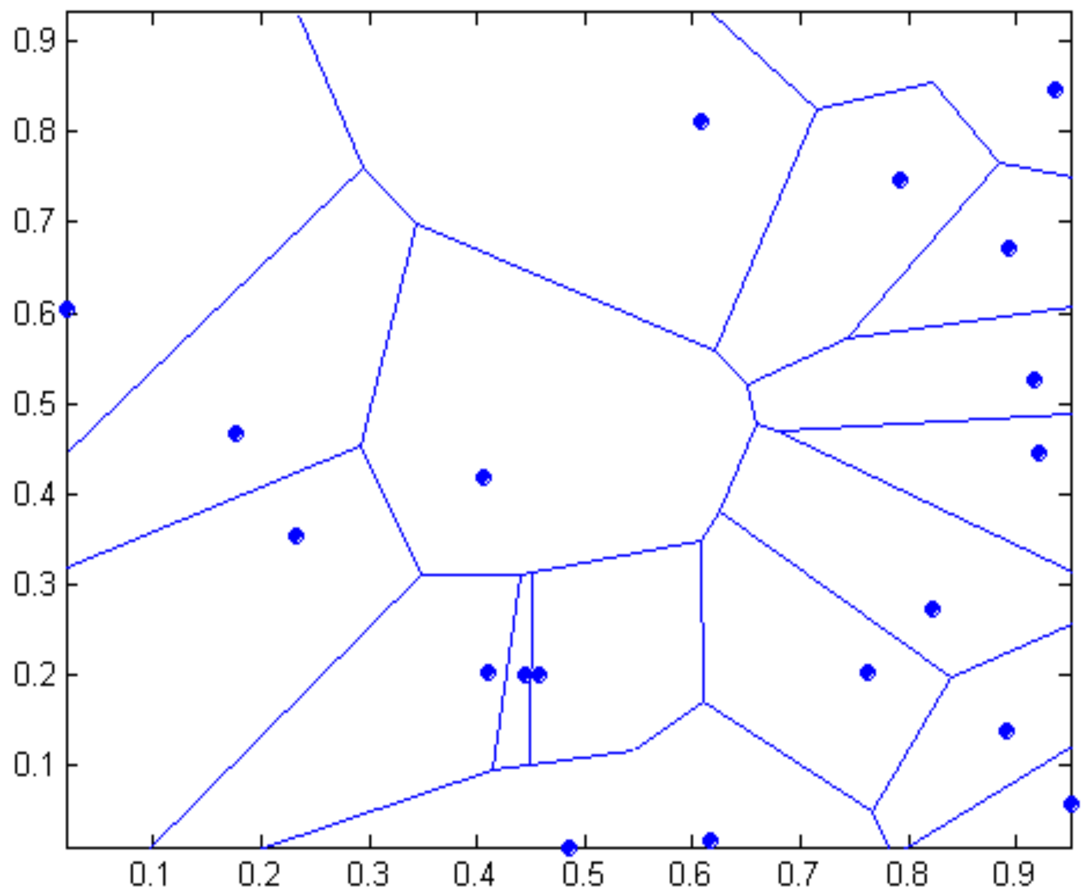


(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-nearest neighbors of a record x are data points that have the k smallest distance to x

1-nearest-neighbor

Voronoi Diagram



Nearest Neighbor Classification

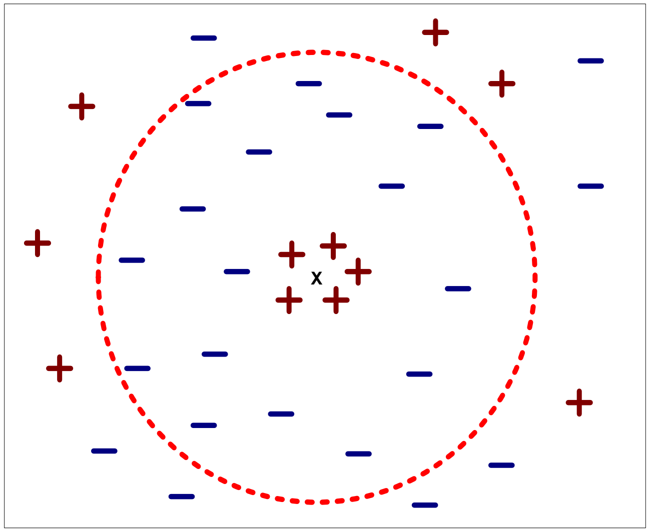
- Compute distance between two points:
  - Euclidean distance

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

- Determine the class from nearest neighbor list
  - take the majority vote of class labels among the k-nearest neighbors
  - Weigh the vote according to distance
    - weight factor, $w = 1/d^2$

Nearest Neighbor Classification…

- Choosing the value of k:
  - If k is too small, sensitive to noise points
  - If k is too large, neighborhood may include points from other classes
  - What if we have a tie?

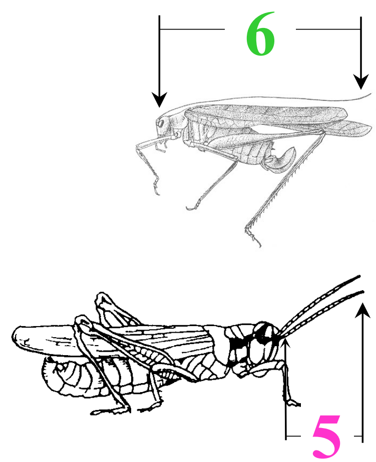


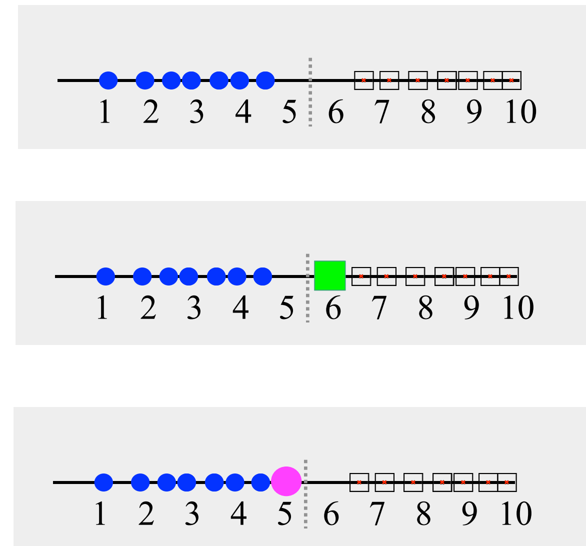
The nearest neighbor algorithm is sensitive to irrelevant features…

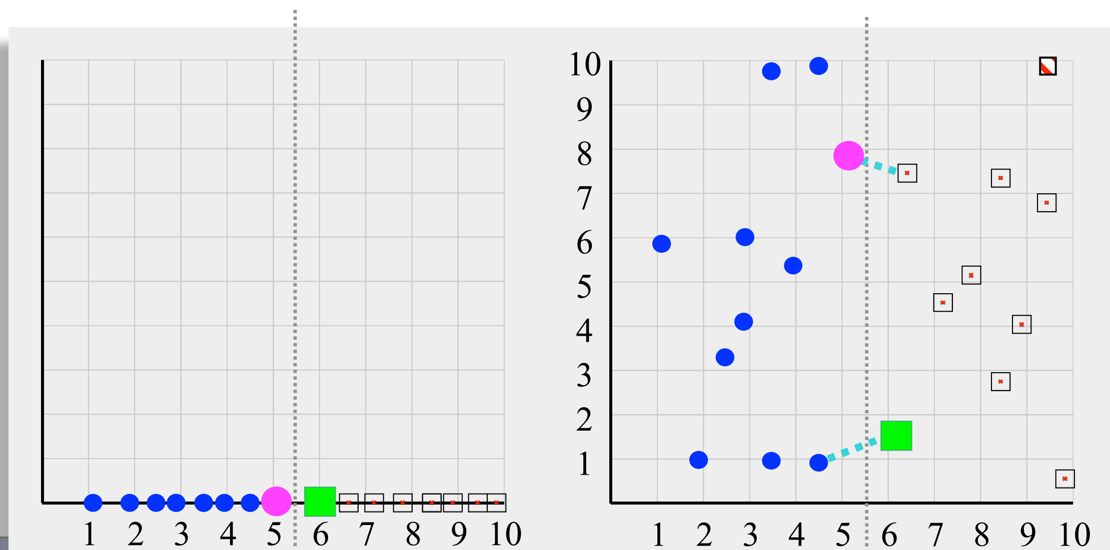
Suppose the following is true, if an insects antenna is longer than 5.5 it is a **Katydid**, otherwise it is a **Grasshopper**.

