CPE412 Pattern Recognition

Week 4

Bayesian Decision Theory

Dr. Nehad Ramaha,
Computer Engineering Department
Karabük Universities

Bayesian Decision Theory

Week 3

- Bayesian Decision Theory is a fundamental statistical approach that quantifies the tradeoffs between various decisions using probabilities and costs that accompany such decisions.
- First, we will assume that all probabilities are known.
- Then, we will study the cases where the probabilistic structure is not completely known.

Bayesian Decision Theory

- Design classifiers to recommend decisions that minimize some total expected ”risk”.
- The simplest risk is the classification error (i.e., costs are equal).
- Typically, the risk includes the cost associated with different decisions.

Fish Sorting Example Revisited

- State of nature is a random variable.
- Define w as the type of fish we observe (state of nature, *class*) where
  - $w = w_1$ for sea bass,
  - $w = w_2$ for salmon.
  - $P(w_1)$ is the *a priori probability* that the next fish is a sea bass.
  - $P(w_2)$ is the a priori probability that the next fish is a salmon.

Prior Probabilities

- Prior probabilities reflect our knowledge of how likely each type of fish will appear before we actually see it.
- How can we choose $P(w_1)$ and $P(w_2)$?
  - Set $P(w_1) = P(w_2)$ if they are equiprobable (*uniform priors*).
  - May use different values depending on the fishing area, time of the year, etc.
- Assume there are no other types of fish

$$P(w_1) + P(w_2) = 1$$

(exclusivity and exhaustivity).

Making a Decision

- How can we make a decision with only the prior information?

$$\text{Decide} \quad \begin{cases} w_1 & \text{if } P(w_1) > P(w_2) \\ w_2 & \text{otherwise} \end{cases}$$

- What is the *probability of error* for this decision?

$$P(error) = \min\{P(w_1), P(w_2)\}$$

Class-Conditional Probabilities

Week 3

- Let's try to improve the decision using the lightness measurement x.
- Let x be a continuous random variable.
- Define $p(x|w_j)$ as the *class-conditional probability density* (probability of x given that the state of nature is w_j for $j = 1, 2$).
- $p(x|w_1)$ and $p(x|w_2)$ describe the difference in lightness between populations of sea bass and salmon.

Class-Conditional Probabilities

Week 3

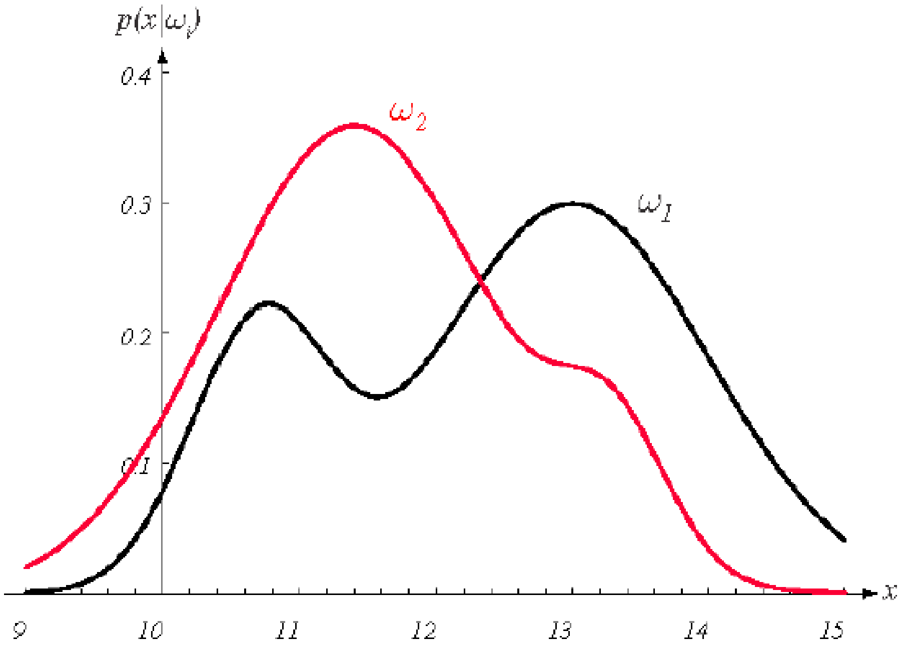


Figure 1: Hypothetical class-conditional probability density functions for two classes.

Bayes' Theorem:

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

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

Posterior Probabilities

- Suppose we know $P(w_j)$ and $p(x|w_j)$ for $j = 1, 2$, and measure the lightness of a fish as the value x.
- Define $P(w_j|x)$ as the *a posteriori probability* (probability of the state of nature being w_j given the measurement of feature value x).
- We can use the *Bayes formula* to convert the prior probability to the posterior probability

$$P(w_j|x) = \frac{p(x|w_j)P(w_j)}{p(x)}$$

where $p(x) = \sum_{j=1}^{2} p(x|w_j)P(w_j)$.

