CLASSIFICATION BY STATISTICAL MODELING NAIVE BAYES

Vic Ciesielski
Department of Computer Science
RMIT
vic.ciesielski@rmit.edu.au

SUMMARY

1. The univariate (one variable) normal

2. Classification with Bayes rule

3. Bivariate (two variables) and multivariate (two or more) normal

4. Naive Bayes - Numeric attributes

5. Naive Bayes - Nominal attributes

6. Evaluation of Classifiers

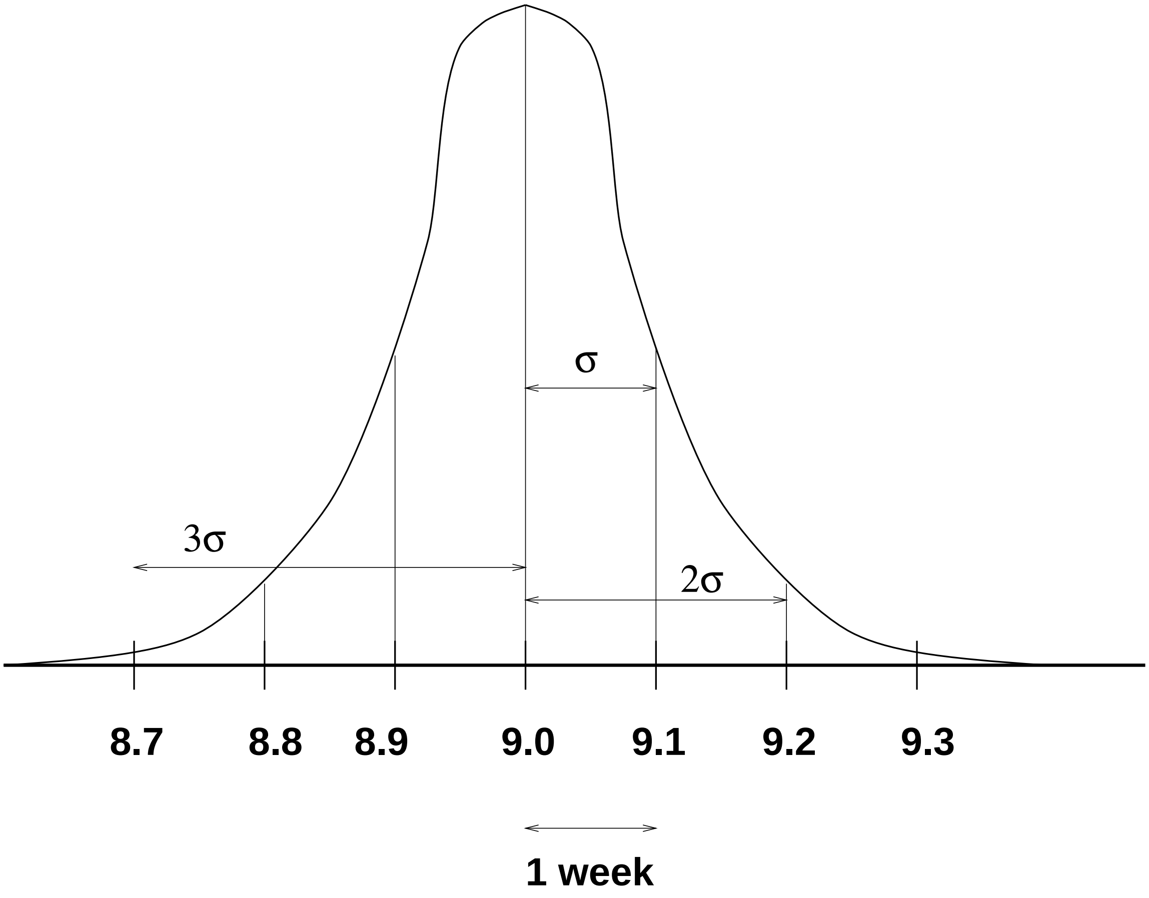
7. Missing Values

8. Unbalanced Data

THE NORMAL (GAUSSIAN) DISTRIBUTION

Most naturally occuring data, if enough of it is collected, fits the normal distribution.