Using just the antenna length we get perfect classification!

Training data







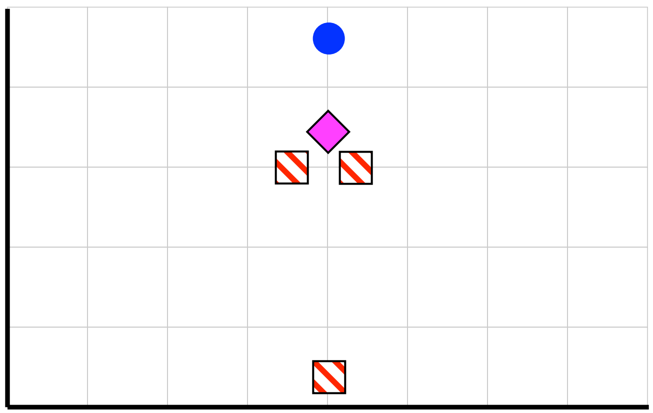
Suppose however, we add in an **irrelevant** feature, for example the insects mass.

Using both the antenna length and the insects mass with the 1-NN algorithm we get the wrong classification!

How do we mitigate the nearest neighbor algorithms sensitivity to irrelevant features?

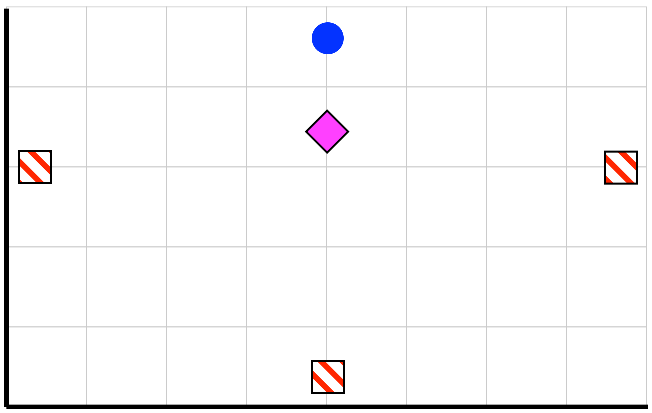
- Use more training instances
- Ask an expert what features are relevant to the task
- Use statistical tests to try to determine which features are useful
- Search over feature subsets

The nearest neighbor algorithm is sensitive to the units of measurement



X axis measured in **centimeters**
Y axis measure in dollars

The nearest neighbor to the **pink** unknown instance is **red**.



X axis measured in **millimeters**
Y axis measure in dollars

The nearest neighbor to the **pink** unknown instance is **blue**.

One solution is to normalize the units to pure numbers.

Scaling Issues

- Attributes may have to be scaled to prevent distance measures from being dominated by one of the attributes
- Example:
  - height of a person may vary from 1.5m to 1.8m
  - weight of a person may vary from 90lb to 300lb
  - income of a person may vary from \$10K to \$1M

Advantages/Disadvantages of Nearest Neighbor

- Advantages:
  - Simple to implement
  - Handles correlated features (Arbitrary class shapes)
  - Defined for any distance measure
  - Handles streaming data trivially
- Disadvantages:
  - Very sensitive to irrelevant features.
  - Slow classification time for large datasets
  - Works best for real valued datasets

– Does not build a model explicitly
  - "Lazy learners", as opposed to eager learners like decision tree induction

Bayesian Methods

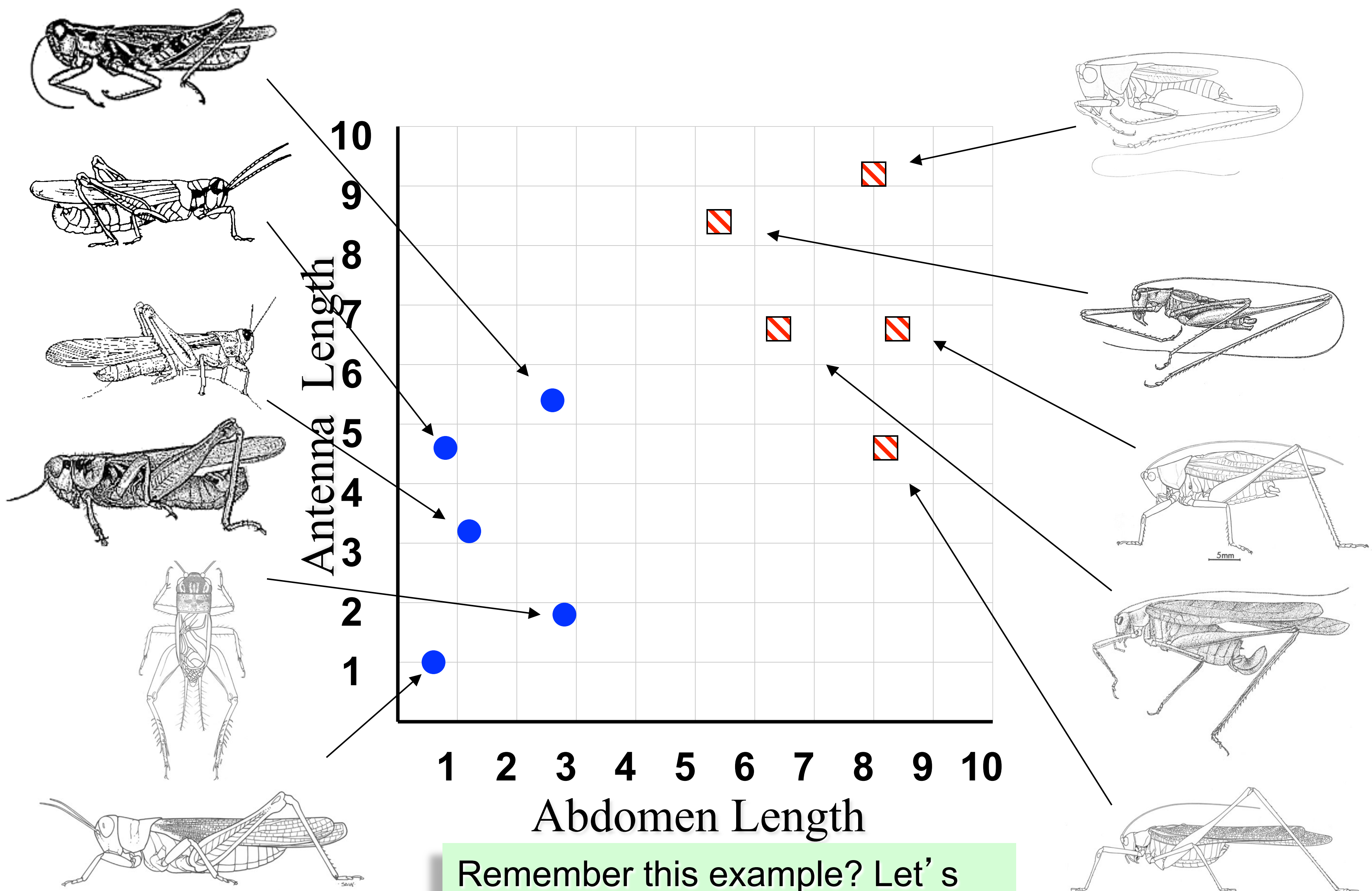
- Learning and classification methods based on probability theory.
- Bayes theorem plays a critical role in probabilistic learning and classification.
- Builds a *generative model* that approximates how data is produced
- Uses *prior* probability of each category given no information about an item.
- Categorization produces a *posterior* probability distribution over the possible categories given a description of an item.