Making a Decision

- $p(x|w_j)$ is called the *likelihood* and $p(x)$ is called the *evidence*.
- How can we make a decision after observing the value of x?

$$\text{Decide} \quad \begin{cases} w_1 & \text{if } P(w_1|x) > P(w_2|x) \\ w_2 & \text{otherwise} \end{cases}$$

- Rewriting the rule gives

$$\text{Decide} \quad \begin{cases} w_1 & \text{if } \frac{p(x|w_1)}{p(x|w_2)} > \frac{P(w_2)}{P(w_1)} \\ w_2 & \text{otherwise} \end{cases}$$

- Note that, at every x, $P(w_1|x) + P(w_2|x) = 1$.

Probability of Error

- What is the probability of error for this decision?

$$P(error|x) = \begin{cases} P(w_1|x) & \text{if we decide } w_2 \\ P(w_2|x) & \text{if we decide } w_1 \end{cases}$$

- What is the average probability of error?

$$P(error) = \int_{-\infty}^{\infty} p(error, x)\, dx = \int_{-\infty}^{\infty} P(error|x)\, p(x)\, dx$$

- *Bayes decision rule* minimizes this error because

using feature x make it more accurate

$$P(error|x) = \min\{P(w_1|x), P(w_2|x)\}.$$

Bayesian Decision Theory

- How can we generalize to
  - more than one feature?
    - replace the scalar x by the feature vector $\mathbf{x}$
  - more than two states of nature?
    - just a difference in notation
  - allowing actions other than just decisions?
    - allow the possibility of rejection
  - different risks in the decision?
    - define how costly each action is

Bayesian Decision Theory

- Let $\{w_1, \ldots, w_c\}$ be the finite set of c states of nature (*classes*, *categories*).
- Let $\{\alpha_1, \ldots, \alpha_a\}$ be the finite set of a possible *actions*.
- Let $\lambda(\alpha_i|w_j)$ be the *loss* incurred for taking action α_i when the state of nature is w_j.
- Let x be the d-component vector-valued random variable called the *feature vector*.

Bayesian Decision Theory

- $p(\mathbf{x}|w_j)$ is the class-conditional probability density function.
- $P(w_j)$ is the prior probability that nature is in state w_j.
- The posterior probability can be computed as

$$P(w_j|\mathbf{x}) = \frac{p(\mathbf{x}|w_j)P(w_j)}{p(\mathbf{x})}$$

where $p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x}|w_j)P(w_j)$.

Conditional Risk

Loss for taking specific wrong action (it's given to you)

- Suppose we observe $\mathbf{x}$ and take action α_i.
- If the true state of nature is w_j, we incur the loss $\lambda(\alpha_i|w_j)$.
- The expected loss with taking action α_i is

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|w_j)P(w_j|\mathbf{x})$$

which is also called the *conditional risk*.

Minimum-Risk Classification

- The general *decision rule* $\alpha(\mathbf{x})$ tells us which action to take for observation $\mathbf{x}$.
- We want to find the decision rule that minimizes the overall risk

$$R = \int_{1}^{a} R(\alpha(\mathbf{x})|\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}.$$

- Bayes decision rule minimizes the overall risk by selecting the action α_i for which $R(\alpha_i|\mathbf{x})$ is minimum.
- The resulting minimum overall risk is called the *Bayes risk* and is the best performance that can be achieved.

Two-Category Classification

- Define
  - α_1: deciding w_1,
  - α_2: deciding w_2,
  - $\lambda_{ij} = \lambda(\alpha_i | w_j)$.

  } 4 loss value
- Conditional risks can be written as

$$R(\alpha_1|\mathbf{x}) = \lambda_{11} P(w_1|\mathbf{x}) + \lambda_{12} P(w_2|\mathbf{x}),$$

$$R(\alpha_2|\mathbf{x}) = \lambda_{21} P(w_1|\mathbf{x}) + \lambda_{22} P(w_2|\mathbf{x}).$$

Two-Category Classification

- The *minimum-risk decision rule* becomes

$$\text{Decide} \begin{cases} w_1 & \text{if } (\lambda_{21} - \lambda_{11}) P(w_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22}) P(w_2|\mathbf{x}) \\ w_2 & \text{otherwise} \end{cases}$$

- This corresponds to deciding w_1 if

$$\frac{p(\mathbf{x}|w_1)}{p(\mathbf{x}|w_2)} > \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \frac{P(w_2)}{P(w_1)}$$

Example-1

- Let's think of a market. Two different brands of eggs come to the market. Information about eggs from the experience gained and the records kept is as follows:
  - Brands: Br1 Egg and Br2 Egg
  - Daily supply amount: Br1 800, Br2 600
  - Broken egg(K) rate: 05%
- The question here is: what is the probability that an egg coming from Br2 will be broken in one day?

Example-1 (Let's apply the logic)

- A total of 1400 eggs,
- 600 of them come from Br2
- 70 broken eggs per day (total * broken rate)
- If we assume equal distribution of the broken parts according to the brands, there are 35 Br2
- Then the probability of Br2 being broken is $35/600 = 0.058333$

Example-1 (Apply Bayesian Decision Theory)

same for Br2 and Br1 (0.5)

given 0.05

$$P(K|Br2) = \frac{P(Br2|K) * P(K)}{P(Br2)}$$

600 out of 1400

Example-1 (Apply Bayesian Decision Theory)

same for Br_2 and Br_1 (0.5)

given 0.5

$$P(K|Br2) = \frac{P(Br2|K) * P(K)}{P(Br2)}$$

600 out of 1400

$$P(K|Br_2) = \frac{0.5 * 0.05}{0.428} = 0.058$$

Example-2

- Members of a consulting company rent a car at a rate of 60% from the 1st enterprise, 30% from the 2nd enterprise and 10% from the 3rd enterprise. If 9% of the vehicles coming from the first enterprise, 20% of the vehicles coming from the second enterprise and 6% of the vehicles coming from the third enterprise require maintenance;
  - a) What is the probability that a vehicle rented to the company will require maintenance?
  - b) What is the probability that the vehicle requiring maintenance came from the second enterprise?
- B: A car requires maintenance.
- A_i: Let the car come from the 1st, 2nd or 3rd enterprise With $i=1, 2, 3$.