Example: Length of a human pregnancy



Parameter	Symbol	Example	Estimator
mean	μ	9 months	average
variance	σ	1 week	standard deviation

THE NORMAL DISTRIBUTION

Equation of the ‘bell’ curve:

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}$$

For a normal distribution with $\mu = 270$ and $\sigma = 7$

Probability of a pregnancy being 280 days (9 months and 1 week) is:

$$\frac{1}{7\sqrt{2\pi}}e^{-\frac{1}{2}(280-270)^2/7^2} = 0.02$$

CLASSIFICATION BASED ON NORMAL DISTRIBUTIONS

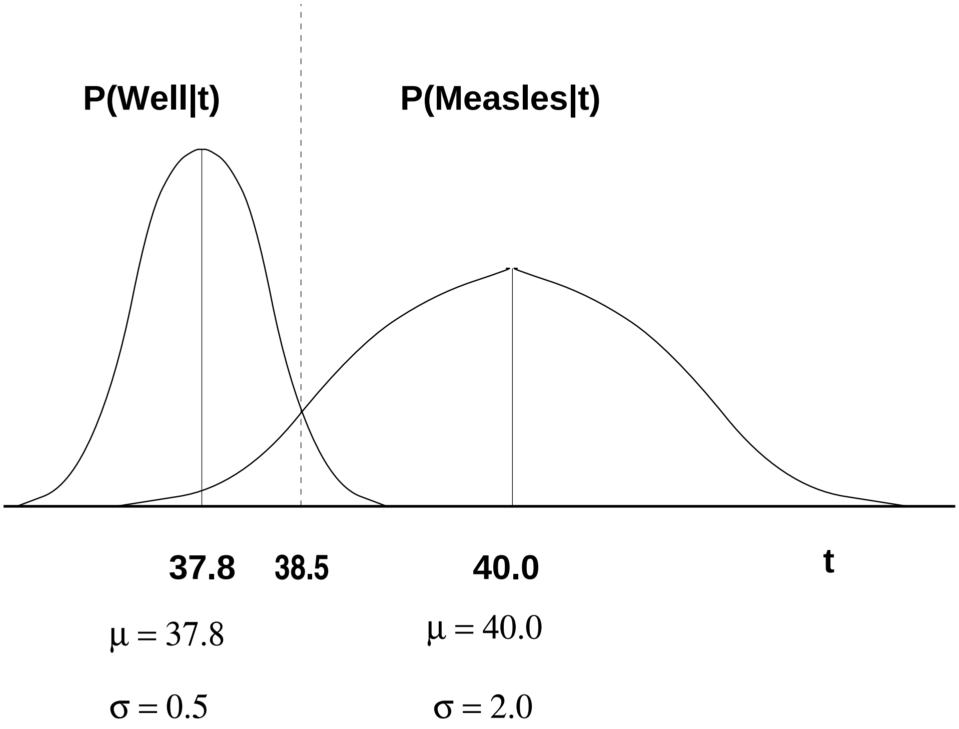
Suppose we want to tell the difference between people who have measles and those who are well. How to proceed?

1. Find 1000 (say) people
2. Divide them into 2 groups, measles and well
   [Why not 500 of each?]
3. Measure the temperature of everybody in both groups
4. Estimate the mean and variance of ‘well’ by computing average and standard deviation

   $P(t|well)$ is normal, mean 37.8; variance 0.5
5. Estimate the mean and variance of ‘measles’ by computing average and standard deviation

   $P(t|measles)$ is normal, mean 40.0; variance 2.0
6. Derive $P(well|t)$ from $P(t|well)$ (Bayes Rule)
7. Derive $P(measles|t)$ from $P(t|measles)$ (Bayes)
8. Plot both curves

