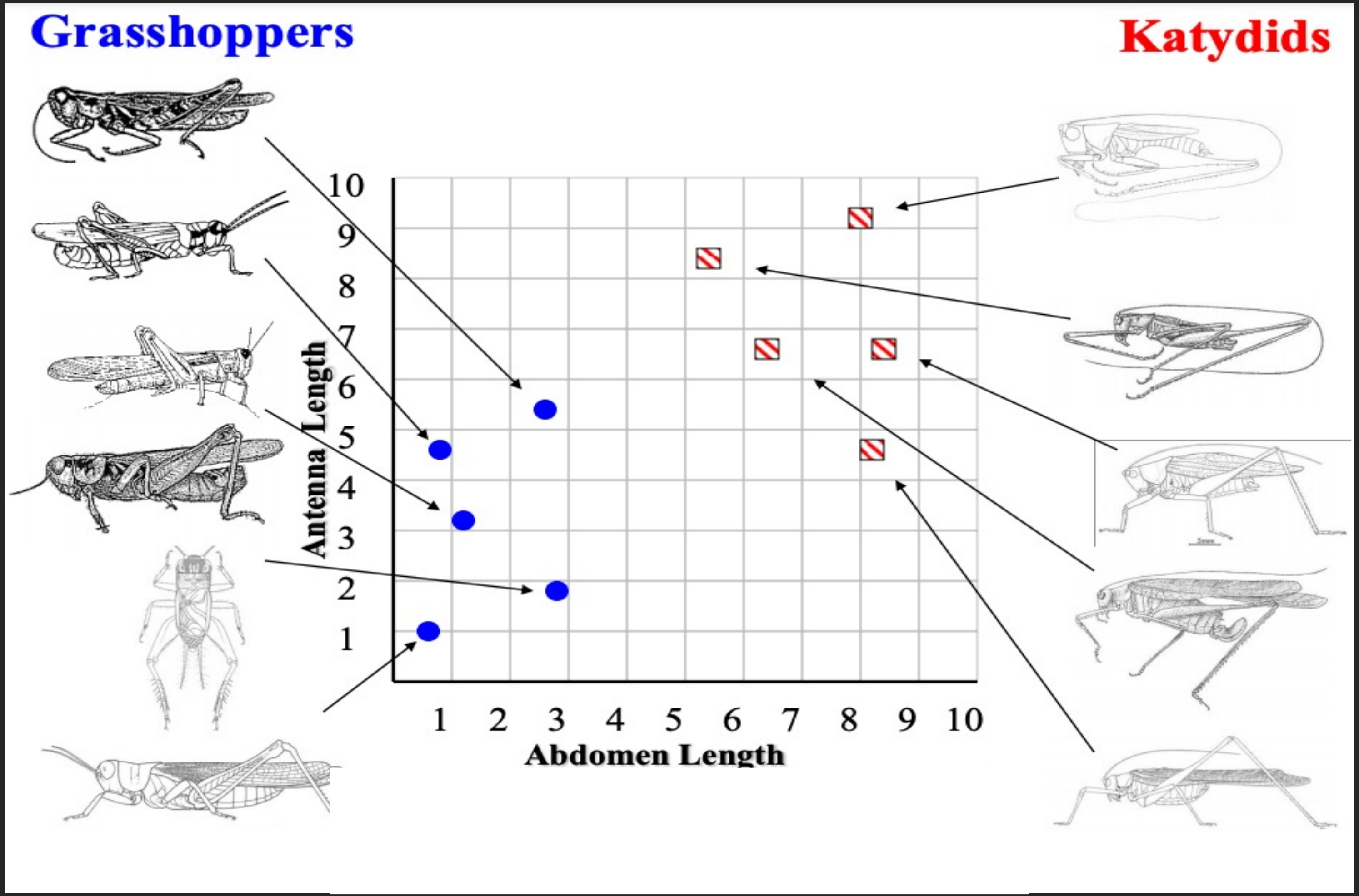
Applied Machine Learning

Bayes Classifier

Computer Science, Fall 2022
Instructor: Xuhong Zhang

Thomas Bayes
1702-1761

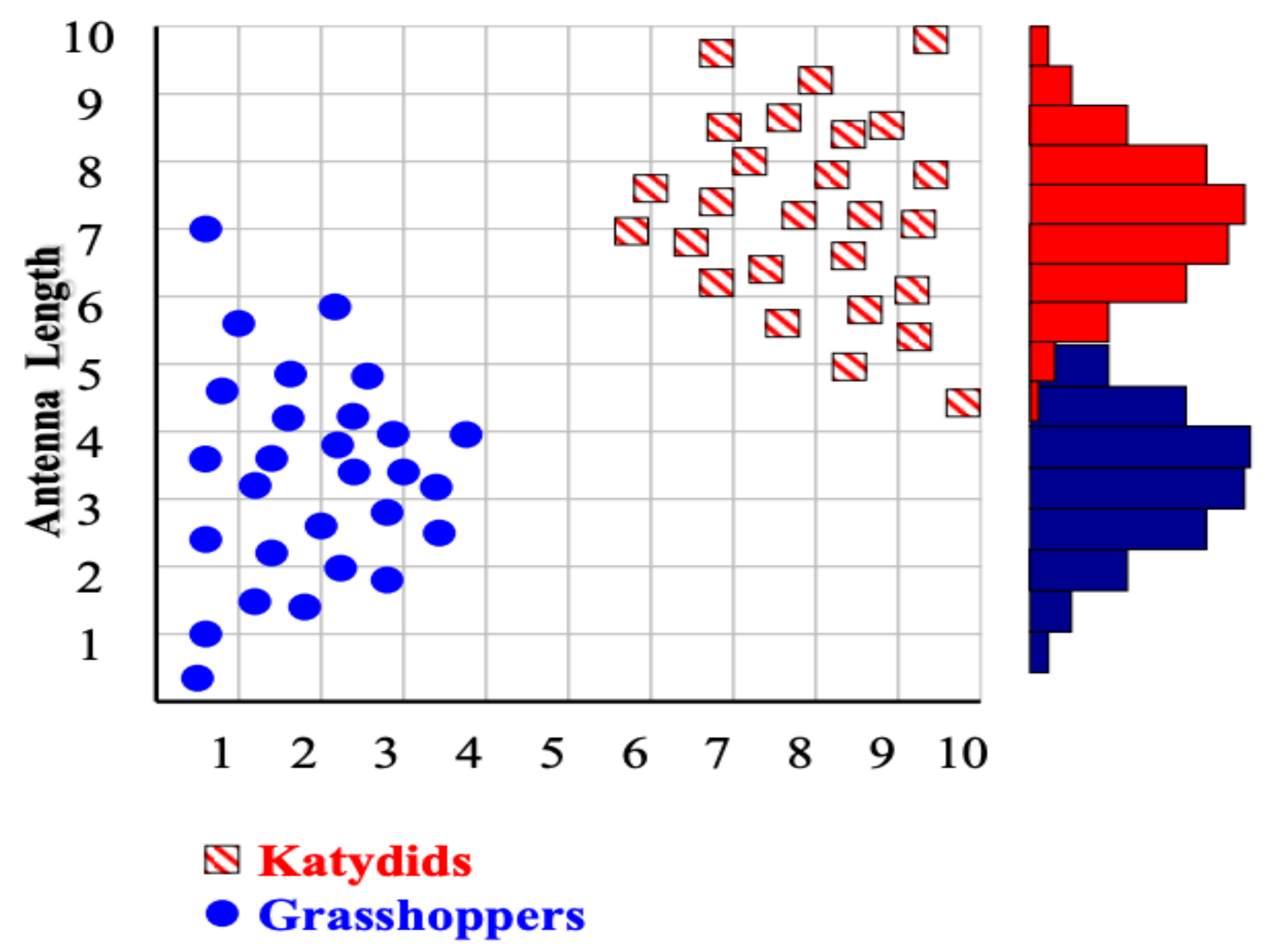