Naïve Bayes Classifier

Thomas Bayes
1702 - 1761

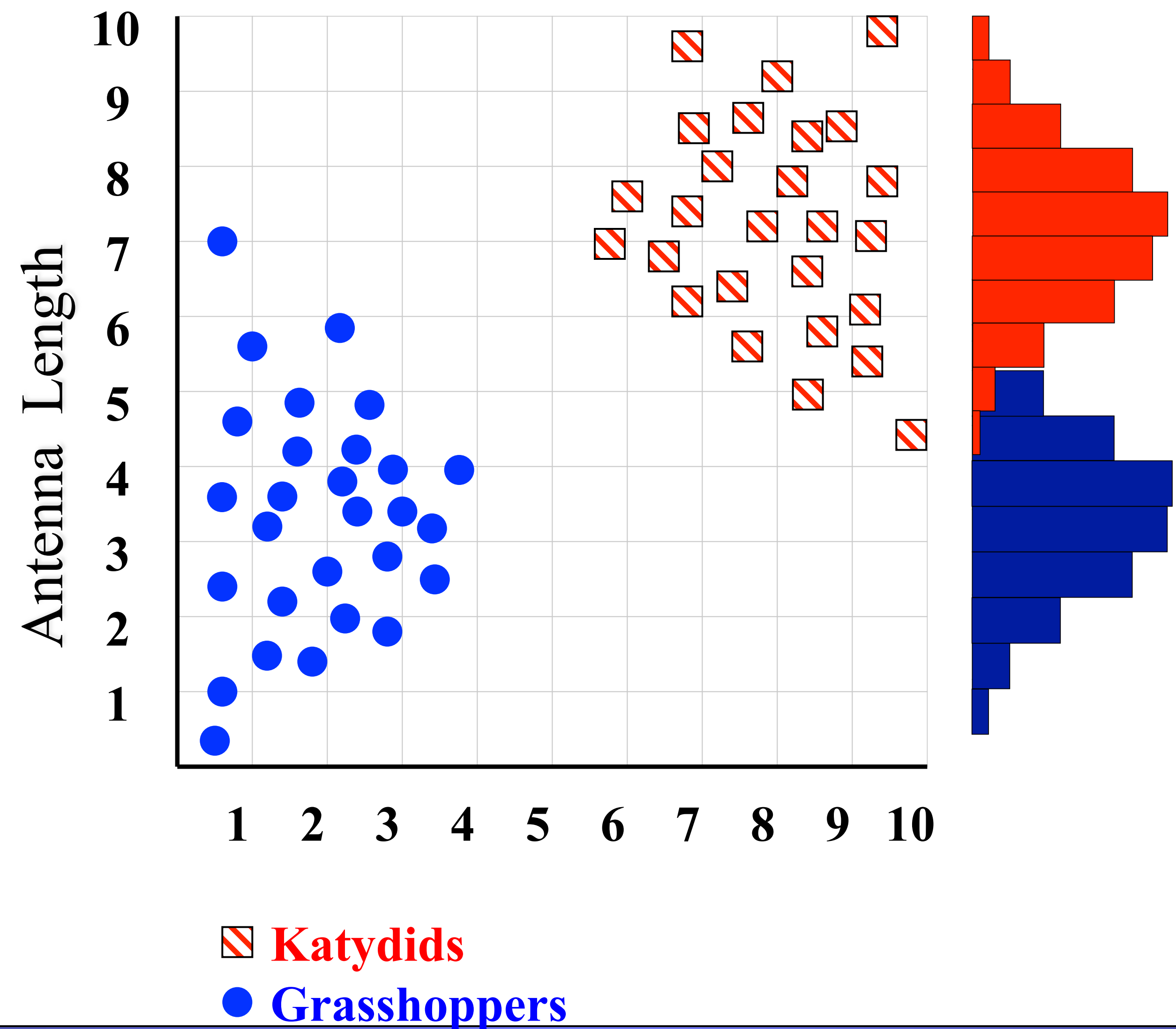
We will start off with a visual intuition, before looking at the math…

Grasshoppers   Katydids

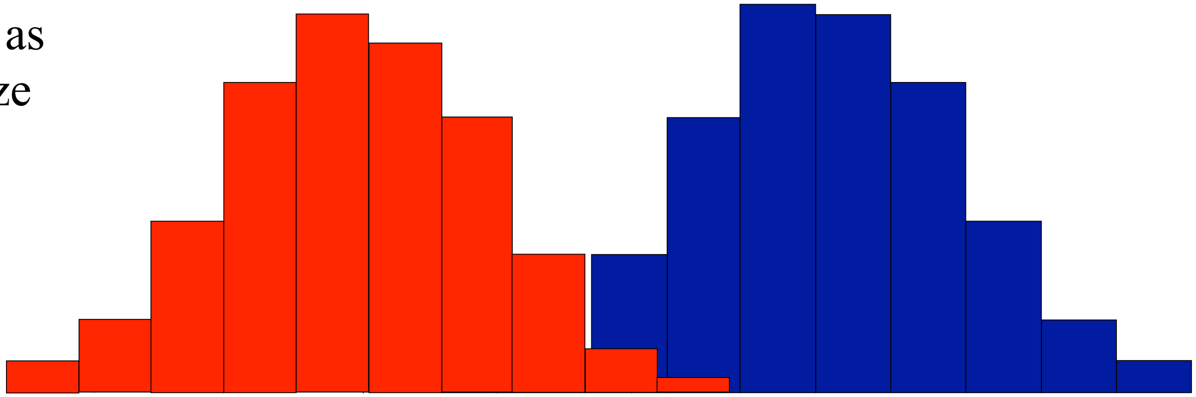


Remember this example? Let's get lots more data…

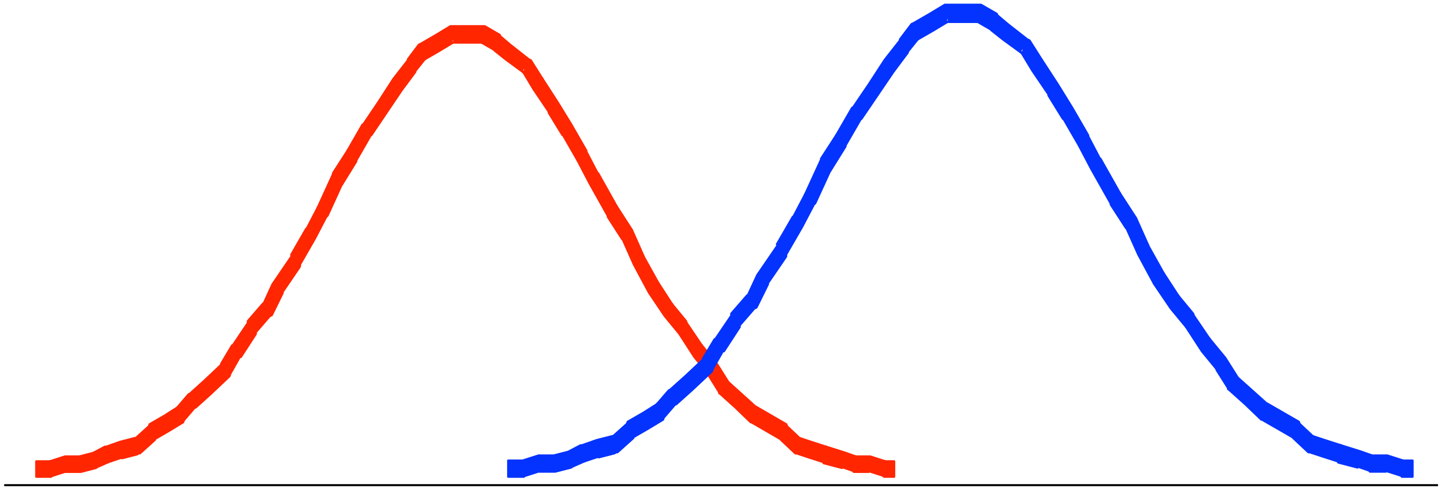
With a lot of data, we can build a histogram. Let us just build one for "Antenna Length" for now…