NORMAL CLASSIFICATION 2



9. Decide on the decision boundary

   (a) Maximum Likelihood

   if $P(measles|t) \geq P(well|t)$ decide measles

   if $P(well|t) > P(measles|t)$ decide well

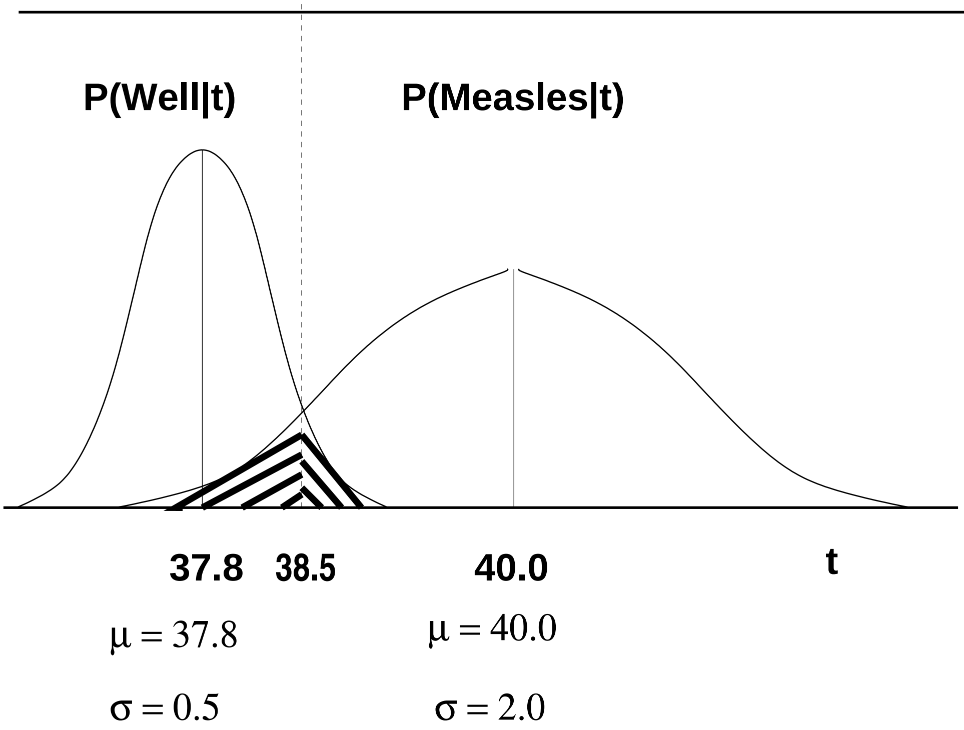
   A person with a temperature of less than 38.5 will be classified as ‘well’

   A person with a temperature of 38.5 or greater will be classified as ‘measles’

   What will be the error rate?

   Is this a good classifier?

MAXIMUM LIKELIHOOD



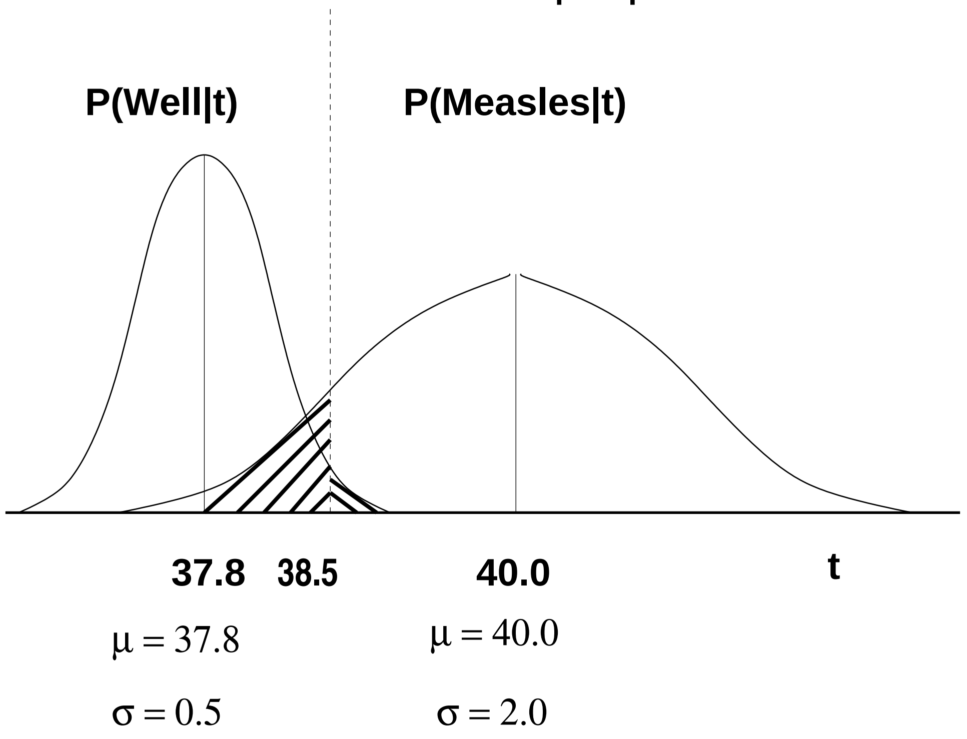
- For any t choose the class with the largest probability
  - False positive error: Conclude measles when person is well (right)
  - False negative error: Conclude well when person has measles (left)
- Confusion Matrix