Example–2

$P(A_1) = 0,60,\quad P(A_2) = 0.30,\quad P(A_3)=0.10$

$P(B \mid A_1) = 0.09\quad P(B \mid A_2) = 0.20,\quad P(B \mid A_3) = 0.06$

- P(B)—> the probability that the car will require maintenance.
- From the total probability is found using: $P(B) = \sum_{i=1}^{3} P(B|A_i).P(A_i)$

$$P(B) = (P(B \mid A_1).P(A_1) + P(B \mid A_2).P(A_2) + P(B \mid A_3).P(A_3)$$
$$= (0.60).(0.09) + (0.30).(0.20) + (0.10).(0.06)$$
$$= 0.12$$

- Then 12% of the vehicles rented by this company will require maintenance.

Example-2

- the probability that the vehicle requiring maintenance came from the second enterprise:

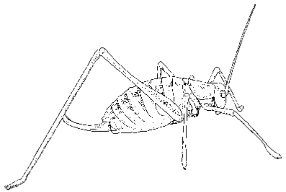
$$P(A_2 \mid B) = \frac{P(B \mid A_2).P(A_2)}{\sum_{i=1}^{3} P(B \mid A_i).P(A_i)} = \frac{(0.30).(0.20)}{0.12} = 0.50$$

$P(B)$

The Classification Problem

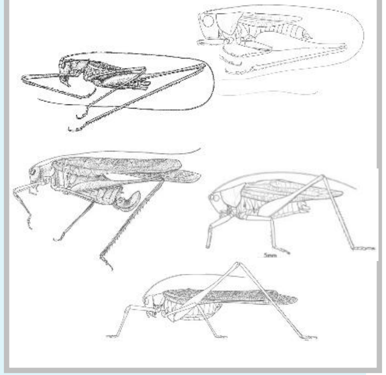
(informal definition)

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.

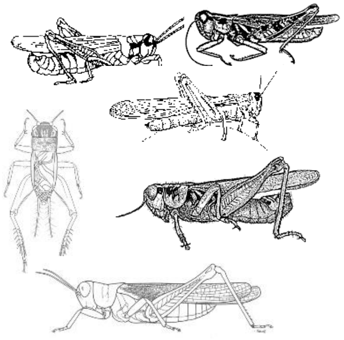


Katydid or **Grasshopper**?