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…

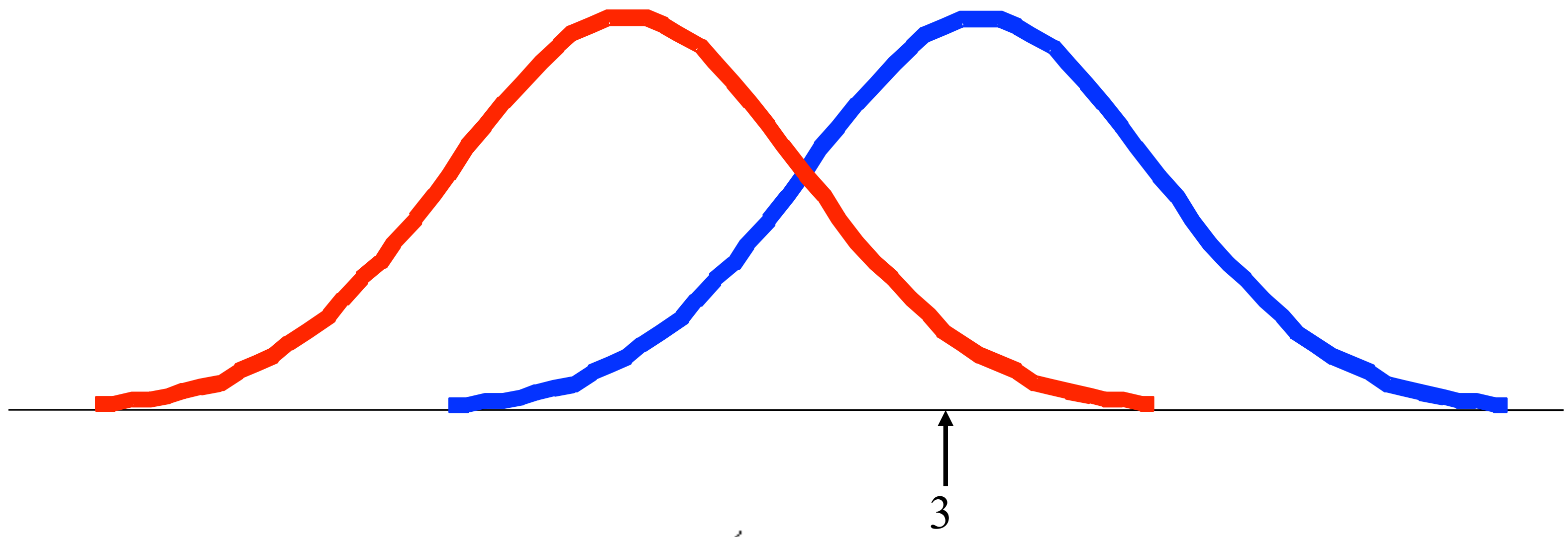


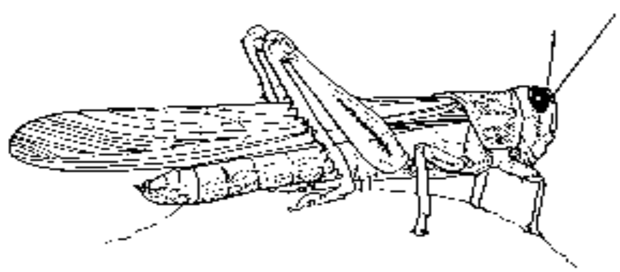
• We want to classify an insect we have found. Its antennae are 3 units long. How can we classify it?

• 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…

$P(C \mid A)$ = probability of class C, *given* that we have observed A



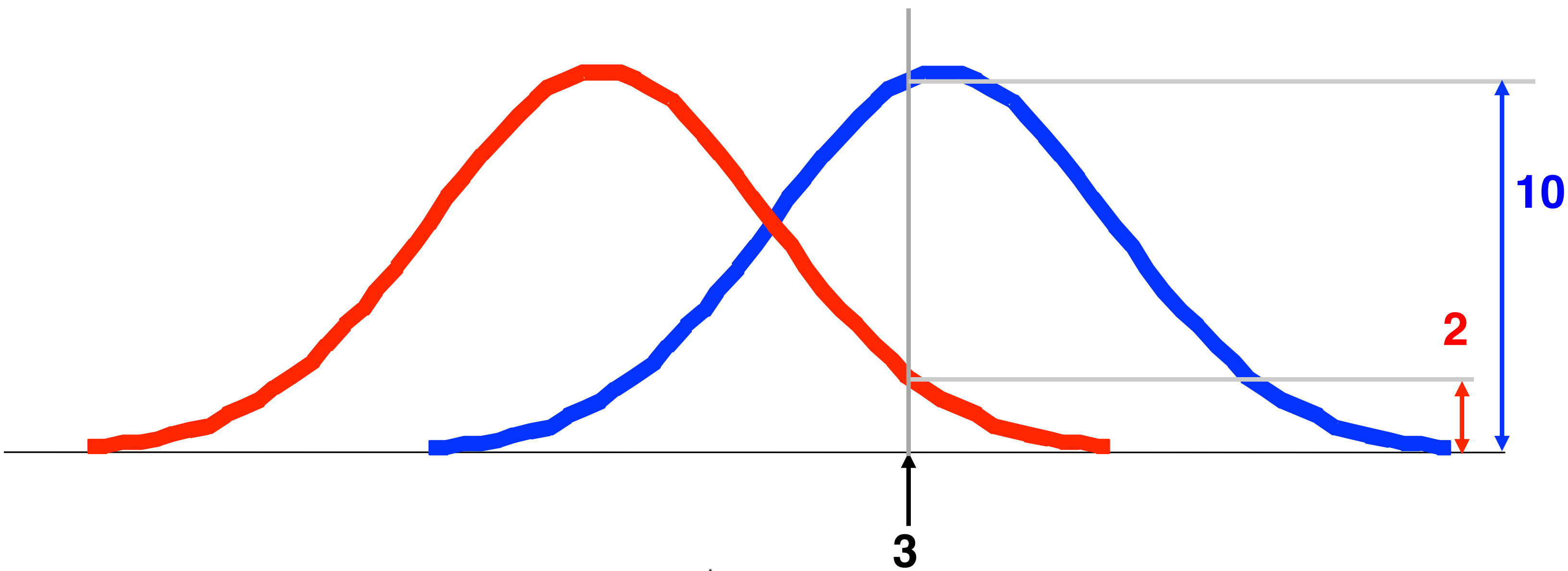


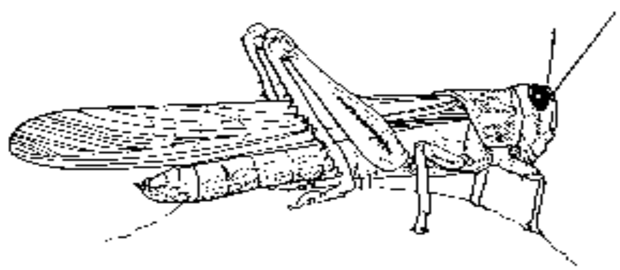
Antennae length is 3

P*(*C*| *A*) = probability of class *C*, given that we have observed *A

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

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



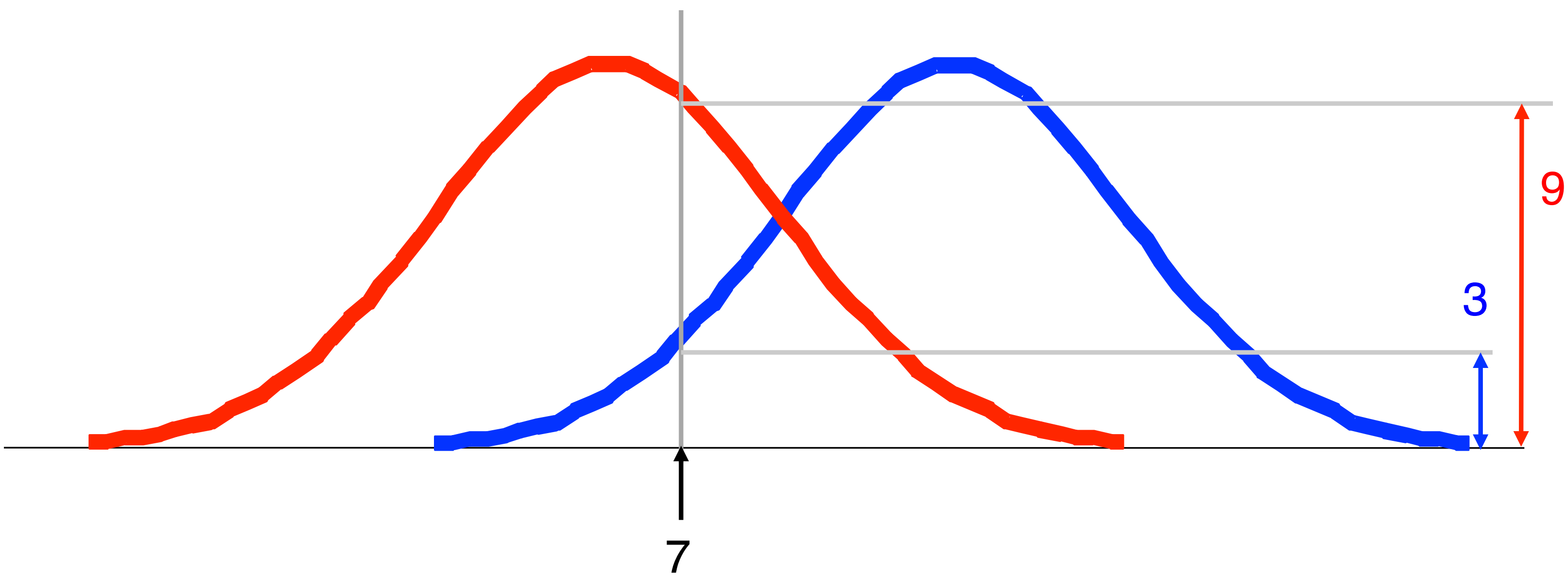


Antennae length is 3

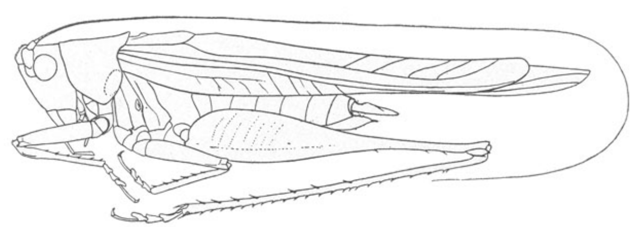
$P(C \mid A)$ = probability of class C, given that we have observed A