	Predicted measles	Predicted well
Actual measles	True Positive (TP)	False Negative (FN)
Actual well	False Positive (FP)	True negative (TN)

MINIMUM COST

(b) Weighted to minimize misclassification cost

- Suppose FP is 2 times as costly as FN.
- Move the treshold in proportion to cost

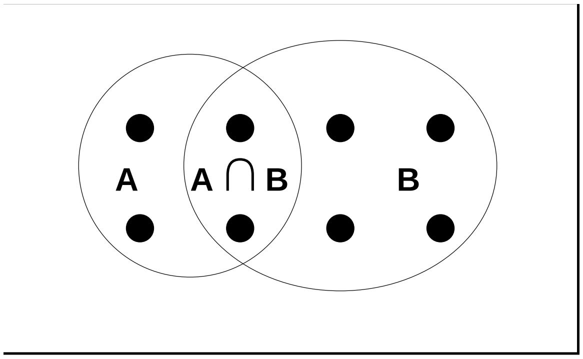


- What would be relative costs for appendicitis?
- Decision boundary shifts left or right based on the cost of errors
- Lower FP error, but higher FN error and higher total error

10. Bayes classifier (*Discriminant Functions* in classical pattern recognition)

11. Note: Depends on assumption of normal distribution.

BAYES THEOREM



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

A little algebra gives

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

Where$P(B) = P(B|A).P(A) + P(B| \sim A).P(\sim A)$

Let $A =$ Measles, $\sim A =$ Well, $t =$ temperature

$$P(M|t) = \frac{P(t|M).P(M)}{P(t|M).P(M) + P(t|W)P(W)}$$

All terms in the RHS are known and curve for $P(Measles|t)$ can be found (analytically)

BAYES THEOREM 2

- A common presentation of Bayes rule relates evidence to a hypothesis

$$P(H|E) = \frac{P(E|H).P(H)}{P(E|H).P(H) + P(E| \sim H).P(\sim H)}$$

- or (equivalent) Let $H = C_1$ and $\sim H = C_2$

$$P(C_1|E) = \frac{P(E|C_1).P(C_1)}{\sum_{j=1}^{2} P(E|C_j).P(C_j)}$$

- If $\vec{x}$ is a vector of data and $C_i, i = 1..k$ are classes we have

$$P(C_i|\vec{x}) = \frac{P(\vec{x}|C_i).P(C_i)}{\sum_j P(\vec{x}|C_j).P(C_j)}$$

BAYES CLASSIFIER

- For a 2 class problem:
  If $P(C_1|\vec{x}) \geq P(C_2|\vec{x})$ we decide C_1 else C_2

- *Key Observation*: When substituting for these probabilities according to equation on previous page we find that the denominators are the *same*.

- So we just need to look at:

  $P(\vec{x}|C_1).P(C_1) \geq P(\vec{x}|C_2).P(C_2)$

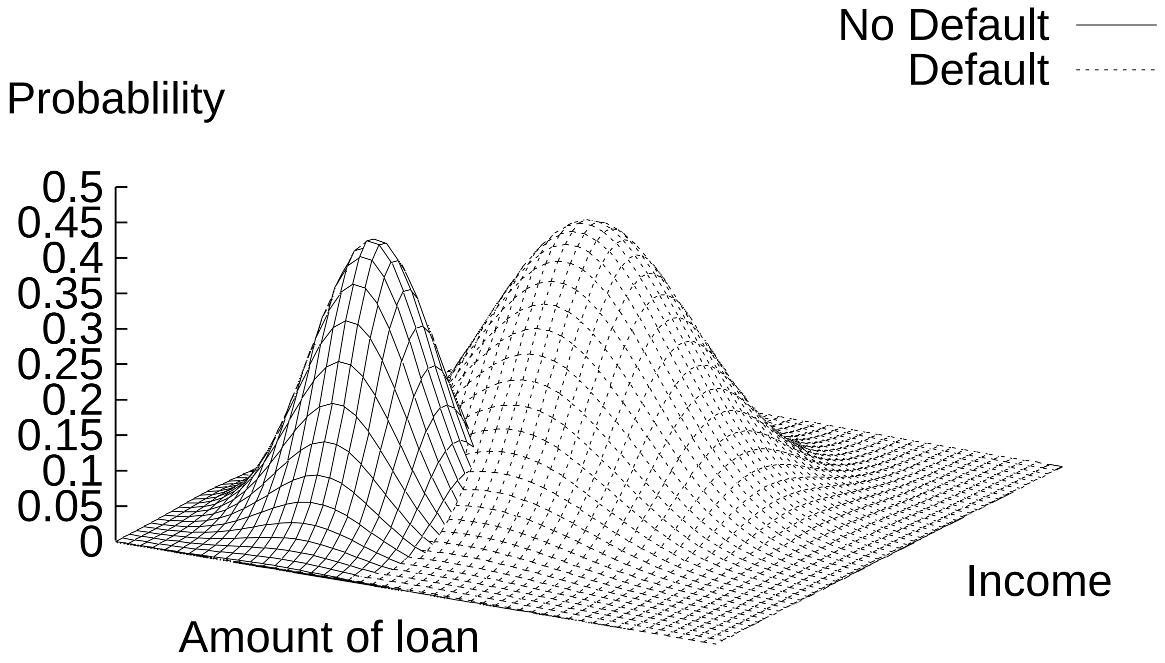
  $P(C_1|\vec{x})$ A posteriori probability of C_1
  After evidence

  $P(C_1)$ Prior (apriori) probability of C_1.
  Before evidence
  Estimated by counting in training data

- Same number in each class: $P(C_1) = P(C_2)$
  Calculation reduces to

  $P(\vec{x}|C_1) \geq P(\vec{x}|C_2)$

CLASSIFICATION WITH BIVARIATE NORMAL



- We have $P(No\ Default)|Amount, Income)$

  and $P(Default)|Amount, Income)$

- We need equations for both of these curves to work out the decision curve

- Note: Income and loan are not independent

CLASSIFICATION USING STATISTICAL MODELS

The general approach is:

1. Collect the data
2. Assign the classes
3. Decide on the probability distributions that best 'model' the data
   - Normal/gaussian for numeric data
   - Binomial for binary (yes/no) data
   - Multinomial for other symbolic data
4. Using the available data, get estimates of the model parameters. For each class we will have

   $P(\vec{x}|CLASS_i)$
   $P(CLASS_i)$ (The prior probability)
5. Using Bayes rule get

   $P(CLASS_i|\vec{x})$

   (Could be difficult to do this analytically. Usually requires assuming attributes are independent)
6. Construct decision rule (Max likelihood, min cost)

NON NUMERIC ATTRIBUTES 1

- Given the sunburn data the task is determine whether Jack will get sunburnt.

name	Hair	Height	Weight	Lotion	Result
Sara	blonde	average	light	no	SB
Dana	blonde	tall	average	yes	NB
Alex	brown	short	average	yes	NB
Annie	blonde	short	average	no	SB
Emily	red	average	heavy	no	SB
Peter	brown	tall	heavy	no	NB
John	brown	average	heavy	no	NB
Katie	blonde	short	light	yes	NB
Jack	blonde	average	light	no	??

- Jack's feature vector (j) is $[Hair = blonde, Height = average, Weight = light, Lotion = no]$
- Decision rule
if $P(SB|j) > P(NB|j)$ then SB else NB
i.e
$P(j|SB).P(SB) > P(j|NB).P(NB)$
- We can easily get $P(NB)$ and $P(SB)$ from counting occurrences in the training data.

  $P(NB) = 5/8$, $P(SB) = 3/8$

NON NUMERIC ATTRIBUTES 2

- If we assume the attributes are independent we replace

  $$P(hair = blonde \;\&\; height = average \;\&\; weight = light \;\&\; lotion = no|SB)$$

  with

  $$P(hair = blonde|SB) * P(height = average|SB) * P(weight = light|SB) * P(lotion = no|SB)$$

- We can get each term by counting in the training data

  $P(hair = blonde|SB) = 2/3$
  $P(height = average|SB) = 2/3$
  $P(weight = light|SB) = 1/3$
  $P(lotion = no|SB) = 3/3$

- Thus $P(SB|jack) = (2/3)*(2/3)*(1/3)*(3/3)*(3/8) = 1/18$

- Similarly $P(NB|jack) = (2/5)*(1/5)*(1/5)*(2/5)*(5/8) = 1/250$

  Thus Jack is SB.