Visual Intuition



Slide credit: Eamonn Keogh

Visual Intuition

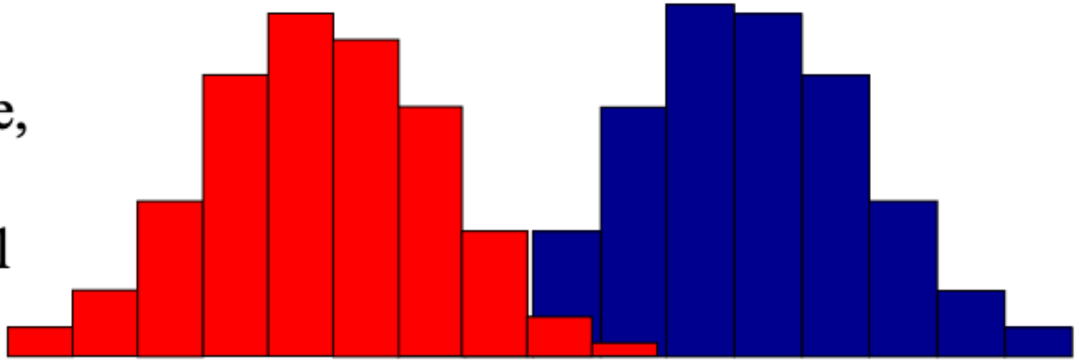
- We can build a histogram for "Antenna Length" with more data