$$P(\text{Grasshopper} \mid 7) = 3 / (3 + 9) = 0.250$$

$$P(\text{Katydid} \mid 7) = 9 / (3 + 9) = 0.750$$



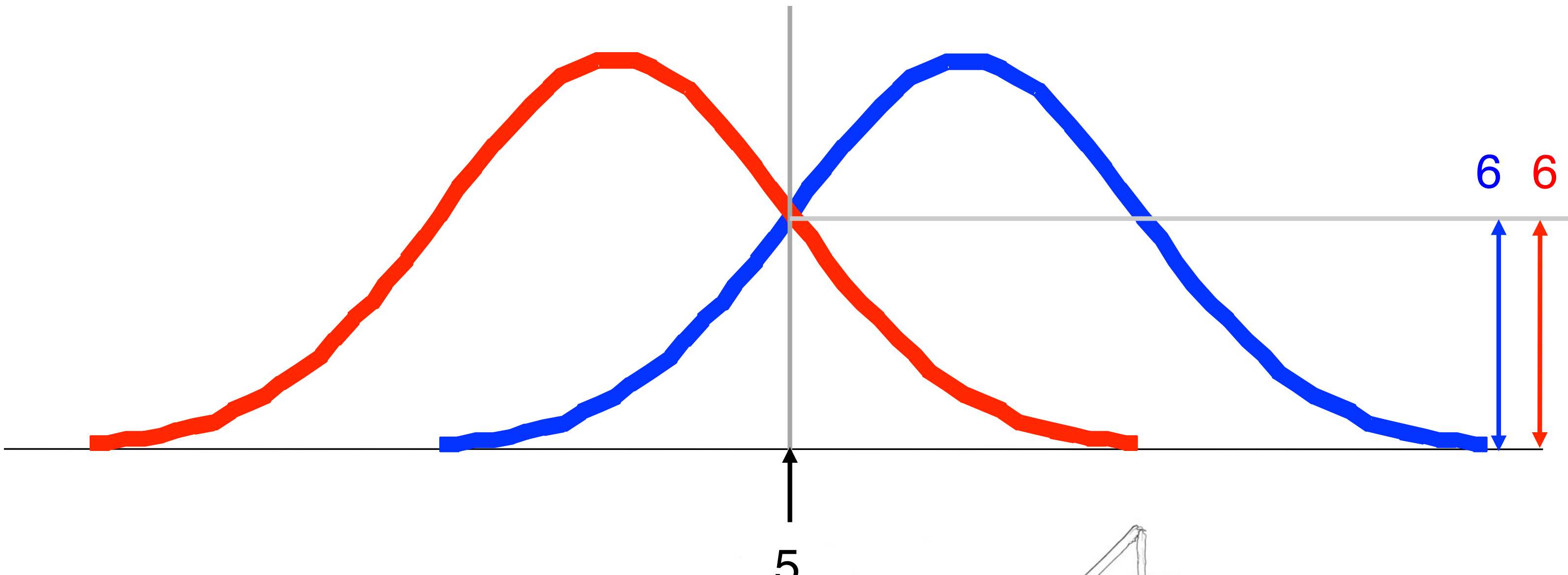
Antennae length is 7

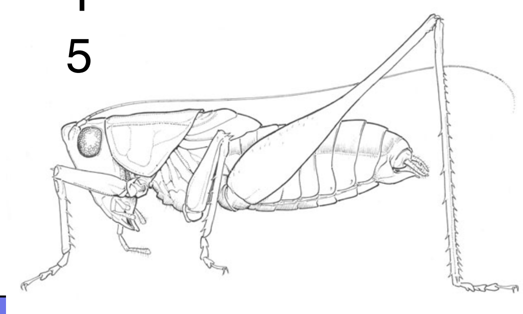


P*(*C*| *A*) = probability of class *C*, given that we have observed *A

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.

Bayes Classifier

- A probabilistic framework for solving classification problems
- Conditional Probability:

$$P(C \mid A) = \frac{P(A, C)}{P(A)}$$

$$P(A \mid C) = \frac{P(A, C)}{P(C)}$$

- Bayes theorem:

$$P(C \mid A) = \frac{P(A \mid C) P(C)}{P(A)}$$

Example of Bayes Theorem

- Given:
  - A doctor knows that meningitis causes stiff neck 50% of the time
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20

- If a patient has stiff neck, what's the probability he/she has meningitis?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Bayesian Classifiers

- Consider each attribute and class label as random variables
- Given a record with attributes (A_1, A_2,…, A_n)
  - Goal is to predict class C
  - Specifically, we want to find the value of C that maximizes $P(C \mid A_1, A_2, \ldots, A_n)$
- Can we estimate $P(C \mid A_1, A_2, \ldots, A_n)$ directly from data?

Bayesian Classifiers

- Approach:
  - compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the Bayes theorem

$$\boxed{P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C)P(C)}{P(A_1 A_2 \ldots A_n)}}$$

  - Choose value of C that maximizes
    $P(C \mid A_1, A_2, \ldots, A_n)$

  - Equivalent to choosing value of C that maximizes
    $P(A_1, A_2, \ldots, A_n|C)\ P(C)$

- How to estimate $P(A_1, A_2, \ldots, A_n \mid C)$?

Closer Look At Bayes Theorem

$$P(C \mid A) = \frac{P(A \mid C)\, P(C)}{P(A)}$$

- $P(C \mid A)$ = probability of instance A being in class C,
  This is what we are trying to compute
- $P(A \mid C)$ = probability of generating instance A given class C,
  We can imagine that being in class C, causes you to have feature *A* with some probability
- $P(C)$ = probability of occurrence of class C,
  This is just how frequent the class C, is in our database
- $P(A)$ = probability of instance A occurring
  This can actually be ignored, since it is the same for all classes

How to Estimate Probabilities from Data?

Tid	Home Owner	Marital Status	Annual Income	Defaulted
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Class: $P(C) = N_c/N$
  - i.e., $P(No) = 7/10$,
    $P(Yes) = 3/10$

- For discrete attributes:

$$P(A_i \mid C_k) = |A_{ik}| / N_c$$

  - where $|A_{ik}|$ is number of instances having attribute A_i and belongs to class C_k
  - Examples:

    P(MaritalStatus=Married|No) = 4/7
    P(HomeOwner=Yes|Yes)=0

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 A.

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 $p(\text{male} \mid drew)$ or $p(\text{female} \mid drew)$

(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 the same for all classes)

This is Officer Drew. Is Officer Drew a Male or Female?

Officer Drew

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

$$P(C \mid A) = \frac{P(A \mid C)\, P(C)}{P(A)}$$

Officer Drew

$$P(C \mid A) = \frac{P(A \mid C)\, P(C)}{p(A)}$$

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

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

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

Officer Drew is more likely to be 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?

$$P(C \mid A) = \frac{P(A \mid \mathrm{C}\,)\,P(C)}{p(A)}$$

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

$$P(A|C) = P(A_1|C) * P(A_2|C) * \ldots.* P(\mathrm{A}_n|C)$$

The probability of class *C* generating instance *A*, equals….

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

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

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

$$P(A|C) = P(A_1|C) * P(A_2|C) * \ldots * P(A_n|C)$$

New point is classified to C if $P(C) \prod P(A_i| C)$ is *maximal*.

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

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

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

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

How to Estimate Probabilities from Data?

- For continuous attributes:
  - Discretize the range into bins
    - one ordinal attribute per bin
    - violates independence assumption

k

  - Two-way split: (A < v) or (A > v)
    - choose only one of the two splits as new attribute
  - Probability density estimation:
    - Assume attribute follows a normal distribution
    - Use data to estimate parameters of distribution (e.g., mean and standard deviation)
    - Once probability distribution is known, can use it to estimate the conditional probability $P(A_i|C)$

How to Estimate Probabilities from Data?

