Bayesian Classifier

Prepared by

Dr. Siddique Ibrahim

VIT-AP University

Supervised Classification Technique

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attributes as a function of the values of other attributes.
- Goal: Previously unseen records should be assigned a class as accurately as possible.
  - Satisfy the property of "mutually exclusive and exhaustive"

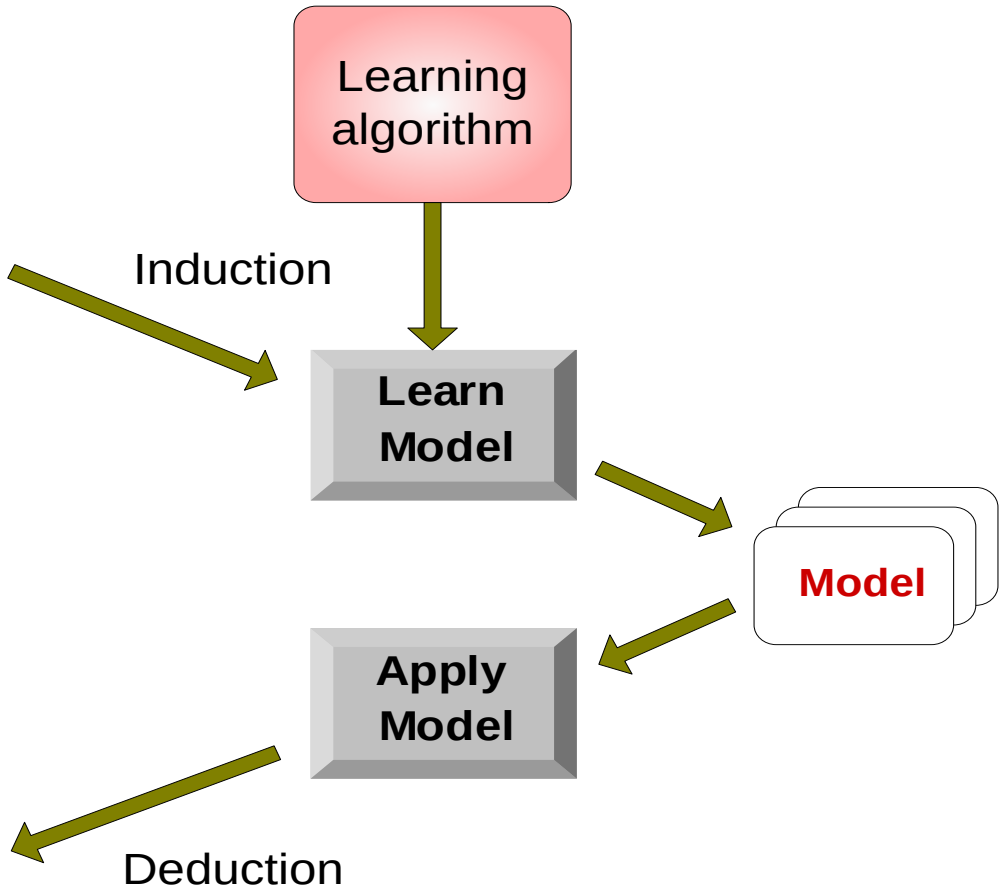
Illustrating Classification Tasks

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Classification Problem

- More precisely, a classification problem can be stated as below:

Definition : **Classification Problem**

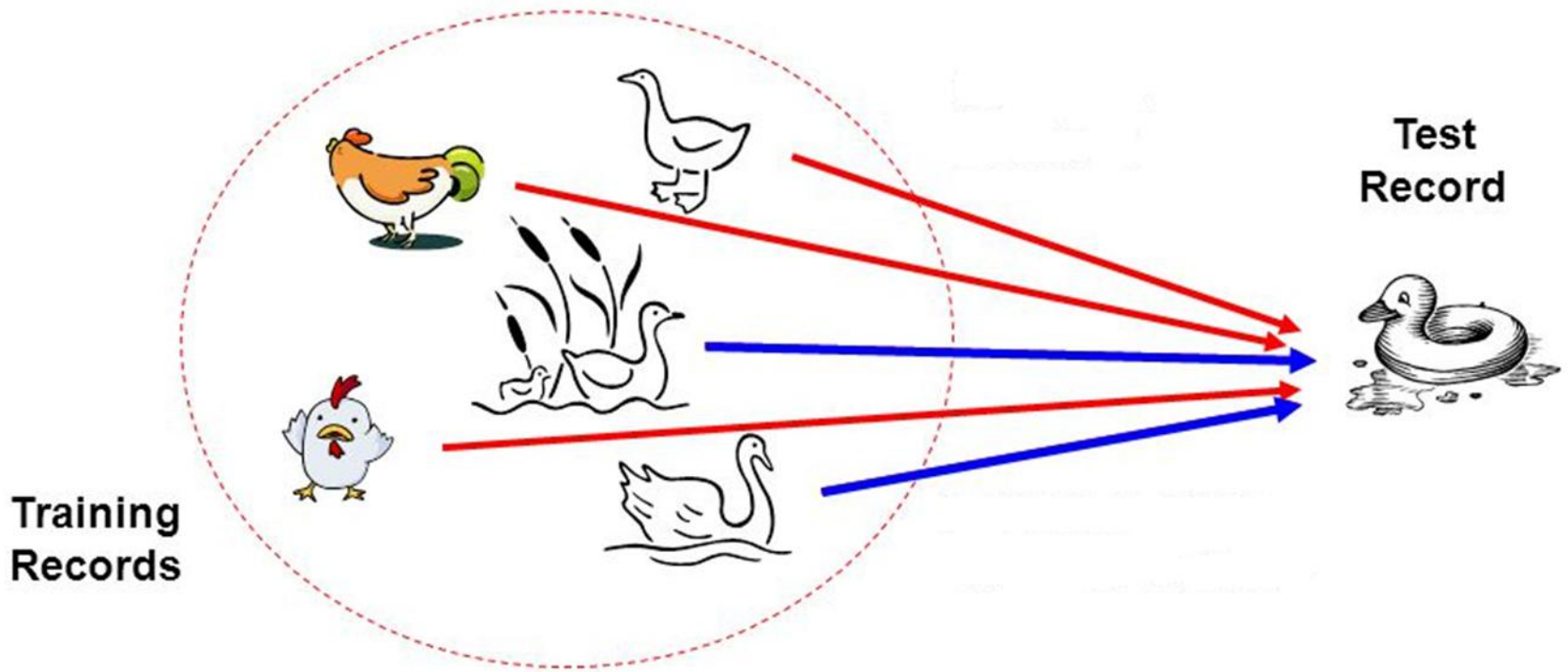
Given a database D = $\{t_1, t_2, \ldots\ldots, t_m\}$ of tuples and a set of classes C = $\{c_1, c_2, \ldots\ldots, c_k\}$, the classification problem is to defie a mapping $\mathrm{f} : \mathrm{D} \rightarrow C$,

Where each t_i is assigned to one class.

Note that tuple $t_i \in D$ is defined by a set of attributes $A = \{A_1, A_2, \ldots\ldots, A_n\}$.

Bayesian Classifier

- Principle
  - If it walks like a duck and quacks like a duck, then it is probably a duck.



Bayesian Classifier

- A statistical classifier
  - Performs probabilistic prediction, i.e., predicts class membership probabilities.
- Foundation
  - Based on Bayes' Theorem.
- Assumptions
  - The classes are mutually exclusive and exhaustive.
  - The attributes are independent given the class.