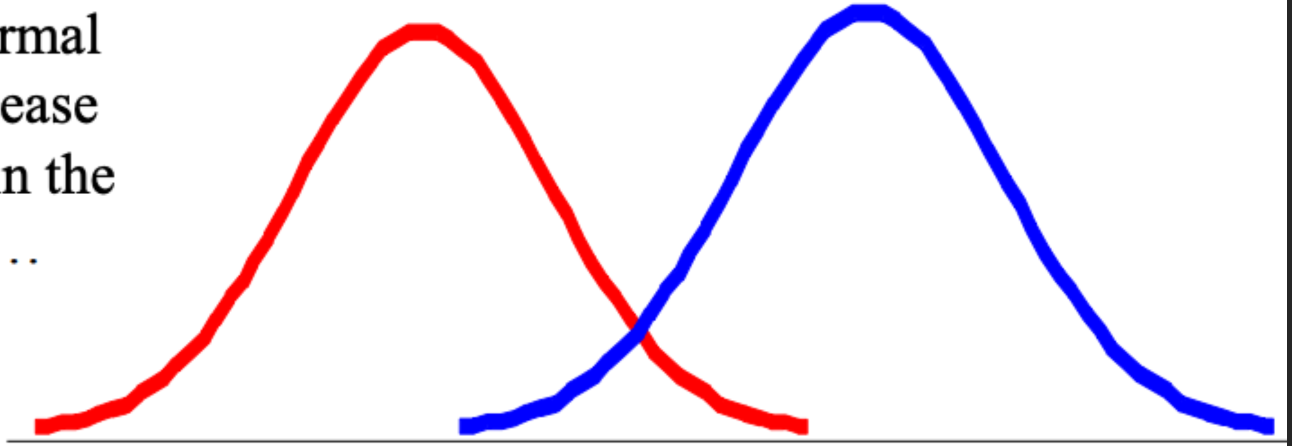
Slide credit: Eamonn Keogh

Visual Intuition

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…



Slide credit: Eamonn Keogh

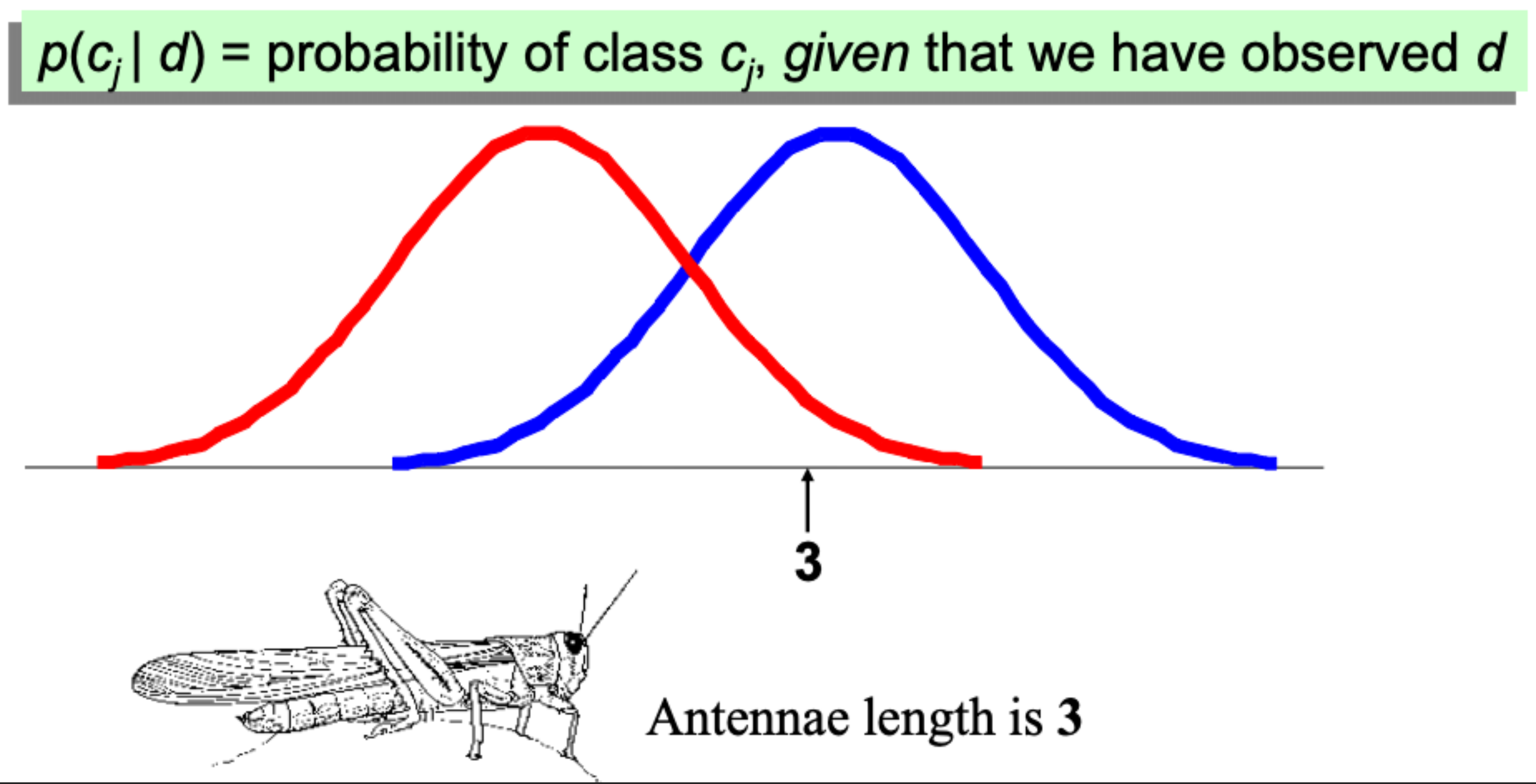
Visual Intuition

- Q: 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***

Visual Intuition

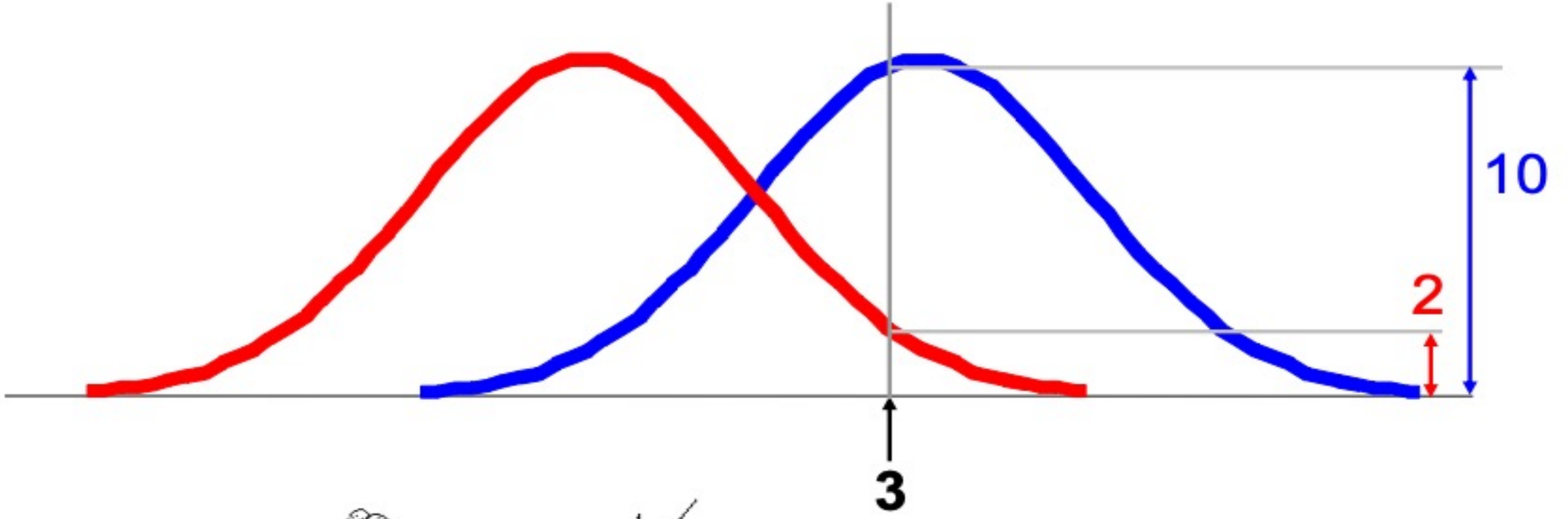


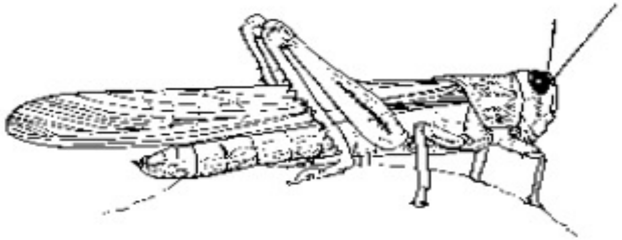