COMBINING NUMERIC AND NOMINAL ATTRIBUTES

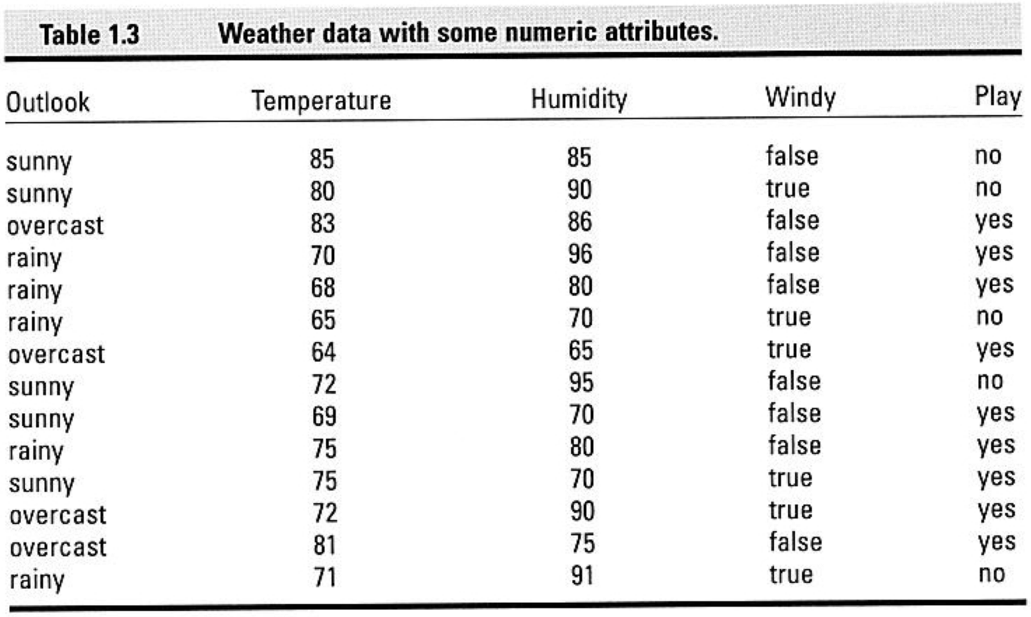


Table 1.3 **Weather data with some numeric attributes.**

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

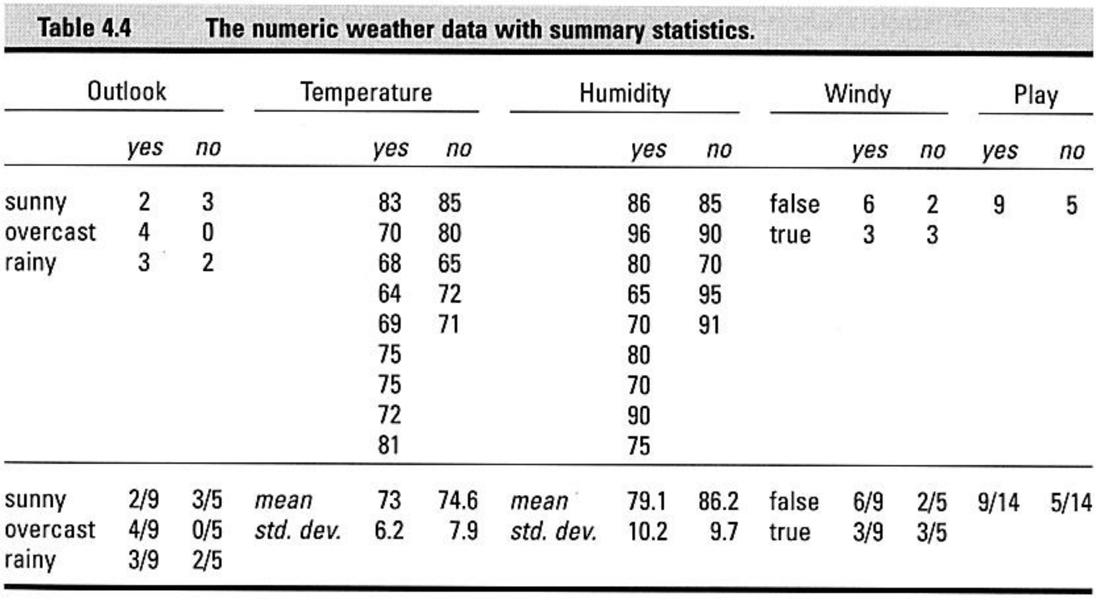


Table 4.4 **The numeric weather data with summary statistics.**

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3		83	85		86	85	false	6	2	9	5
overcast	4	0		70	80		96	90	true	3	3		
rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
sunny	2/9	3/5	*mean*	73	74.6	*mean*	79.1	86.2	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	*std. dev.*	6.2	7.9	*std. dev.*	10.2	9.7	true	3/9	3/5		
rainy	3/9	2/5											

How would Outlook=sunny,Temperature=66,Humidity=90,Windy=true be classified?

NUMERIC AND NOMINAL ATTRIBUTES

Calculation of $P(Play = yes|sunny, 66, 90, true)$

$$P(Outlook = sunny|yes) = 2/9$$

$$P(Temp = 66|yes) = \frac{1}{6.2\sqrt{2\pi}}e^{-\frac{1}{2}(66-73)^2/2.6^2} = 0.034$$

$$P(Humid = 90|yes) = \frac{1}{10.2\sqrt{2\pi}}e^{-\frac{1}{2}(90-79.1)^2/10.2^2} = 0.022$$

$$P(Windy = true|yes) = 3/9$$

$$P(Play = yes) = 9/14)$$

Likelihood of yes:
$2/9 \times 0.034 \times 0.022 \times 3/9 \times 9/14 = 0.000036$

Calculation of $P(Play = no|sunny, 66, 90, true)$
Likelihood of no:
$3/5 \times 0.022 \times 0.38 \times 3/5 \times 5/14 = 0.000108$

Classification is “Play=no”

Note: Multiplying small numbers can lead to floating underflow. Use logs.

NAIVE BAYES CLASSIFIER

- Assume attributes are independent so probabilities can be multiplied [Origin of ‘Naive’]
- Assume each attribute contributes equally to the decision
- What if some of the probabilities are zero?
  - Use the Laplace correction, add 1 to all counts
- Works well with attribute selection
- Good for missing values, don’t include that attribute.
- Fast
- Remember
  - From the data we can get $P(Evidence|Hypothesis)$
  - To make the decision we need $P(Hypothesis|Evidence)$