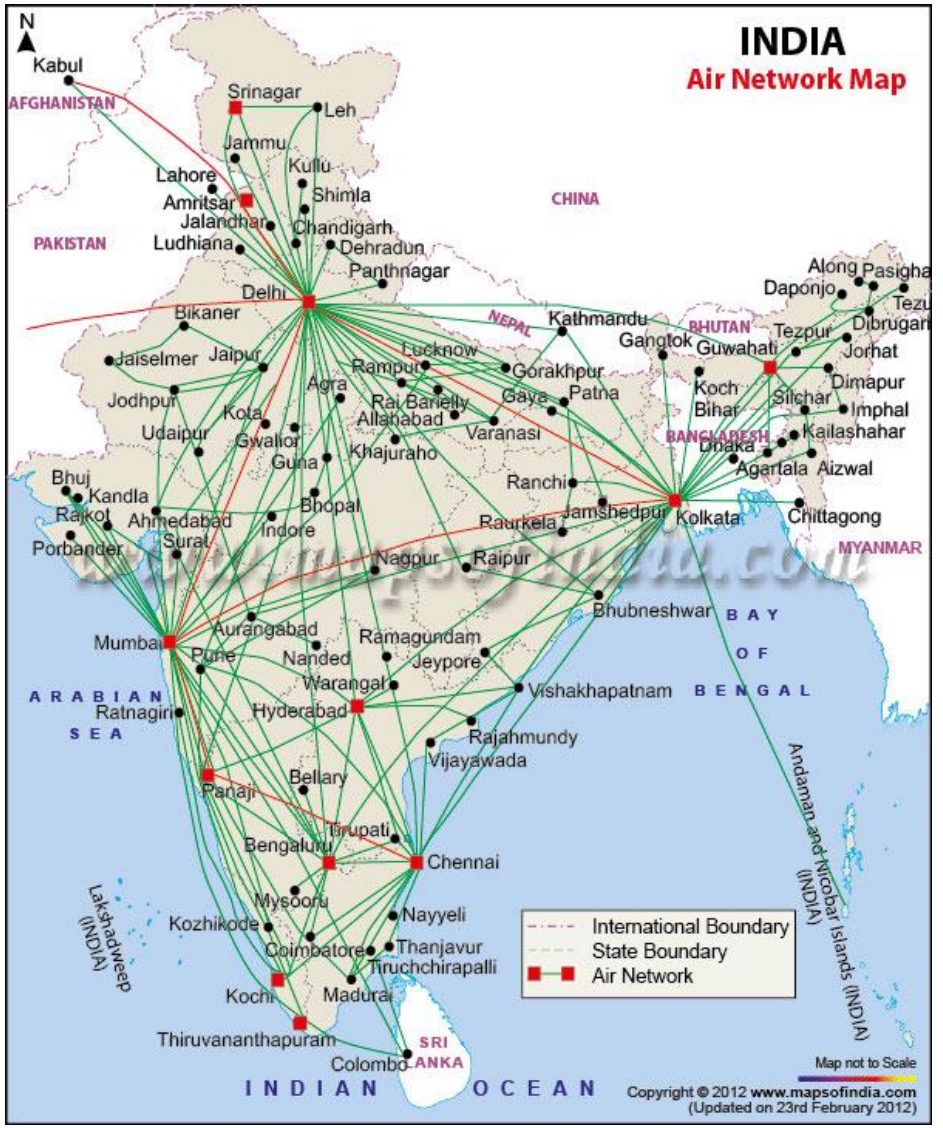
- Called "Naïve" classifier because of these assumptions.
  - Empirically proven to be useful.
  - Scales very well.

Naïve Bayes Classifier Algorithm

- Naïve Bayes algorithm is a supervised learning algorithm, which is based on the Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that includes a high-dimensional training dataset.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

Example: Bayesian Classification

- **Example**: Air Traffic Data
  - Let us consider a set of observations recorded in a database
    - Regarding the arrival of airplanes on the routes from any airport to New Delhi under certain conditions.



Air-Traffic Data

Days	Season	Fog	Rain	Class
Weekday	Spring	None	None	On Time
Weekday	Winter	None	Slight	On Time
Weekday	Winter	None	None	On Time
Holiday	Winter	High	Slight	Late
Saturday	Summer	Normal	None	On Time
Weekday	Autumn	Normal	None	Very Late
Holiday	Summer	High	Slight	On Time
Sunday	Summer	Normal	None	On Time
Weekday	Winter	High	Heavy	Very Late
Weekday	Summer	None	Slight	On Time

Air-Traffic Data

Cond. from previous slide...

Days	Season	Fog	Rain	Class
Saturday	Spring	High	Heavy	Cancelled
Weekday	Summer	High	Slight	On Time
Weekday	Winter	Normal	None	Late
Weekday	Summer	High	None	On Time
Weekday	Winter	Normal	Heavy	Very Late
Saturday	Autumn	High	Slight	On Time
Weekday	Autumn	None	Heavy	On Time
Holiday	Spring	Normal	Slight	On Time
Weekday	Spring	Normal	None	On Time
Weekday	Spring	Normal	Heavy	On Time

Air-Traffic Data

- In this database, there are four attributes

  A = [Day, Season, Fog, Rain]

  with 20 tuples.
- The categories of classes are:

  C= [On Time, Late, Very Late, Cancelled]

- Given this is the knowledge of data and classes, we are to find the most likely classification for any other unseen instance, for example:

Week Day	Winter	High	None	???

- Classification technique eventually to map this tuple into an accurate class

Bayesian Classifier

- In many applications, the relationship between the attributes set and the class variable is non-deterministic.
  - In other words, a test cannot be classified to a class label with certainty.
  - In such a situation, the classification can be achieved probabilistically.

- The Bayesian classifier is an approach for modeling probabilistic relationships between the attribute set and the class variable.
- More precisely, the Bayesian classifier uses Bayes' Theorem of Probability for classification.
- Before going to discuss the Bayesian classifier, we should have a quick look at the Theory of Probability and then Bayes' Theorem.

Bayes' Theorem

- Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on conditional probability.

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

Where,

- P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

Working Steps of Naïve Bayes' Classifier

- Step 1: Convert the given dataset into frequency tables.
- Step 2:Generate Likelihood table by finding the probabilities of given features.
- Step 3:Now, use the Bayes theorem to calculate the posterior probability.