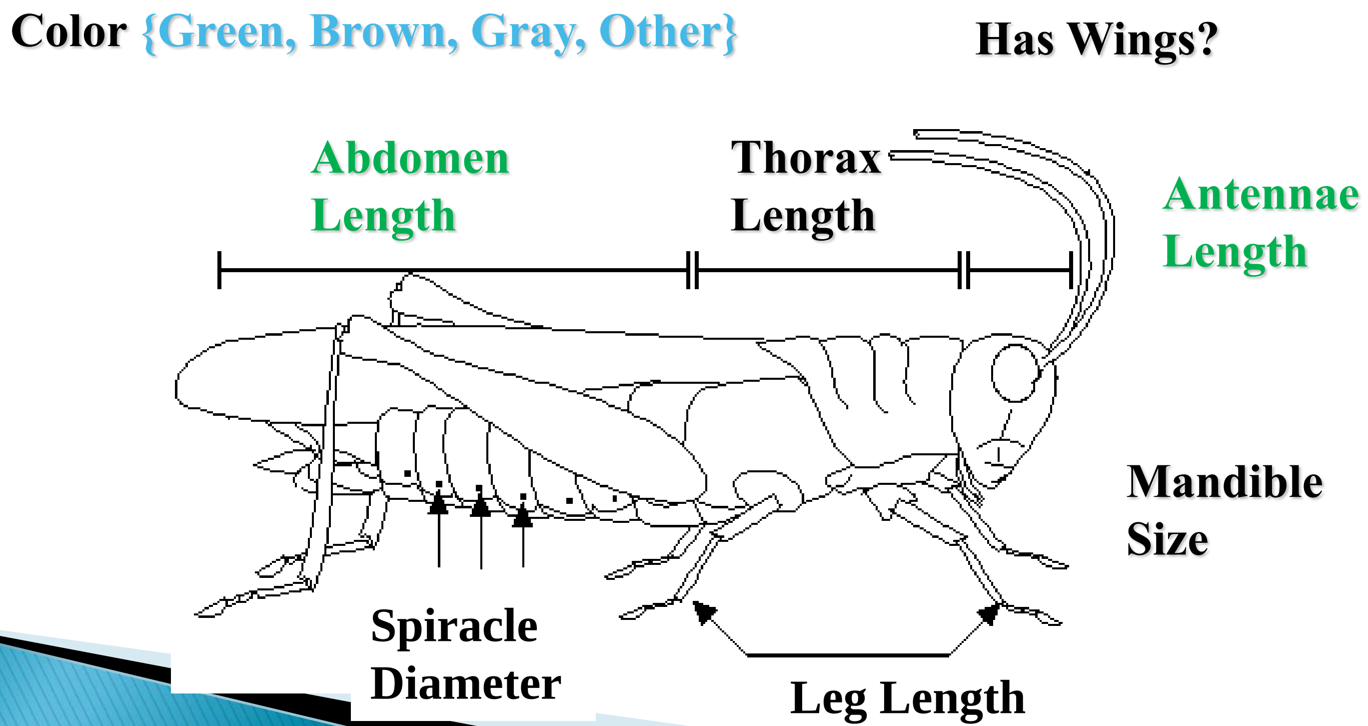
Katydids



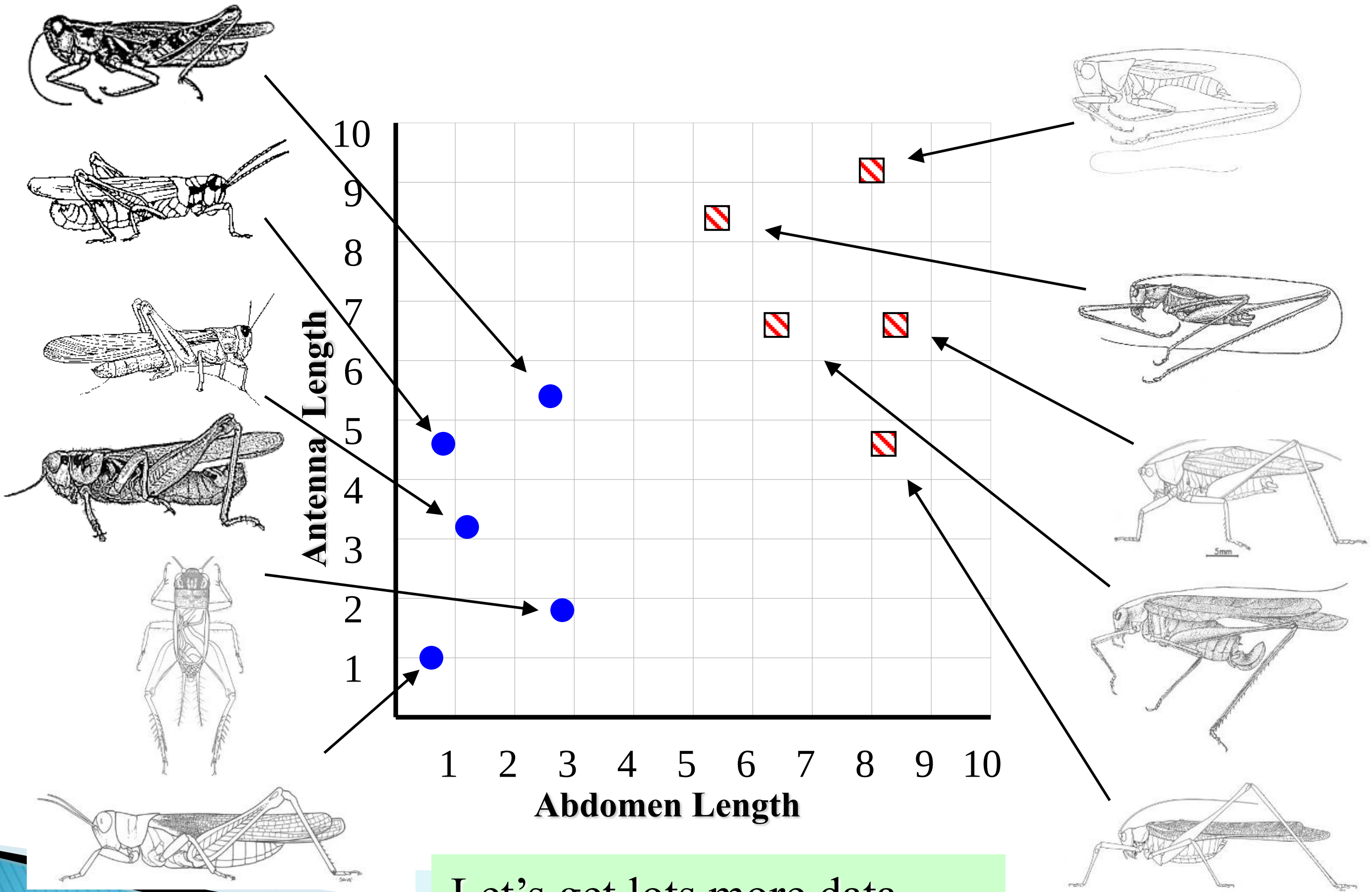
Grasshoppers



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

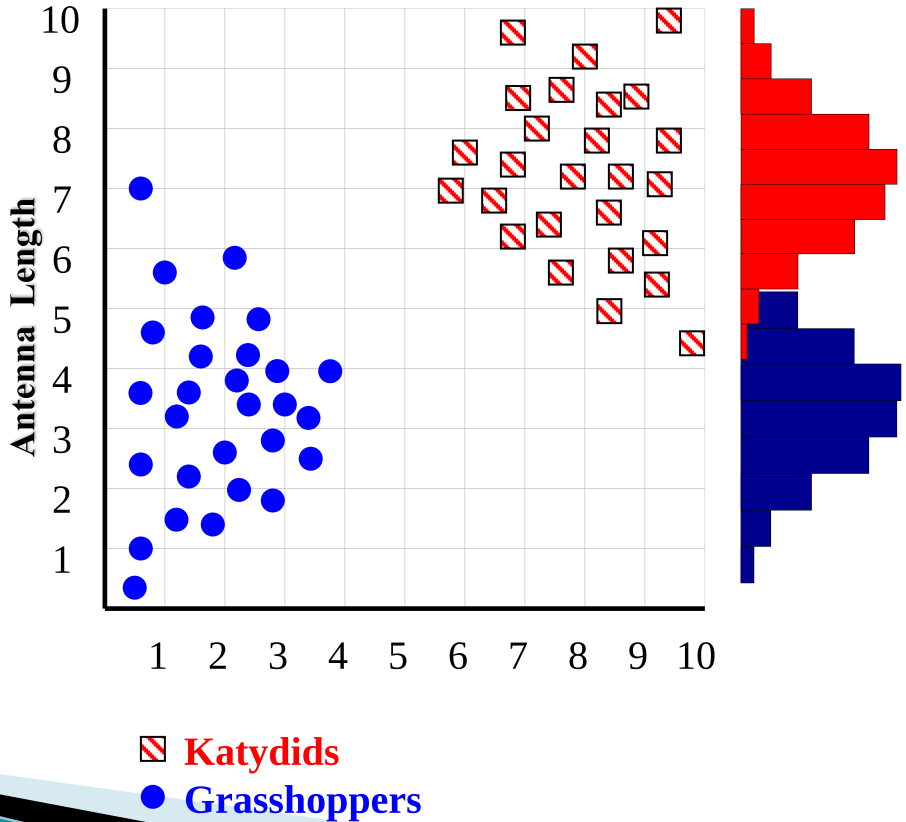


Grasshoppers **Katydids**