Tid	**Home Owner**	**Marital Status**	**Annual Income**	**Defaulted**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Normal distribution:

$$P(A_i \mid C) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

  – One for each (A_i, C) pair

- For (Income, Class=No):
  – If Class=No
    - sample mean = 110
    - sample variance = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

More Example

Given a Test Record:

$$X = (\text{HomeOwner} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

naive Bayes Classifier:

P(HomeOwner=Yes|No) = 3/7
P(HomeOwner = No|No) = 4/7
P(HomeOwner = Yes|Yes) = 0
P(HomeOwner = No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

- P(X|Class=No) = P(HomeOwner=No|Class=No)
  × P(Married| Class=No)
  × P(Income=120K| Class=No)
  = 4/7 × 4/7 × 0.0072 = 0.0024

- P(X|Class=Yes) = P(HomeOwner=No| Class=Yes)
  × P(Married| Class=Yes)
  × P(Income=120K| Class=Yes)
  = 1 × 0 × 1.2 × 10^{-9} = 0

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)
=> Class = No

Naïve Bayes Classifier

- If one of the conditional probability is zero, then the entire expression becomes zero
- Probability estimation:

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m - estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: number of classes

p: prior probability

m: parameter

Another Example

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

A: attributes

M: mammals

N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

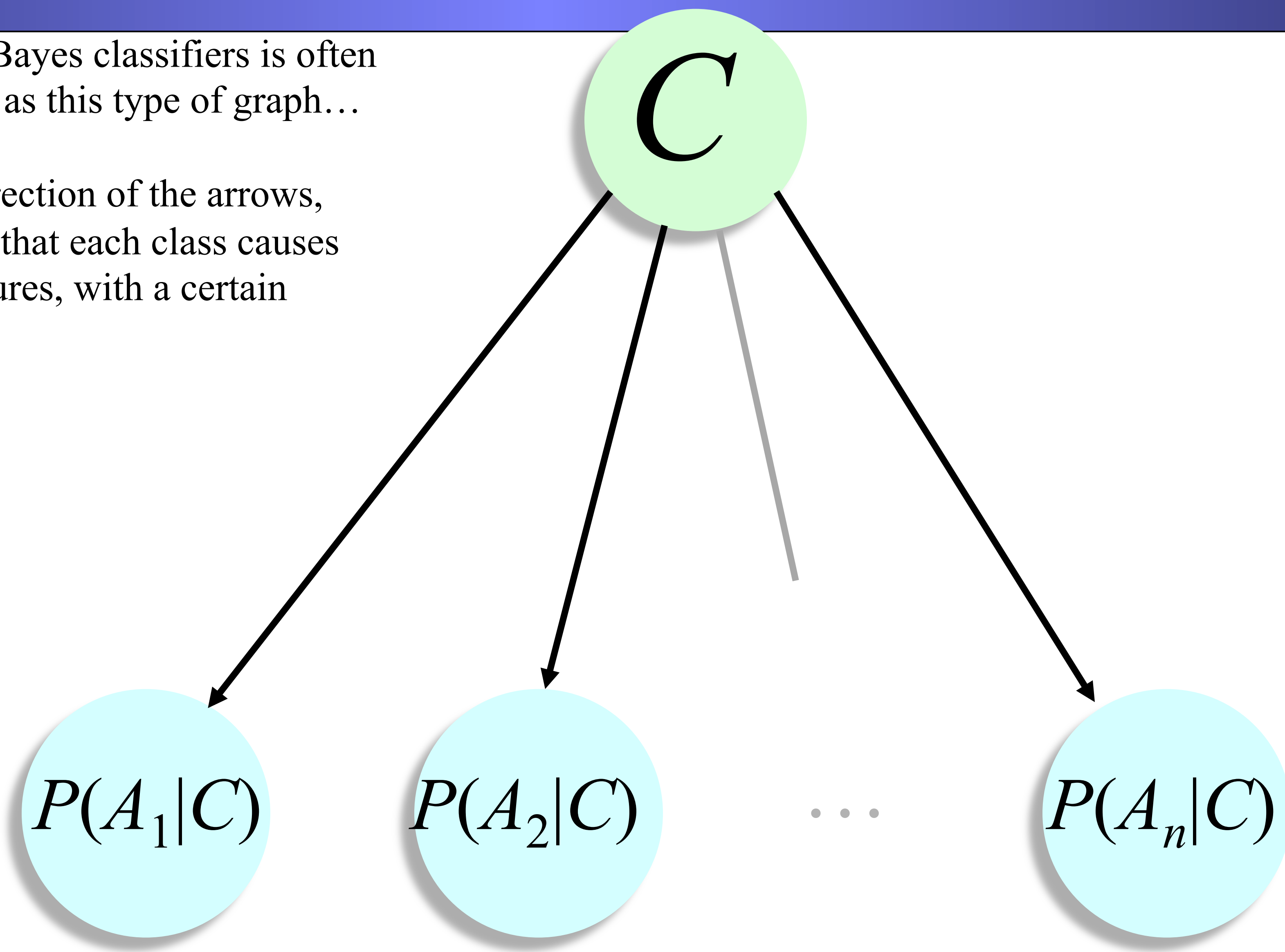
Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

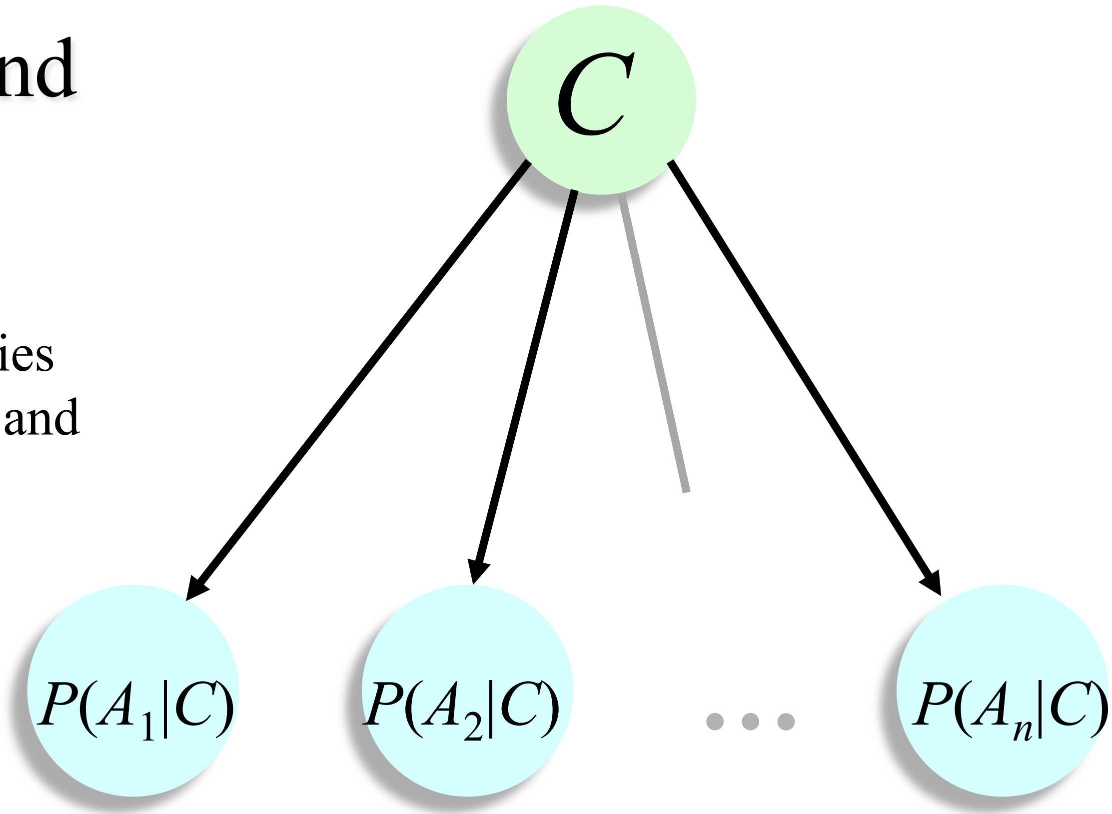
The Naive Bayes classifiers is often represented as this type of graph…

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 compute 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 person's sex based on several features, including eye color. (Of course, eye color is completely irrelevant to a persons gender)

$$P(\text{Jessica} \mid C) = P(\text{eye} = \text{brown} \mid C) * P(\text{wears\_dress} = \text{yes} \mid C) * \ldots.$$