Recollect Prior and Posterior Probabilities

- P(A) and P(B) are called prior probabilities
- P(A|B), P(B|A) are called posterior probabilities

Example: Prior versus Posterior Probabilities

- This table shows that the event *Y* has two outcomes namely *A* and *B*, which is dependent on another event *X* with various outcomes like x_1, x_2 and x_3.

- **Case1:** Suppose, we don't have any information of the event *A*. Then, from the given sample space, we can calculate

$$P(Y = A) = \frac{5}{10} = 0.5$$

- **Case2:** Now, suppose, we want to calculate $P(X = x_2 \mid Y = A)$

$$= \frac{2}{5} = 0.4 .$$

The later is the conditional or posterior probability, where as the former is the prior probability.

X	Y
x_1	A
x_2	A
x_3	B
x_3	A
x_2	B
x_1	A
x_1	B
x_3	B
x_2	B
x_2	A

Naïve Bayesian Classifier

- Suppose, Y is a class variable and X = {$X1$,X_2,……,X_n } is a set of attributes, with an instance of Y.

INPUT (X)	CLASS(Y)
… … …	
… … …	…
x_1 , x_2 , … , x_n	y_i
… … …	…

The classification problem, then can be expressed as the class-conditional probability

$$P\left(Y = y_i | (X_1 = x_1) \text{ AND } (X_2 = x_2) \text{ AND } \ldots\ldots (X_n = x_n)\right)$$

Naïve Bayesian Classifier

- Naïve Bayesian classifier calculates this posterior probability using Bayes' theorem, which is as follows.
- From Bayes' theorem on conditional probability, we have

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

$$= \frac{P(X|Y)\cdot P(Y)}{P(X|Y=y_1)\cdot P(Y=y_1) + \cdots + P(X|Y=y_k)\cdot P(Y=y_k)}$$

where,

$$P(X) = \sum_{i=1}^{k} P(X|Y=y_i)\cdot P(Y=y_i)$$

Note:

- $P(X)$ is called the evidence (also the total probability) and it is a constant.
- The probability $P(Y|X)$ (also called class conditional probability) is therefore proportional to $P(X|Y)\cdot P(Y)$.
- Thus, $P(Y|X)$ can be taken as a measure of Y given that X.

$$P(Y|X) \approx P(X|Y)\cdot P(Y)$$

Naïve Bayesian Classifier

- Suppose, for a given instance of X (say x = ($X_1 = x_1$) and ….. ($X_n = x_n$)).
- There are any two class conditional probabilities namely $P(Y = y_i | X=x)$ and $P(Y = y_j | X=x)$.
- If $P(Y = y_i | X=x) > P(Y = y_j | X=x)$, then we say that y_i is more stronger than y_j for the instance $X = x$.
- The strongest y_i is the classification for the instance $X = x$.

Bayesian Classifier Solved Problems

Naïve Bayesian Classifier

Algorithm: Naïve Bayesian Classification

Input: Given a set of *k* mutually exclusive and exhaustive classes $C = \{c_1, c_2, \ldots\ldots, c_k\}$, which have prior probabilities $P(C_1)$, $P(C_2)$,..... $P(C_k)$.

There are *n*-attribute set $A = \{A_1, A_2, \ldots\ldots, A_n\}$, which for a given instance have values $A_1 = a_1$, $A_2 = a_2$,....., $A_n = a_n$

Step: For each $c_i \in C$, calculate the class condition probabilities, $i = 1,2,\ldots\ldots,k$

$$p_i = P(C_i) \times \prod_{j=1}^{n} P(A_j = a_j | C_i)$$

$$p_x = \max\{p_1, p_2, \ldots\ldots, p_k\}$$

Output: C_x is the classification

Note: $\sum p_i \neq 1$, because they are not probabilities rather a proportion values (to posterior probabilities)

X= (Outlook=”Sunny”, Temperature=”Mild”, Humidty =”Normal”, Wind=”Weak”

Example:

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

Working Steps of Naïve Bayes' Classifier

- Step 1: Calculate the prior probability for given class labels
- Step 2: Find the Likelihood probability with each attribute for each class
- Step 3: Put these values in Bayes Formula and calculate posterior probability.
- Step 4: See which class has a higher probability, given the input belongs to the higher probability class.