Slide credit: Eamonn Keogh

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

$$P(\textbf{Grasshopper} \mid 3) = 10 / (10 + 2) = 0.833$$

$$P(\textbf{Katydid} \mid 3) = 2 / (10 + 2) = 0.166$$



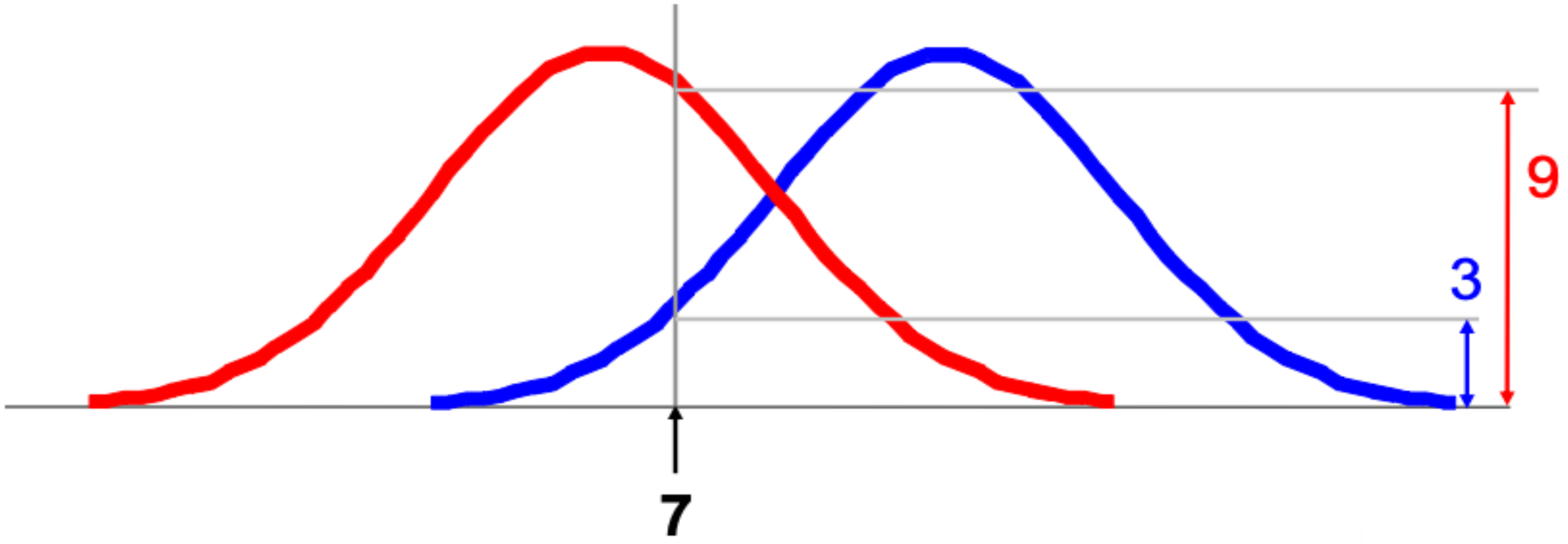


Antennae length is **3**

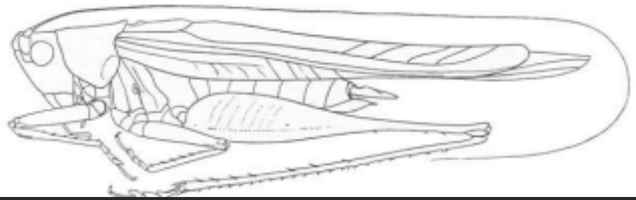
$p(c_j \mid d)$ = probability of class c_j, given that we have observed d