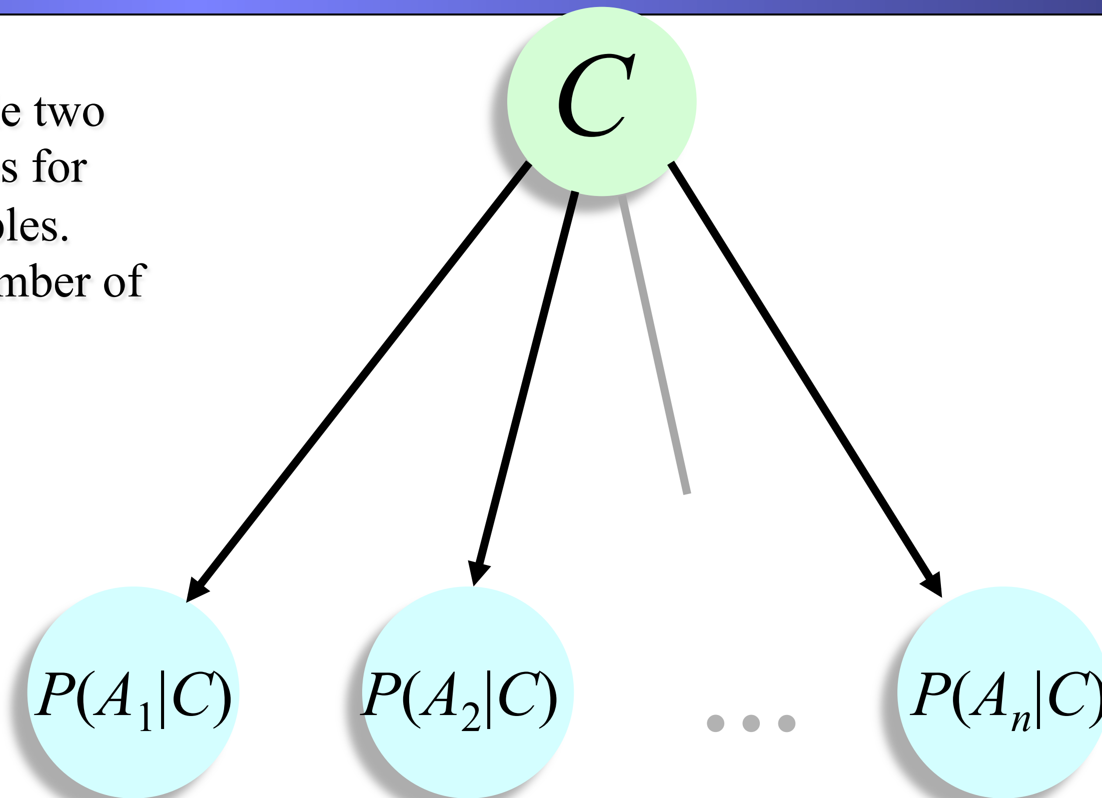
p(Jessica | Female) = *9,000/10,000* * *9,975/10,000* * ….

p(Jessica | Male) = *9,001/10,000* * *2/10,000* * ….

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



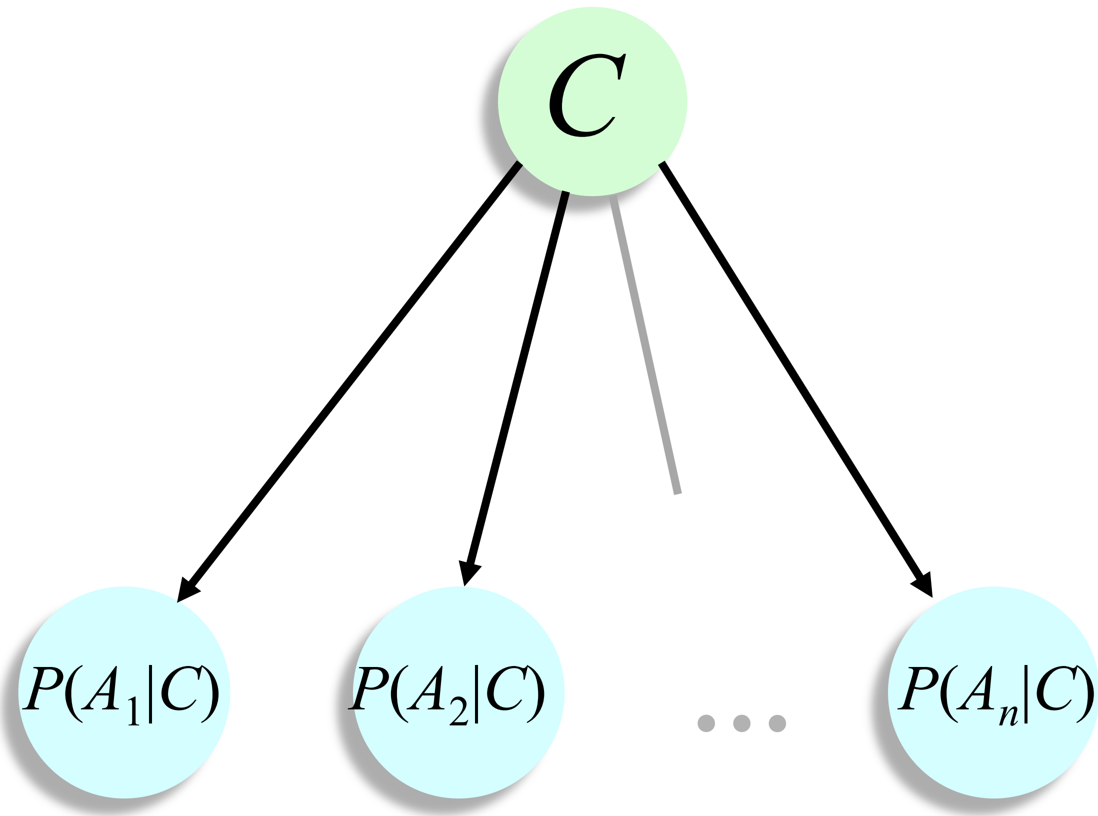
Animal	Mass >10_{kg}	
Cat	Yes	0.15
	No	0.85
Dog	Yes	0.91
	No	0.09
Pig	Yes	0.99
	No	0.01

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

Animal
Cat
Dog
Pig

Problem!