Example Cont'd

- Learning Phase(frequency tables)

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(Play=Yes) = 9/14 P(Play=No) = 5/14

Example Cont'd

- Remembering Bayes Rule:

$$p(h|D) = \frac{p(D|h) * p(h)}{p(D)}$$

- We write our f(x) in that form:

$$P(\text{Play Tennis} \mid \text{Attributes}) = \frac{P(\text{Attributes} \mid \text{Play Tennis}) * P(\text{Play Tennis})}{P(\text{Attributes})}$$

Or

$$P(v|a) = \frac{P(a|v) * P(v)}{P(a)}$$

Lets look closely at P(a|v)

$$P(a|v) = P(a0 \ldots a3 \mid v0,1)$$

Or

P(a|v) = P(Outlook, Temperature, Humidity, Wind | Play tennis, Don't Play tennis)

Example Cont'd

- In order to get a table with reliable measurements every combination of each attribute a0…a3 for each hypothesis v0,1 our table would have to be of size 3*3*2*2*2 = 72 and each combination would have to be observed multiple times to ensure its reliability.
- Why, because we are assuming an inter-dependence of the attributes (probably a good assumption).
- The Naïve Bayes classifier is based on simplifying this assumption. That is to say, cool temperature is completely independent of it being sunny and so on.

Example Cont'd

- So :

$$P(a0 \ldots a3 \mid v_j=0,1) = P(a0|v0) * P(a1|v0) * P(an|v0)$$
$$= P(a0|v1) * P(a1|v1) * P(an|v1)$$

or

$$P(a0 \ldots an \mid v_j) = \prod_i P(ai \mid v_j)$$

P(outlook = sunny, temperature = cool, humidity = normal, wind = strong | Play tennis) =

P(outlook = sunny | Play tennis) * P(temperature = cool | Play tennis) *
P(humidity = normal | Play tennis) * P(wind = strong | Play tennis)

Example Cont'd

- Using the table of 14 examples we can calculate our overall probabilities and conditional probabilities.

First, we estimated the probability of playing tennis:

P(Play Tennis = Yes) = 9/14 = .64

P(Play Tennis = No) = 5/14 = .36

- Then we estimate the conditional probabilities of the individual attributes.
- Remember this is the step in which we are assuming that the attributes are independent of each other:
- Outlook:

P(Outlook = Sunny | Play Tennis = Yes) = 2/9 = .22

P(Outlook = Sunny | Play Tennis = No) = 3/5 = .6

P(Outlook = Overcast | Play Tennis = Yes) = 4/9 = .44

P(Outlook = Overcast | Play Tennis = No) = 0/5 = 0

P(Outlook = Rain | Play Tennis = Yes) = 3/9 = .33

P(Outlook = Rain | Play Tennis = No) = 2/5 = .4

Example Cont'd

Temperature

P(Temperature = Hot | Play Tennis = Yes) = 2/9 = .22
P(Temperature = Hot | Play Tennis = No) = 2/5 = .40
P(Temperature = Mild | Play Tennis = Yes) = 4/9 = .44
P(Temperature = Mild | Play Tennis = No) = 2/5 = .40
P(Temperature = Cool | Play Tennis = Yes) = 3/9 = .33
P(Temperature = Cool | Play Tennis = No) = 1/5 = .20

Humidity

P(Humidity = Hi | Play Tennis = Yes) = 3/9 = .33
P(Humidity = Hi | Play Tennis = No) = 4/5 = .80
P(Humidity = Normal | Play Tennis = Yes) = 6/9 = .66
P(Humidity = Normal | Play Tennis = No) = 1/5 = .20

Wind

P(Wind = Weak | Play Tennis = Yes) = 6/9 = .66
P(Wind = Weak | Play Tennis = No) = 2/5 = .40
P(Wind = Strong | Play Tennis = Yes) = 3/9 = .33
P(Wind = Strong | Play Tennis = No) = 3/5 = .60

Example Cont'd

- Test Phase

  Given a new instance,

  x'=(Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong)

- $P(Yes|(sunny, cool, high, strong)) = \frac{P((sunny, cool, high, strong) \mid Yes) * P(Yes)}{P(sunny, cool, high, strong)}$

$$= \frac{P(sunny|Yes)*P(cool|Yes)*P(high|Yes)*P(strong|Yes) * P(yes)}{P((sunny, cool, high, strong) \mid Yes) + P((sunny, cool, high, strong) \mid No)}$$

$$= \frac{(.22 * .33 * .33 * .33) * .64}{(.22 * .33 * .33 * .33)*.64 + (.6 * .2 * .8 * .6)*.36}$$

$$= \frac{.0051}{.0051 + .0207}$$

$$= .1977$$

Example Cont'd

- Similarly,

$$(\text{No}|(\text{sunny, cool, high, strong})) = \frac{P((\text{sunny, cool, high, strong}) \mid \text{No}) * P(\text{No})}{P(\text{sunny, cool, high, strong})}$$

$$= \frac{.0207}{.0051 + .0207}$$

$$= .8023$$

- Apply MAP (Maximum Likelihood) rule

The 20% for playing tennis in the described conditions, and a value of 80% for not playing tennis in these conditions, therefore the prediction is that no tennis will be played if the day is like these conditions.