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

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



Antennae length is 7

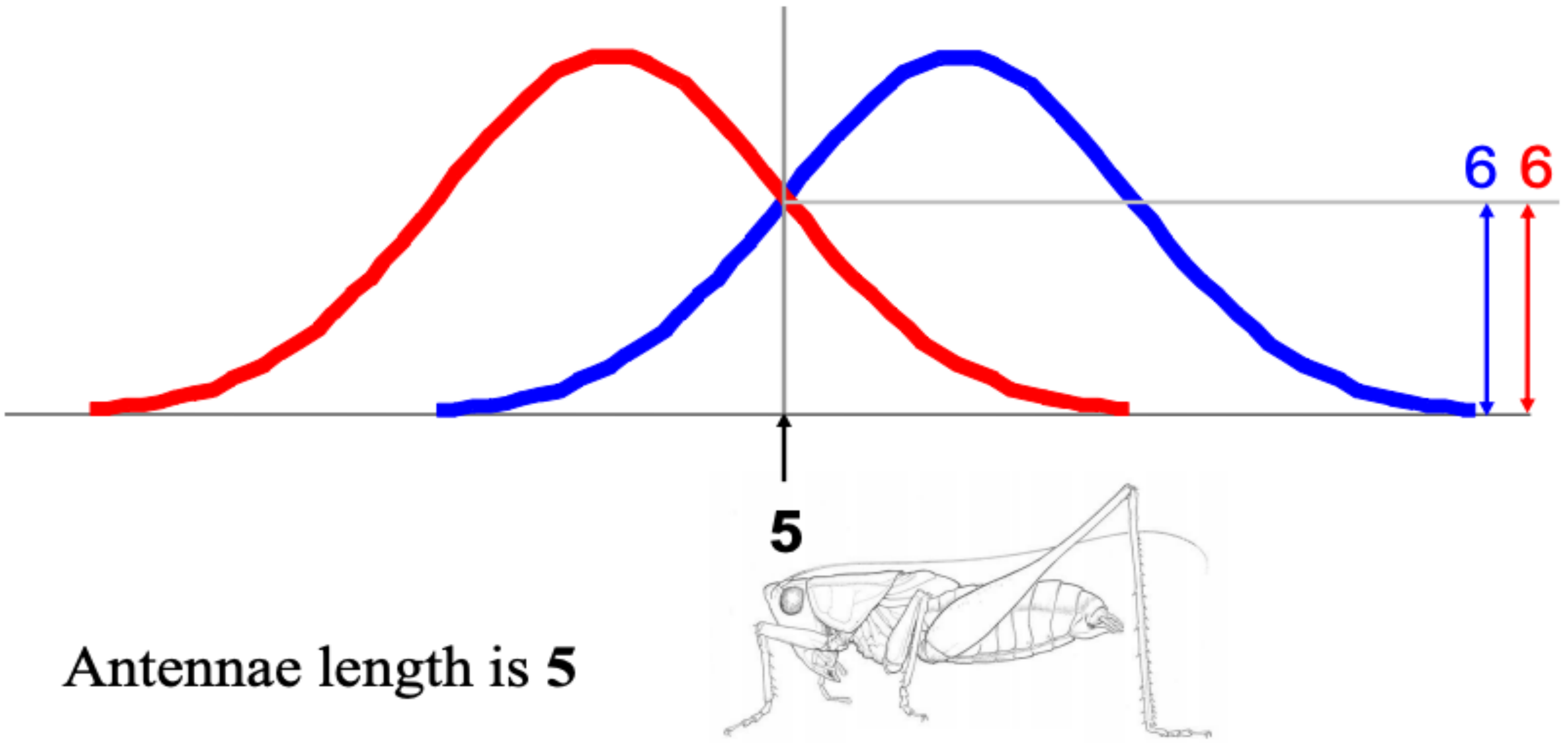


Slide credit: Eamonn Keogh

$p(c_j \mid d)$ = 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



Slide credit: Eamonn Keogh

Bayes Classifier

- Naïve Bayes
- Simple Bayes
- Idiot Bayes

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

Essential Probability Concepts

- Marginalization: $P(B) = \sum_{v \in Val(A)} P(B \wedge A = v)$
- Conditional Probability: $P(A|B) = \frac{P(A \wedge B)}{P(B)}$
- Bayes' Rule: $P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$
- Independence:

$$A \perp B \leftrightarrow P(A \wedge B) = P(A) \times P(B)$$
$$\leftrightarrow P(A|B) = P(A)$$
$$A \perp B|C \leftrightarrow P(A \wedge B|C) = P(A|C) \times P(B|C)$$

Bayes Classifier

- Bayesian classifiers use **Bayes theorem**, which says
  - $p(c_j|d) = \frac{p(d|c_j)p(c_j)}{p(d)}$
    - ✓ $p(c_j|d)$: probability of instance d being in class c_j
    - ✓ $p(d|c_j)$: probability of generating instance d given class c_j
    - ✓ $p(c_j)$: probability of occurrence of class c_j
    - ✓ $p(d)$: probability of instance d occurring

This is what we are trying to compute

We can imagine that being in class c_j, because we have feature d with some probability

This is just how frequent the class c_j is in our database

It is the same for all classes, so we can ignore it

Bayes Classifier

- Q: Assume we have two classes

$$C_1 = \text{male} \text{ and } C_2 = \text{female}$$

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

- Classifying drew as male or female is equivalent to ask is it more probable that drew is male or female, i.e. which is greater $p(\text{male}|drew)$ or $p(\text{female}|drew)$

Bayes Classifier

Note: Some name can be neutral. For example, “Taylor” can be a male or female name

- Female:
  - Taylor Mayne Pearl Brooks
  - Taylor-Anne Crichton
- Male:
  - Taylor Daniel Lautner
  - Tayler Michel Momsen

What is the probability of being called “drew” given that you are a male?

What is the probability of being a male?

What is the probability of being calle “drew” given at you are a male

$$p(\mathrm{male}|drew) = \frac{p(drew|\mathrm{male})p(\mathrm{male})}{o(drew)}$$

What is the probability of being named “drew”?

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

We can apply Bayes rule to figure out which category is more likely:

$$p(\text{male}|drew) = \frac{p(drew|\text{male})p(\text{male})}{o(drew)}$$

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

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

Bayes Classifier

- So far, we considered one feature (the "antennae length" or the "name") of the data.

- What if we have more than one feature? And how to use all the features?

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

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

Bayes Classifier

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

$$p(x|c_j) = p(x^1|c_j) \times p(x^2|c_j) \times \cdots \times p(x^d|c_j)$$

The probability of class Cj generating instance x, equals…

The probability of class Cj generating the observed value for feature 1, multiplied by…

The probability of class Cj generating the observed value for feature 2, multiplied by…

Bayes Classifier

$$p(x|c_j) = p(x^1|c_j)\times p(x^2|c_j)\times \cdots \times p(x^d|c_j)$$

In case for Officer Drew, the features are blue-eyed, over 170 cm tall, and has long hair