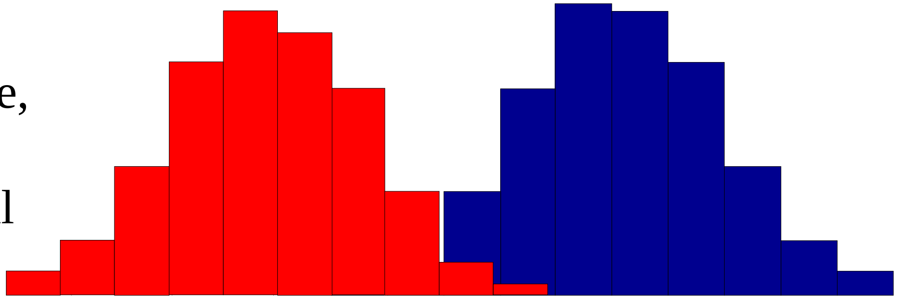


Let’s get lots more data…

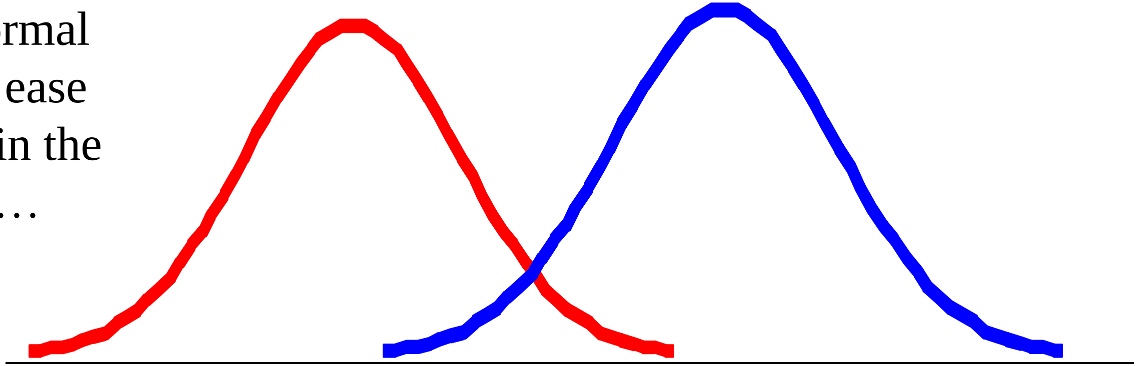
With a lot of data, we can build a histogram. Let us just build one for "Antenna Length" for now…