Car theft Example

Data set

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Attributes are Color , Type , Origin, and the subject, stolen can be either yes or no.

We want to classify a Red, Domestic,and SUV

There is no example of a Red Domestic SUV in our data set

Car theft Example

- Calculate the probabilities:

  P(Red|Yes), P(SUV|Yes), P(Domestic|Yes) ,

  P(Red|No) , P(SUV|No), and P(Domestic|No)

Color	Stolen=*Yes*	Stolen=*No*
Red	3/5	2/5

Type	Stolen=*Yes*	Stolen=*No*
SUV	1/5	2/5

Origin	Stolen=*Yes*	Stolen=*No*
Domestic	2/5	3/5

P(Stolen= Yes) = 5/10 = .5
P(Stolen= No) = 5/10 = .5

- $P(Yes|(Red,Domestic, SUV)) = \frac{P((Red,Domestic, SUV) \mid Yes) * P(Yes)}{P(Red,Domestic, SUV)}$

$$= \frac{P(red|Yes)*P("Domestic"|Yes)*P(SUV|Yes)* P(yes)}{P((Red,Domestic, SUV) \mid Yes) + P((Red,Domestic, SUV) \mid No)}$$

$$= \frac{(.6*.5*.2)*.5}{(.6*.5*.2)*.5+(.4*.6*.6)*.5} = 0.295$$

- $P(No|(Red,Domestic, SUV)) = \frac{P((Red,Domestic, SUV) \mid No) * P(No)}{P(Red,Domestic, SUV)}$

$$= \frac{P(red|No)*P("Domestic"|No)*P(SUV|No)* P(No)}{P((Red,Domestic, SUV) \mid Yes) + P((Red,Domestic, SUV) \mid No)}$$

$$= \frac{(.4*.6*.6))*.5}{(.6*.5*.2)*.5+(.4*.6*.6)*.5} = 0.705$$

The example gets classified as 'NO'

Problem 3- for Naive Bayes Classification

Blood pressure	Weight	Family history	Age	Diabetes
Average	Above average	Yes	50+	1
Low	Average	Yes	0 50	0
High	Above average	No	50+	1
Average	Above average	Yes	50+	1
High	Above average	Yes	50+	0
Average	Above average	Yes	0 50	1
Low	Below average	Yes	0 50	0
High	Above average	No	0 50	0
Low	Below average	No	0 50	0
Average	Above average	Yes	0 50	0
High	Average	No	50+	0
Average	Average	Yes	50+	1
High	Above average	No	50+	1
Average	Average	No	0 50	0
Low	Average	No	50+	0
Average	Above average	Yes	0 50	1
High	Average	Yes	50+	1
Average	Above average	No	0 50	0
High	Above average	No	50+	1
High	Average	No	0 50	0

Classify from the following Data:

Blood Pressure: High

Wight: Above average

Family History: Yes

Age: 50+

Diabetes?

Problem 3 Cont'd

P(Diabetes=*Yes*) = 9/20 =0.45

P(Diabetes=*No*) = 11/20=0.55

"BP"	Diabetes=Yes	Diabetes=*No*
Low	0/9	4/11
Average	5/9	3/11
High	4/9	4/11

"Weight"	Diabetes=Yes	Diabetes=*No*
Below Average	?	?
Average		
Above average	7/9	4/11

"Family history"	Diabetes=Yes	Diabetes=*No*
Yes	6/9	4/11
No		

"Age"	Diabetes=Yes	Diabetes=*No*
0-50		
50+	7/9	3/11

Problem 3 Cont'd

Classify from the following Data:

Blood Pressure: High Wight: Above average Family History: Yes Age: 50+ Diabetes?