KERNEL DENSITY

- Naive Bayes works by assuming normal (numeric) and multinomial (symbolic) distributions and estimating the parameters from the data

- What if the data does not fit a normal distribution and we don't know the what kind of distribution it might be?

- Use some curve fitting techniques to fit a curve to the data

- Use [statistical] techniques to fit a probability density function to the data
  - Note: same problem occurs in clustering

- Use this function in Bayes rule

TEXT MINING

1. Document Classification
   - (a) Junk email vs Good
   - (b) Relevant vs Not Relevant documents, web pages
2. Document Clustering
3. Document Summarisation
4. Sentiment Analysis (Twitter)
5. Data Models
   - (a) Link to field of Information Retrieval (IR)
   - (b) Bag of words (Vector Space Model)
   - (c) TF-IDF Term frequency - inverse document frequency

TEXT MINING IN WEKA

- Use Weka “String” attribute for text attributes
- Use StringToWordVector filter
- Might need:
  1. Stemmer
  2. Stop words handler
  3. Tokenizer

MISSING VALUES

- Ignore any record with a missing value?
- Easy for Naive Bayes
  - Don't include missing attribute in probability calculation
- Symbolic Attributes
  - Use 'Missing' as another value
  - Use the most common value [Weka]
  - Use 0 0 0 0 for 1 out of n coding [Neural Networks]
- Numeric Attributes
  - Use 0 [Could be dangerous]
  - Use 0 [Neural Networks, usually OK]
  - Use the average [Weka]
- Use machine learning techniques to predict missing value

COMPARISON CRITERIA FOR CLASSIFIERS

Accuracy Various measures

Speed Computational cost of learning and using

Robustness Deal with noisy data, missing values

Scalability Deal with large amounts of data

Interpretability Level of understanding and insight

	K-NN	J48
Accuracy	Similar	Similar
Speed Learning	Very Good	OK
Speed Using	Bad	Very Good
Robustness	OK	OK
Scalability	Very Bad	OK
Interpretability	Bad	Good