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



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

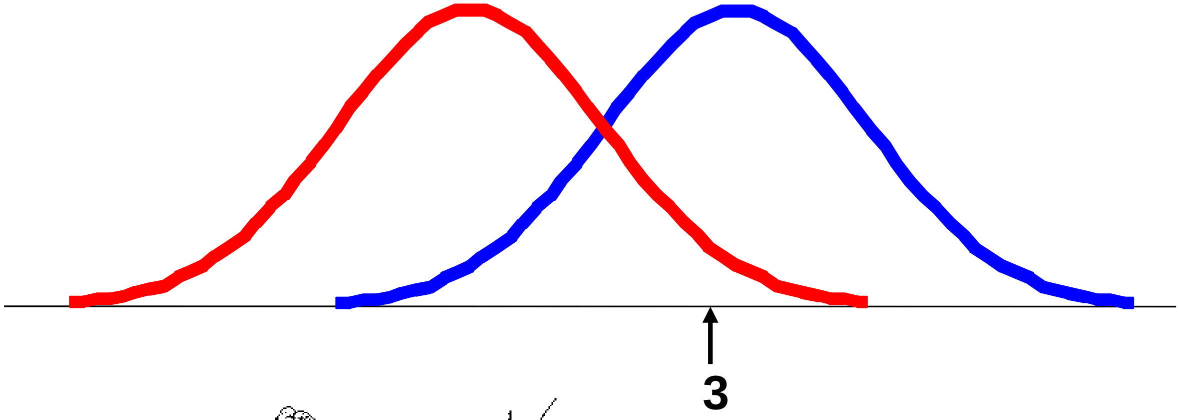


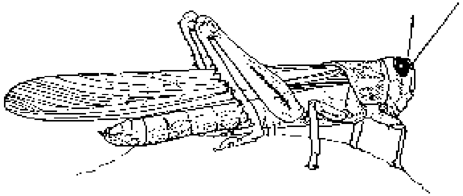
• We want to classify an insect we have found. Its antennae are 3 units long. How can we classify it?