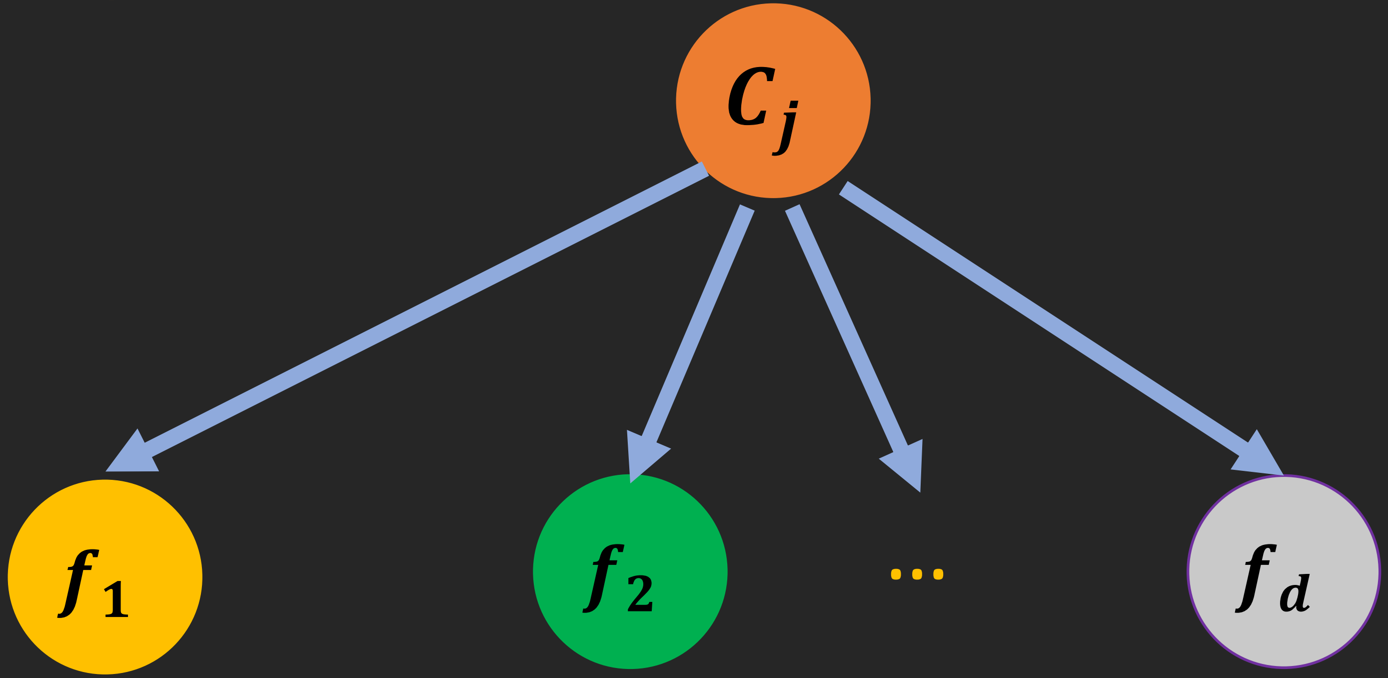
$$p(\text{officer drew}|c_j) = p(\text{over } 170cm = yes|c_j)\times p(\text{eye color} = \text{blue}|c_j)\times \cdots$$

$$p(\text{officer drew}|\text{Female}) = \frac{2}{5}\times\frac{3}{5}\times\cdots$$

$$p(\text{officer drew}|\text{male}) = \frac{2}{3}\times\frac{2}{3}\times\cdots$$

Bayes Classifier

- Graphic representation



The arrow indicates a class condition, and each class causes certain features with a certain probability

Bayes Classifier

- **Properties**
  - Naïve Bayes is not sensitive to irrelevant features

Suppose we are trying to classify a person's sex based on some features, including eye color. (eye color is completely irrelevant to a a person's gender)

$$p(\text{Jessica}|c_j) = p(\text{wears\_dress} = yes|c_j) \times p(\text{eye color} = \text{brown}|c_j) \times \cdots$$

$$p(\text{Jessica}|\text{Female}) = \frac{9{,}975}{10{,}000} \times \frac{9{,}000}{10{,}000} \times \cdots$$

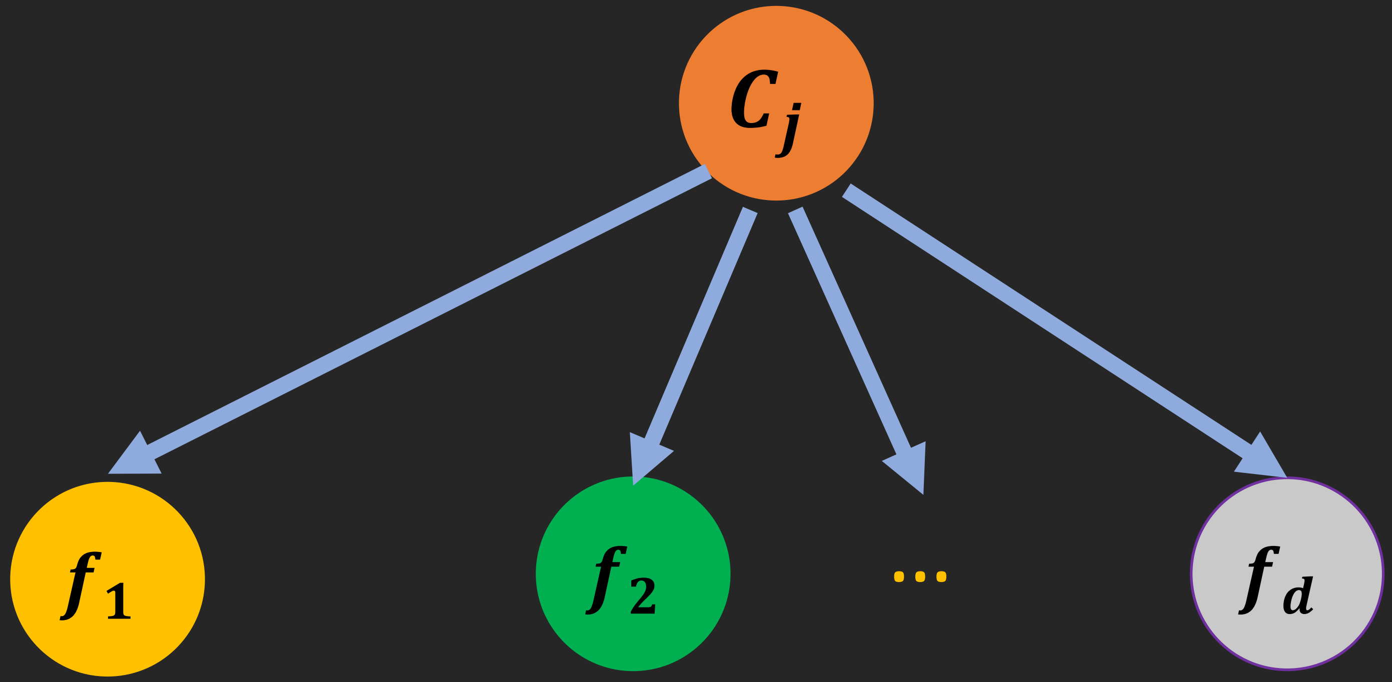
$$p(\text{Jessica}|\text{male}) = \frac{2}{10{,}000} \times \frac{9{,}001}{10{,}000} \times \cdots$$

Almost the same

Assumption: good enough estimates of the probabilities -- the more data the better.