- P(Yes|(BP=High, Weight=Above average, Family history=yes,Age=50+)) =

$$\frac{P((BP=High,Wieght=Above\ average,\ Family\ history=yes,Age=50+))\ |\ Yes) * P(Yes)}{P("BP=High,Weight=Above\ average,\ Family\ history=yes,Age=50+")}$$

  - P(Yes|(BP=High, Weight=Above average, Family history=yes,Age=50+))
  - $= \frac{(.44*.78*.67*.78)*.45}{(.44*.78*.67*.78)*.45+(.36*.36*.36*.27)*.55}$
  - $= \frac{.080}{.0869} = .920$

- Similarly,
- P(No|(BP=High,Weight=Above average, Family history=yes,Age=50+)) =

$$\frac{P((BP=High,Weight=Above\ average,\ Family\ history=yes,Age=50+))\ |\ No) * P(No)}{P("BP=High,Wieht=Above\ average,\ Family\ history=yes,Age=50+")}$$

- P(No|(BP=High, Weight=Above average, Family history=yes,Age=50+))
- $= \frac{(.36*.36*.36*.27)*.55}{(.44*.78*.67*.78)*.45+(.36*.36*.36*.27)*.55}$
- $=\frac{.0069}{.0869} = .080$

Problem 4
Air-Traffic Data

Given this is the knowledge of data and classes, we are to find the most likely classification for any other unseen instance, for example:

Week Day	Winter	High	None	???

Days	Season	Fog	Rain	Class
Weekday	Spring	None	None	On Time
Weekday	Winter	None	Slight	On Time
Weekday	Winter	None	None	On Time
Holiday	Winter	High	Slight	Late
Saturday	Summer	Normal	None	On Time
Weekday	Autumn	Normal	None	Very Late
Holiday	Summer	High	Slight	On Time
Sunday	Summer	Normal	None	On Time
Weekday	Winter	High	Heavy	Very Late
Weekday	Summer	None	Slight	On Time
Saturday	Spring	High	Heavy	Cancelled
Weekday	Summer	High	Slight	On Time
Weekday	Winter	Normal	None	Late
Weekday	Summer	High	None	On Time
Weekday	Winter	Normal	Heavy	Very Late
Saturday	Autumn	High	Slight	On Time
Weekday	Autumn	None	Heavy	On Time
Holiday	Spring	Normal	Slight	On Time
Weekday	Spring	Normal	None	On Time
Weekday	Spring	Normal	Heavy	On Time

Problem 4 Cont'd

		Class			
Attribute		On Time	Late	Very Late	Cancelled
Day	Weekday	9/14 = 0.64	½ = 0.5	3/3 = 1	0/1 = 0
	Saturday	2/14 = 0.14	½ = 0.5	0/3 = 0	1/1 = 1
	Sunday	1/14 = 0.07	0/2 = 0	0/3 = 0	0/1 = 0
	Holiday	2/14 = 0.14	0/2 = 0	0/3 = 0	0/1 = 0
Season	Spring	4/14 = 0.29	0/2 = 0	0/3 = 0	0/1 = 0
	Summer	6/14 = 0.43	0/2 = 0	0/3 = 0	0/1 = 0
	Autumn	2/14 = 0.14	0/2 = 0	1/3= 0.33	0/1 = 0
	Winter	2/14 = 0.14	2/2 = 1	2/3 = 0.67	0/1 = 0

Problem 4 Cont'd

Attribute		Class			
		On Time	Late	Very Late	Cancelled
Fog	None	5/14 = 0.36	0/2 = 0	0/3 = 0	0/1 = 0
	High	4/14 = 0.29	1/2 = 0.5	1/3 = 0.33	1/1 = 1
	Normal	5/14 = 0.36	1/2 = 0.5	2/3 = 0.67	0/1 = 0
Rain	None	5/14 = 0.36	1/2 = 0.5	1/3 = 0.33	0/1 = 0
	Slight	8/14 = 0.57	0/2 = 0	0/3 = 0	0/1 = 0
	Heavy	1/14 = 0.07	1/2 = 0.5	2/3 = 0.67	1/1 = 1
Prior Probability		14/20 = 0.70	2/20 = 0.10	3/20 = 0.15	1/20 = 0.05

Problem 4 Cont'd

Instance:

Week Day	Winter	High	Heavy	???

Case1: Class = On Time : $0.70 \times 0.64 \times 0.14 \times 0.29 \times 0.07 = 0.0013$

Case2: Class = Late : $0.10 \times 0.50 \times 1.0 \times 0.50 \times 0.50 = 0.0125$

Case3: Class = Very Late : $0.15 \times 1.0 \times 0.67 \times 0.33 \times 0.67 = 0.0222$

Case4: Class = Cancelled : $0.05 \times 0.0 \times 0.0 \times 1.0 \times 1.0 = 0.0000$

Case3 is the strongest; Hence correct classification is **Very Late**

Naïve Bayesian Classifier

Pros and Cons

- The Naïve Bayes' approach is a very popular one, which often works well.
- However, it has a number of potential problems
  - It relies on all attributes being categorical.
  - If the data is less, then it estimates poorly.