• 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



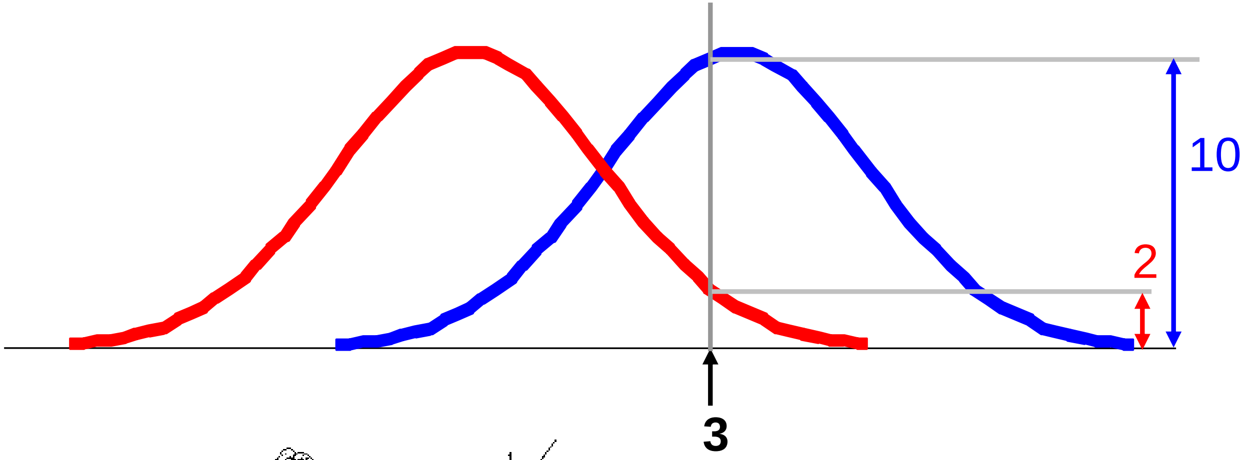


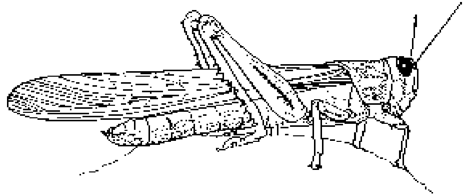
Antennae length is **3**

$p(c_j \mid d)$ = 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



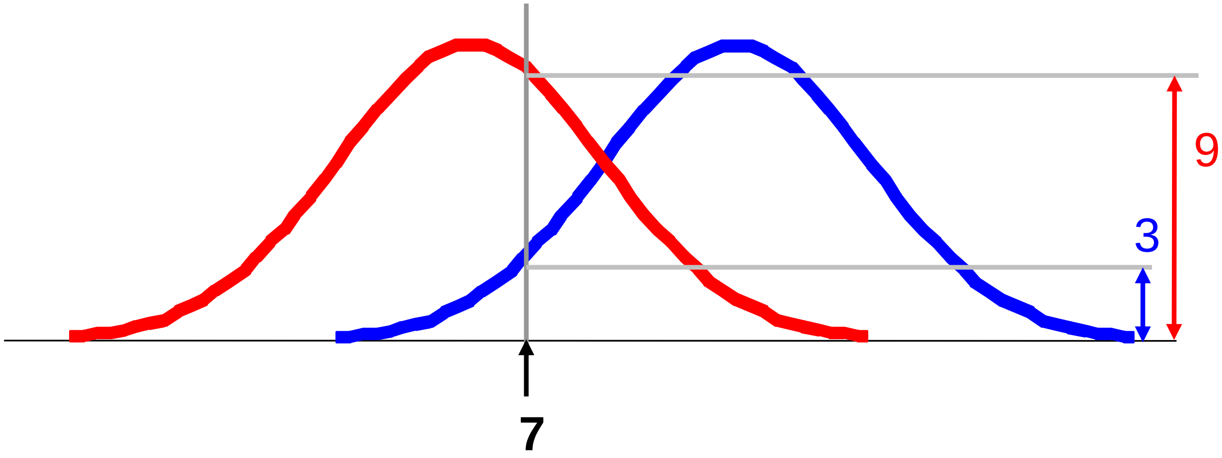


Antennae length is **3**

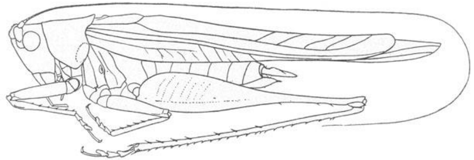
$p(c_j \mid d)$ = probability of class c_j, given that we have observed d

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

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



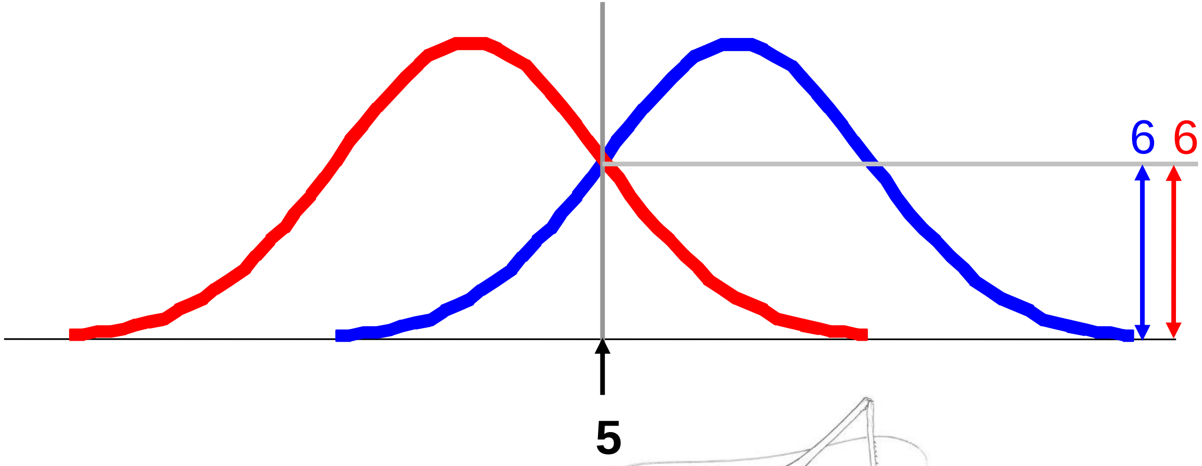
Antennae length is **7**



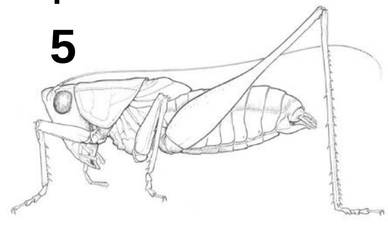
$p(c_j \mid d)$ = probability of class c_j, given that we have observed d

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.

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 $p(\text{male} \mid drew)$ 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?

What is the probability of being a male?

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

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

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

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

Officer Drew is more likely to be a Female.

Officer Drew

Officer Drew IS a female!

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

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

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

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

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

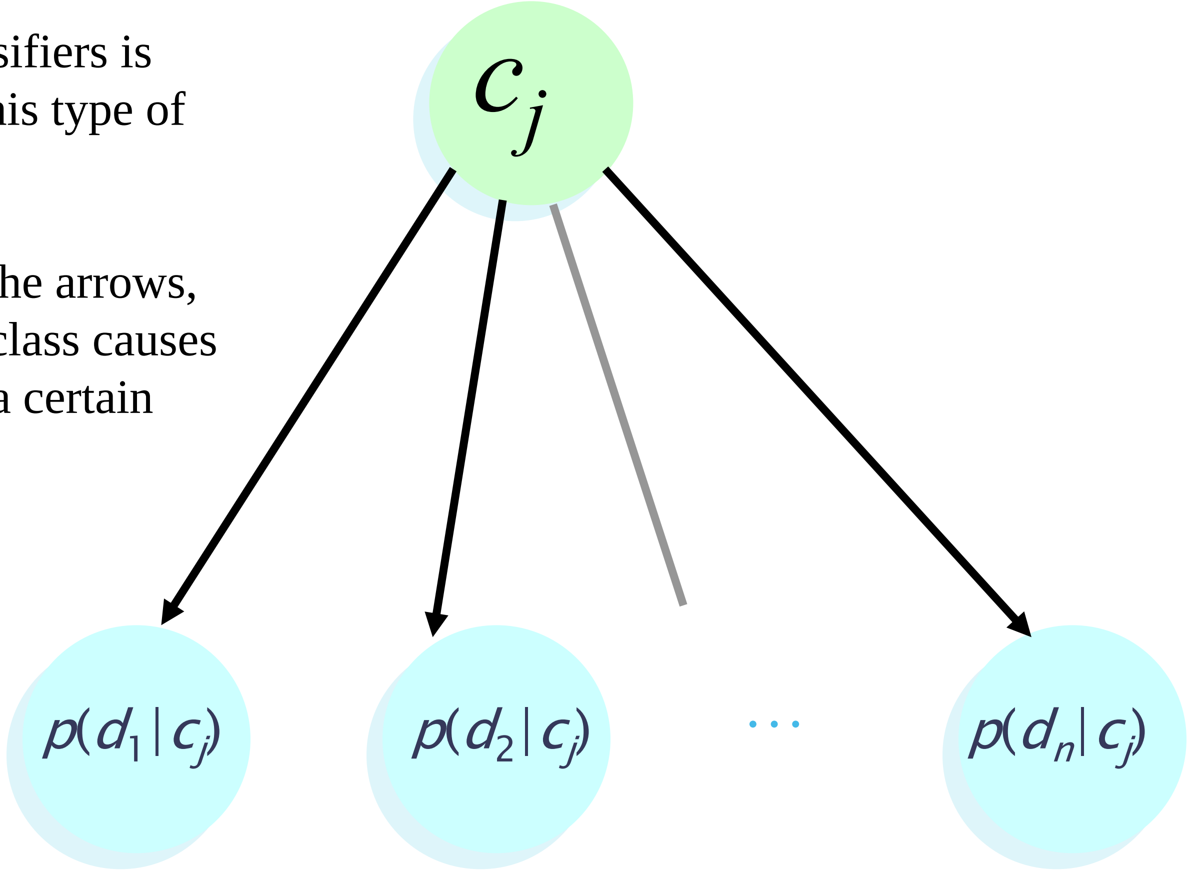
Officer Drew is blue-eyed, over 170_{cm} tall, and has long hair