Bayes Classifier

- **Properties (cont.)**
  - It is fast and space efficient



With a single scan of the entire dataset, we can look up all the probabilities and store them in a table.

For d_1, d_2,:

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

Similarly for all the other features.

Bayes Classifier

- But be cautious here:
  Naïve Bayes assumes independence of features

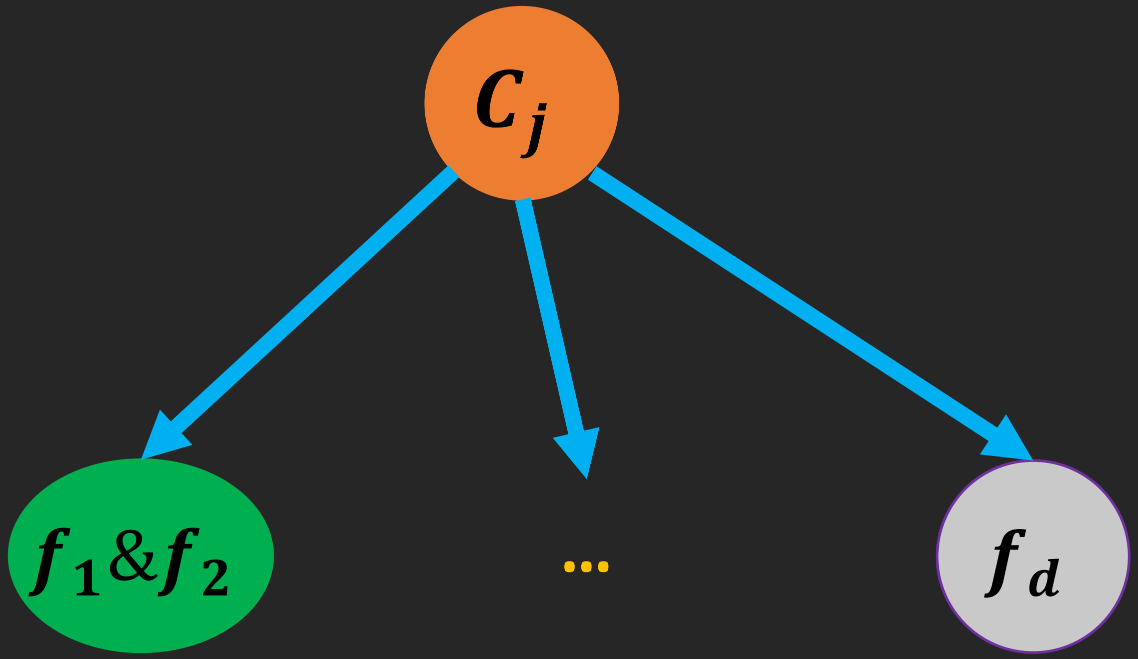
Sex	Over 6 foot	
Male	Yes	0.15
	No	0.85
Female	Yes	0.01
	No	0.99

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

Question:
Can the feature "Over 6 foot" and "Over 200 pounds" be completely independent to each other?

Bayes Classifier

- Solution: Consider the relationships between features.



Sex	Over 200 pounds & Over 6 foot	
Male	Yes and Yes	0.11
	No and Yes	0.59
	Yes and No	0.05
	No and No	0.35

Bayes Classifier

- Another disadvantage:
  - When we estimate probabilities, sometimes we estimate them by counting from the training data. **But counting might be zero.**
  - Fix by using Laplace smoothing : adding 1 to each count

$$P(d_i = v | C_j = k) = \frac{c_v + 1}{\sum_{v' \in Val(d_i)} c_{v'} + |values(d_i)|}$$

- ❑ c_v is the count of training instance with a value of v for attribute i and class label k.
- ❑ $\sum_{v' \in Val(d_i)} c_{v'}$ is the number of instances for class k.
- ❑ $|values(d_i)|$ is the number of values d_i can take on

Bayes Classifier

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	false	no
Sunny	80	90	true	no
Overcast	83	86	false	yes
Rainy	70	96	false	yes
Rainy	68	80	false	yes
Rainy	65	70	true	no
Overcast	64	65	true	yes
Sunny	72	95	false	no
Sunny	69	70	false	yes
Rainy	75	80	false	yes
Sunny	75	70	true	yes
Overcast	72	90	true	yes
Overcast	81	75	false	yes
Rainy	71	91	true	no

Q:
x = {overcast, 66, 90, true}
Play or not play?

- Prior probability for target variable : Play

P(play = yes) = $9/14$
P(play = no) = $5/14$

Bayes Classifier

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	false	no
Sunny	80	90	true	no
Overcast	83	86	false	yes
Rainy	70	96	false	yes
Rainy	68	80	false	yes
Rainy	65	70	true	no
Overcast	64	65	true	yes
Sunny	72	95	false	no
Sunny	69	70	false	yes
Rainy	75	80	false	yes
Sunny	75	70	true	yes
Overcast	72	90	true	yes
Overcast	81	75	false	yes
Rainy	71	91	true	no

Q:
x = {overcast, 66, 90, true}
Play = yes or no?

- Likelihood $p(x|play)$

$$P(x|yes) = p(overcast|yes) \times p(66|yes) \times p(90|yes) \times p(true|yes)$$

Bayes Classifier

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	false	no
Sunny	80	90	true	no
Overcast	83	86	false	yes
Rainy	70	96	false	yes
Rainy	68	80	false	yes
Rainy	65	70	true	no
Overcast	64	65	true	yes
Sunny	72	95	false	no
Sunny	69	70	false	yes
Rainy	75	80	false	yes
Sunny	75	70	true	yes
Overcast	72	90	true	yes
Overcast	81	75	false	yes
Rainy	71	91	true	no

Q:

x = {overcast, 66, 90, true}

Play = yes or no?