Estimating the Posterior Probabilities for Continuous Attributes

Approach to overcome the limitations in Naïve Bayesian Classification

- Estimating the posterior probabilities for continuous attributes
  - In real-life situations, all attributes are not necessarily categorical, In fact, there is a mix of both categorical and continuous attributes.
  - In the following, we discuss the schemes to deal with continuous attributes in Bayesian classifier.

1. We can discretize each continuous attribute and then replace the continuous values with their corresponding discrete intervals.
2. We can assume a certain form of a probability distribution for the continuous variable and estimate the parameters of the distribution using the training data. A Gaussian distribution is usually chosen to represent the posterior probabilities for continuous attributes. A general form of Gaussian distribution will look like

$$P(x: \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where, μ and σ^2 denote mean and variance, respectively.

Estimating the Posterior Probabilities for Continuous Attributes

For each class C_i, the posterior probabilities for attribute A_j (it is the numeric attribute) can be calculated following the Gaussian normal distribution as follows.

$$P(A_j = a_j | C_i) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(a_j - \mu ij)^2}{2{\sigma_{ij}}^2}}$$

Here, the parameter μ_{ij} can be calculated based on the sample mean of attribute value of A_j for the training records that belong to the class C_i.

Similarly, ${\sigma_{ij}}^2$ can be estimated from the calculation of variance of such training records.

Estimating the Posterior Probabilities for Continuous Attributes

Example Problem: classify whether a given person is a male or a female based on the measured features. The features include height, weight, and foot size, assuming there exists normalized Gaussian Distribution in the populations.

Person	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

$$P(Male) = 4/8 = 0.5$$

$$P(Female) = 4/8 = 0.5$$

$$Mean\ (Height) = \frac{(6+5.92+5.58+5.92)}{4} = 5.855$$

$$Variance\ (Height) = \frac{\sum(x_i - x)^2}{n-1}$$

$$= \frac{(6-5.855)^2+(5.92-5.855)^2+(5.58-5.855)^2+(5.92-5.855)^2}{4-1}$$

$$= 0.035055$$

Person	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033×10^{-2}	176.25	1.2292×10^{2}	11.25	9.1667×10^{-1}
female	5.4175	9.7225×10^{-2}	132.5	5.5833×10^{2}	7.5	1.6667

Estimating the Posterior Probabilities for Continuous Attributes

- To which class a person with the given inputs will be classified?

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

$$\text{posterior (male)} = \frac{P(\text{male})\,p(\text{height} \mid \text{male})\,p(\text{weight} \mid \text{male})\,p(\text{foot size} \mid \text{male})}{evidence} \quad = 6.1984e^{-09}$$

$$\text{posterior (female)} = \frac{P(\text{female})\,p(\text{height} \mid \text{female})\,p(\text{weight} \mid \text{female})\,p(\text{foot size} \mid \text{female})}{evidence} \quad = 5.3778e^{-04}$$

$$P(x: \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

P(male) = 0.5

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

where
$\mu = 5.855$
and
$\sigma^2 = 3.5033e-02$

P(female) = 0.5
p(height | female) = 2.2346e-1
p(weight | female) = 1.6789e-2
p(foot size | female) = 2.8669e-1

Similarly, $p(\text{weight} \mid \text{male}) = 5.9881e^{-06}$
$p(\text{foot size} \mid \text{male}) = 1.3112e^{-3}$

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

M-estimate of Conditional Probability

- we estimated conditional probabilities Pr(A | B) by n_c /n where n_c is the number of times $A \wedge B$ happened and n is the number of times B happened in the training data.
- This can cause trouble if $n_c = 0$

To avoid this, we fix the following numbers p and m beforehand:

- A nonzero prior estimate p for Pr(A | B), and
- A number m that says how confident we are of our prior estimate p, as measured in the number of samples

M-estimate of Conditional Probability

- The M-estimation is to deal with the potential problem of Naïve Bayesian Classifier when training data size is too poor.
  - If the posterior probability for one of the attributes is zero, then the overall class-conditional probability for the class vanishes.
  - In other words, if training data do not cover many of the attribute values, then we may not be able to classify some of the test records.
- This problem can be addressed by using the M-estimate approach.

M-estimate of Conditional Probability

- M-estimate approach can be stated as follows

$$P(A_j = a_j | C_i) = \frac{n_{c_i} + mp}{n + m}$$

where, n = total number of instances from class C_i

n_{c_i} = number of training examples from class C_i that take the value $A_j = a_j$

m = it is a parameter known as the equivalent sample size, and

p = is a user-specified parameter.

Note:

If $n = 0$, that is, if there is no training set available, then $P(a_i|C_i) = p$,

so, this is a different value, in absence of sample value.