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

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

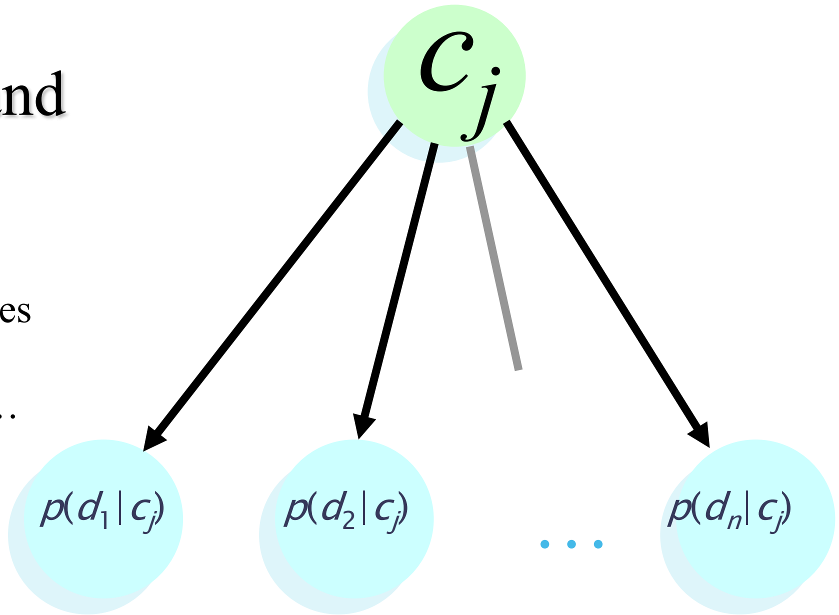
The Naive Bayes classifiers is often represented as this type of graph…

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…

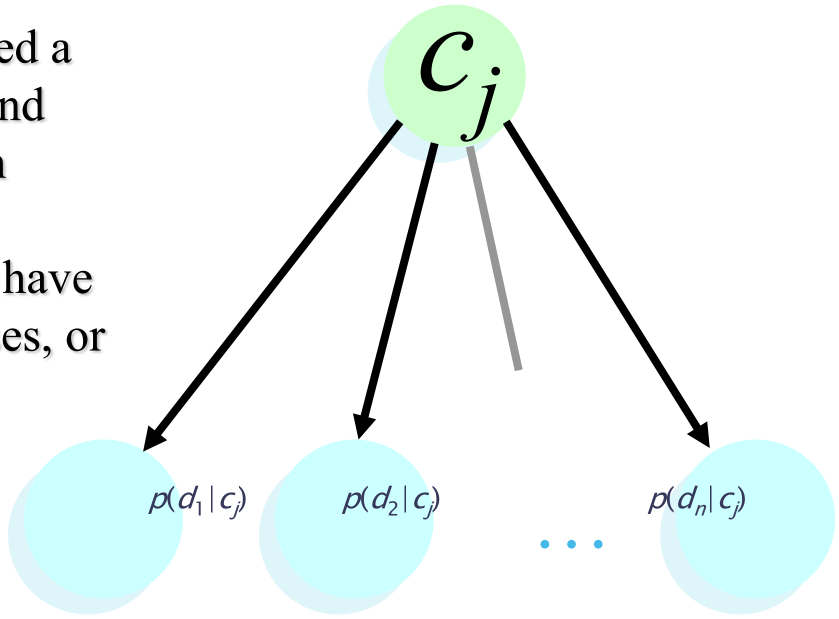


Sex	Over190cm	
Male	Yes	0.15
	No	0.85
Female	Yes	0.01
	No	0.99

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

Animal
Cat
Dog
Pig

Thanks ☺