- Likelihood $\mathrm{p(x|play)}$

$$P(\mathrm{x|no}) = \overbrace{\boxed{\mathrm{p(overcast|no)}}}^{0} \times \mathrm{p(66|no)} \times \mathrm{p(90|no)} \times \mathrm{p(true|no)}$$

Bayes Classifier

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	false	no
Sunny	80	90	true	no
Overcast	83	86	false	yes
Rainy	70	96	false	yes
Rainy	68	80	false	yes
Rainy	65	70	true	no
Overcast	64	65	true	yes
Sunny	72	95	false	no
Sunny	69	70	false	yes
Rainy	75	80	false	yes
Sunny	75	70	true	yes
Overcast	72	90	true	yes
Overcast	81	75	false	yes
Rainy	71	91	true	no

Q:

x = {overcast, 66, 90, true}

Play = yes or no?

- Likelihood $\mathrm{p(x|play)}$

0

$$P(\mathrm{x|no}) = \boxed{\mathrm{p(overcast|no)}} \times \mathrm{p(66|no)} \times \mathrm{p(90|no)} \times \mathrm{p(true|no)}$$

Bayes Classifier

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	false	no
Sunny	80	90	true	no
Overcast	83	86	false	yes
Rainy	70	96	false	yes
Rainy	68	80	false	yes
Rainy	65	70	true	no
Overcast	64	65	true	yes
Sunny	72	95	false	no
Sunny	69	70	false	yes
Rainy	75	80	false	yes
Sunny	75	70	true	yes
Overcast	72	90	true	yes
Overcast	81	75	false	yes
Rainy	71	91	true	no

Q:

x = {overcast, 66, 90, true}

Play = yes or no?

- Likelihood $p(x|play)$

$$p(overcast|no) = (0+1)/(5+3)$$

- 0 instance where

Outlook = overcast and play = no

- Total instances is 5 where play=no
- Outlook has 3 unique values

Bayes Classifier

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	false	no
Sunny	80	90	true	no
Overcast	83	86	false	yes
Rainy	70	96	false	yes
Rainy	68	80	false	yes
Rainy	65	70	true	no
Overcast	64	65	true	yes
Sunny	72	95	false	no
Sunny	69	70	false	yes
Rainy	75	80	false	yes
Sunny	75	70	true	yes
Overcast	72	90	true	yes
Overcast	81	75	false	yes
Rainy	71	91	true	no

Q:

x = {overcast, 66, 90, true}

Play = yes or no?

- Likelihood $\mathrm{p(x|play)}$

$$\mathrm{p(overcast|yes)} = (4 + 1)/(9 + 3)$$

- 4 instances where Outlook = overcast and play = yes
- Total instances is 9 where play=yes
- Outlook has 3 unique values

Bayes Classifier

- In practice, we use ***log-probabilities*** to prevent ***underflow***.

In practice, the independence assumption doesn't often hold true, but Naïve Bayes performs very well despite it

$$\log(P(x)) = \text{argmax}\, P(C = k) + \sum_{i=1}^{d} \log P(x^i | Y = k)$$

- For each class label k
  - Estimate $P(C = k)$ from the data
  - For each value $x^{i,j}$ of each attribute x^i
    - Estimate $P(x^{i,j} | C = k)$

Bayes Classifier: prediction

Play?	P(Play)
yes	3/4
no	1/4

Temp	Play?	P(Temp | Play)
warm	*yes*	4/5
cold	*yes*	1/5
warm	*no*	1/3
cold	*no*	2/3

Sky	Play?	P(Sky | Play)
sunny	*yes*	4/5
rainy	*yes*	1/5
sunny	*no*	1/3
rainy	*no*	2/3

Humid	Play?	P(Humid | Play)
high	*yes*	3/5
norm	*yes*	2/5
high	*no*	2/3
norm	*no*	1/3

Question: Predict label for x = (rainy, warm, normal)

$$P(play|x) \propto \log P(\text{play}) + \log P(\text{rainy}|\text{play}) + \log P(\text{warm}|\text{play}) + \log P(\text{normal}|\text{play})$$
$$\propto \log\frac{3}{4} + \log\frac{1}{5} + \log\frac{4}{5} + \log\frac{2}{5} = -1.319$$

$$P(\neg play|x) \propto \log P(\neg\text{play}) + \log P(\text{rainy}|\neg\text{play}) + \log P(\text{warm}|\neg\text{play}) + \log P(\text{normal}|\neg\text{play})$$
$$\propto \log\frac{1}{4} + \log\frac{2}{3} + \log\frac{1}{3} + \log\frac{1}{3} = -1.732$$

Predict PLAY

Bayes Classifier

- Can also be used for computing probabilities (Not just predicting labels)
  - NB classifier gives predictions, not probabilities, because we ignore $P(X)$ (the denominator in Bayes rule)
  - For each possible class label c_k, the class probability is given by

$$P(C = k|X = x) = \frac{P(C = k)\prod_{i=1}^{d} P(X_i = x_i|C = k)}{\sum_{k'=1}^{\#\, of\ classes} P(C = k')\prod_{i=1}^{d} P(X_i = x_i|C = k')}$$

Bayes Classifier

- For **continuous inputs** X, we can also use the previous argmax as the basis for designing a Naïve Bayes classifier.
  - **One common approach** is to assume that for each discrete value k of C, the distribution of each continuous x^i is Gaussian, and is defined by a mean and standard deviation specific X^i and k.
  - Then we must estimate the mean and standard deviation of each of these Gaussians:

$$\mu_{ik} = E\left[X^i \middle| C = k\right]$$
$$\sigma_{ik}^2 = E[(X^i - \mu_{ik})^2 | C = k]$$

for each attribute $X^i$and each possible value k of C.

Bayes Classifier

- We must also estimate the priors on C as well

$$\pi_k = P(C = k)$$

✓The model summarizes a Gaussian Naïve Bayes classifier, which assumes that the data X is generated by a mixture of class-conditional (i.e., dependent on the value of the class variable Y) Gaussians.
✓ The naïve Bayes assumption introduces the additional constraint that the attribute values X^i are independent of one another within each of these mixture components.
✓ We might introduce additional assumptions to further restrict the number of parameters or the complexity of estimating them.

➢ For example, we can assume that noise in the observed X^i comes From a common source, then all of the σ_{ik} are identical, regardless of the attribute i or class k.