Naïve Bayes assumes independence of features…

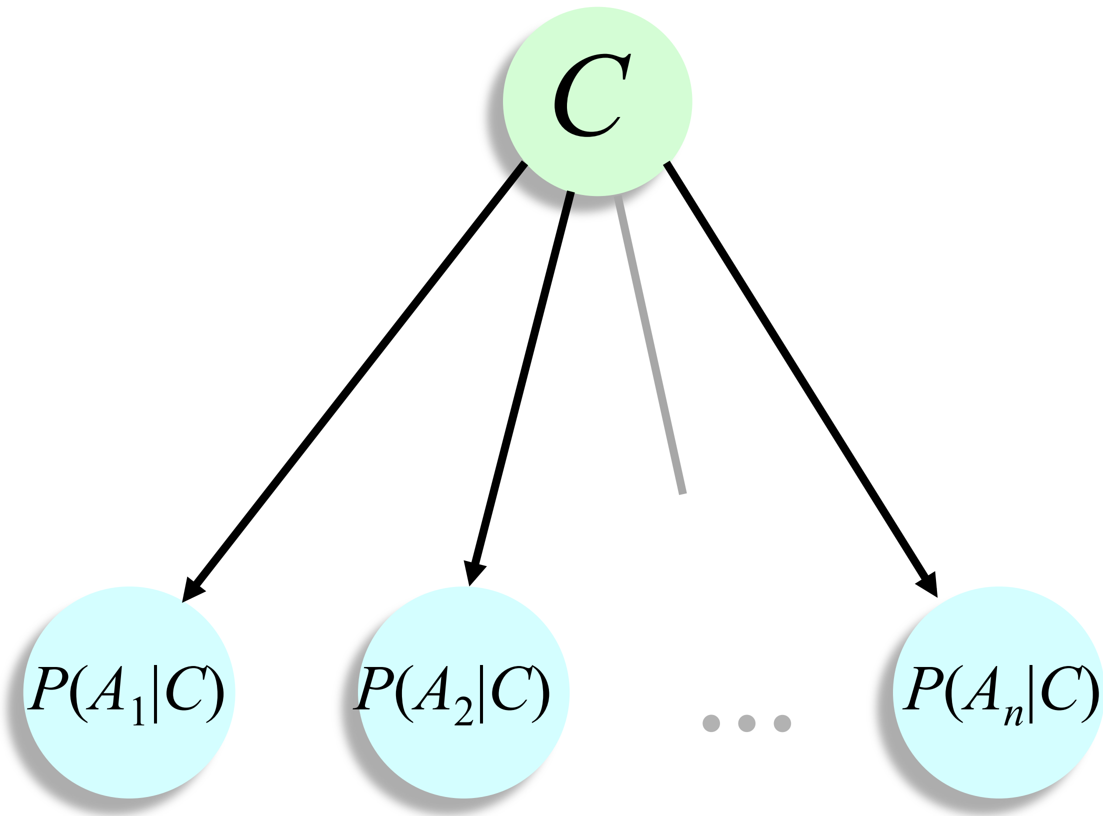


Sex	Over 6 foot	
Male	Yes	0.15
	No	0.85
Female	Yes	0.01
	No	0.99

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

Solution

Consider the relationships between attributes…

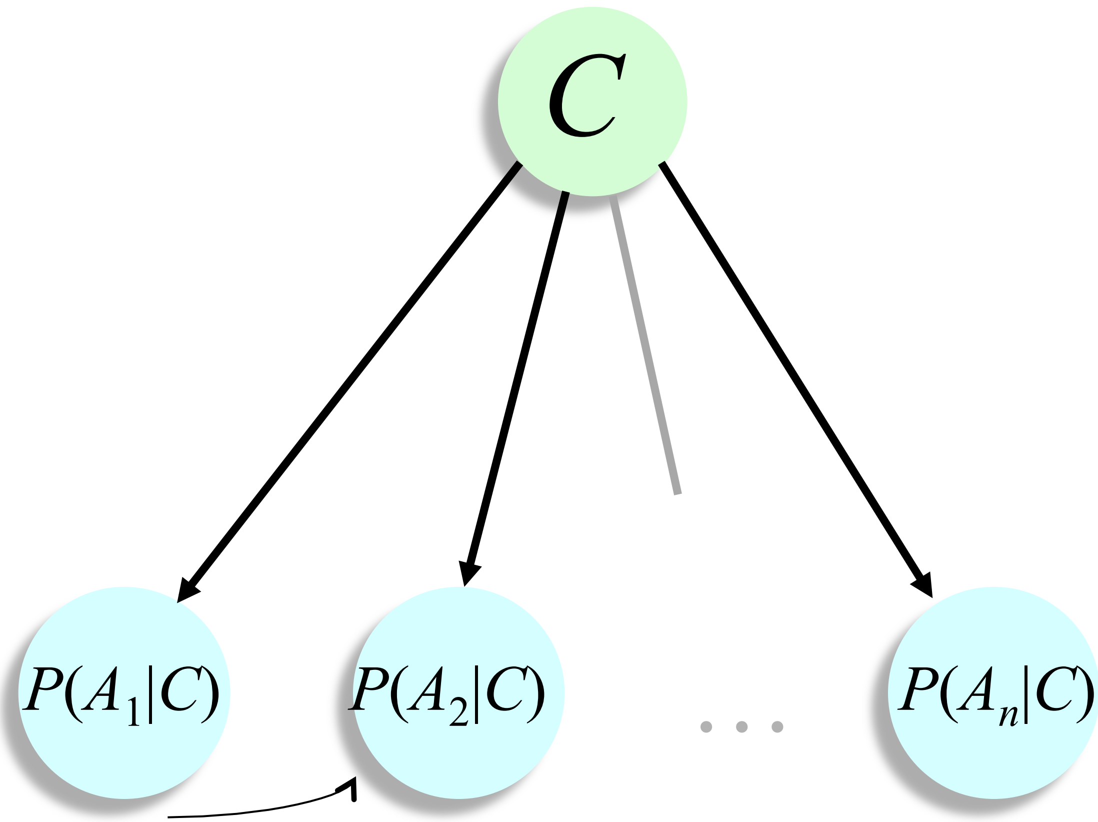


Sex	**Over 6 foot**	
Male	Yes	0.15
	No	0.85
Female	Yes	0.01
	No	0.99

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

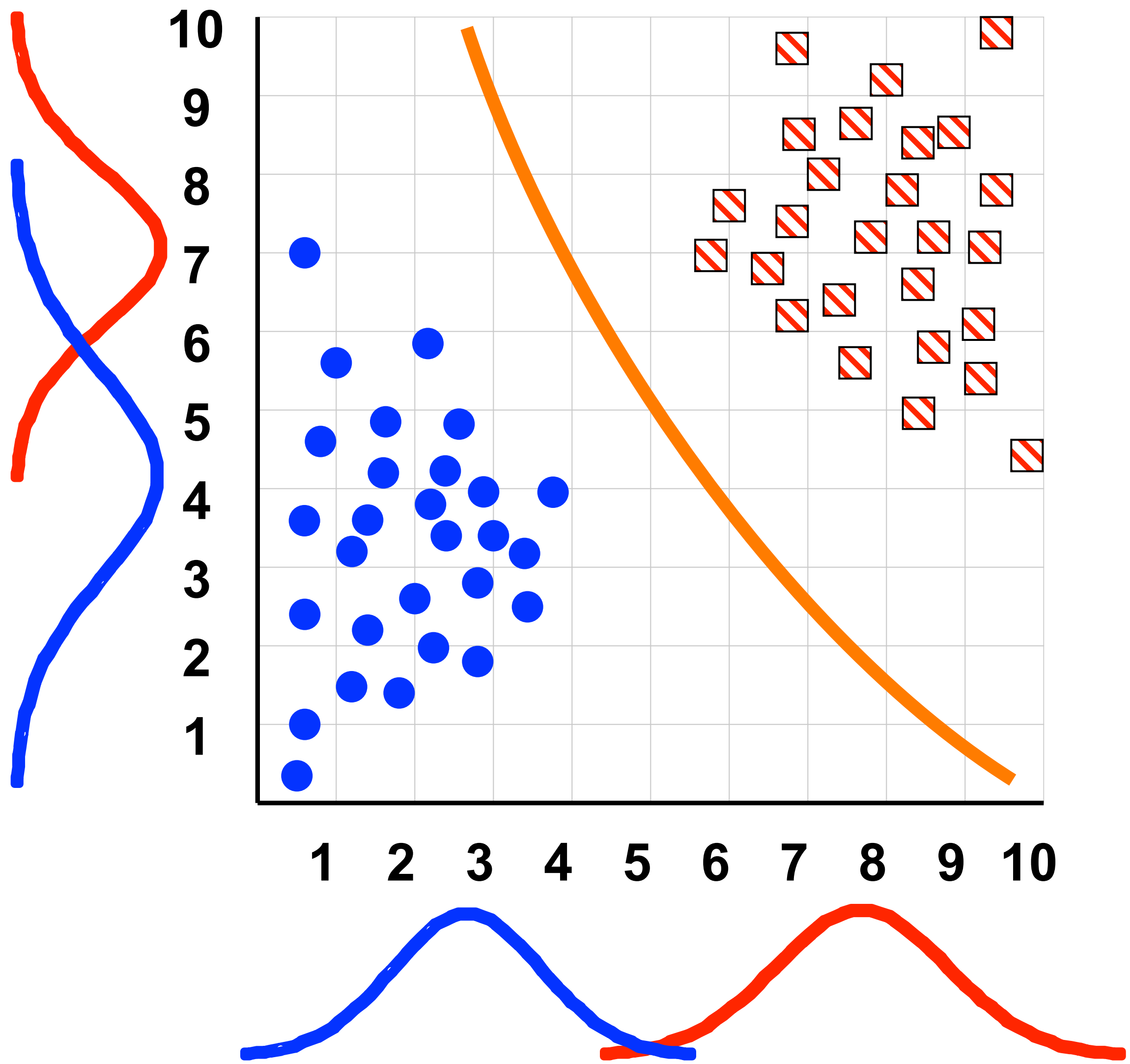
Solution

Consider the relationships between attributes…



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

The Naïve Bayesian Classifier has a quadratic decision boundary



Advantages/Disadvantages of Naïve Bayes

- Advantages:
  - Fast to train (single scan). Fast to classify
  - Not sensitive to irrelevant features
  - Robust to isolated noise points
  - Handles real and discrete data
  - Handles streaming data well
  - Handle missing values by ignoring the instance during probability estimate calculations
- Disadvantages:
  - Independence assumption may not hold for some attributes
  - Use other techniques such as Bayesian Belief Networks