Probabilistic Learning

Module 2

Naïve Bayes Classification

Introduction

- For some problems like weather estimates, they are based on probabilistic methods or those concerned with describing uncertainty. They use data on past events to extrapolate future events.
- In the case of weather, the chance of rain describes the proportion of prior days to similar measurable atmospheric conditions in which precipitation occurred. A 70 percent chance of rain implies that in 7 out of the 10 past cases with similar conditions, precipitation occurred somewhere in the area.
- When a meteorologist provides a weather forecast, precipitation is typically described with terms such as "70 percent chance of rain." Such forecasts are known as probability of precipitation reports.

Understanding Naïve Bayes

- The technique descended from the work of the 18th century mathematician Thomas Bayes, who developed foundational principles to describe the probability of events, and how probabilities should be revised using additional information. These principles formed the foundation for what are now known as Bayesian methods.

- A probability is a number between 0 and 1 (that is, between 0 percent and 100 percent), which captures the chance that an event will occur in the light of the available evidence. The lower the probability, the less likely the event is to occur. A probability of 0 indicates that the event will definitely not occur, while a probability of 1 indicates that the event will occur with 100 percent certainty.

Understanding Naïve Bayes

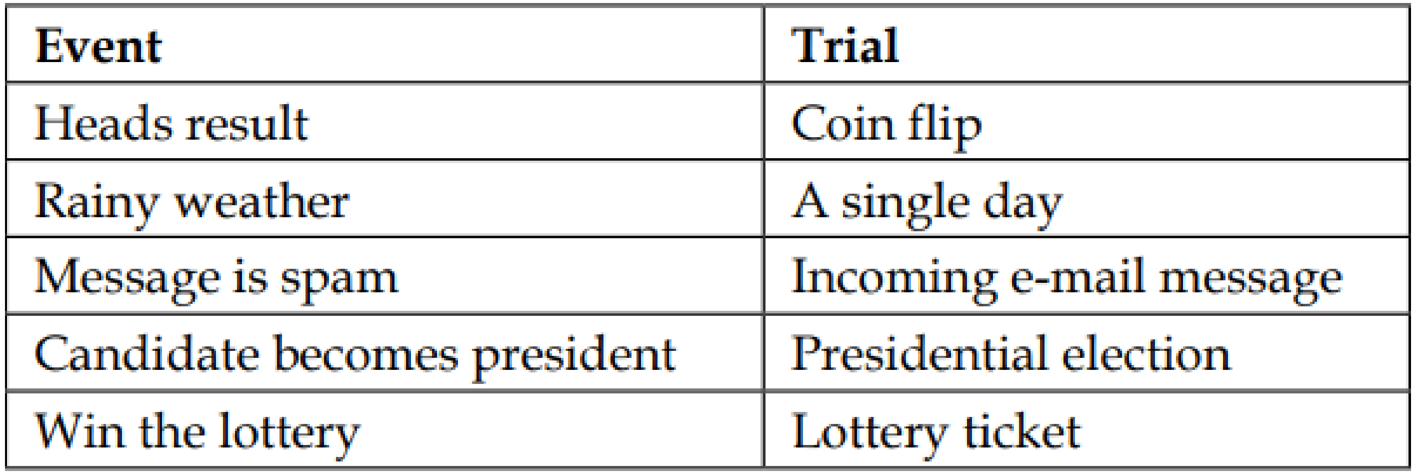
- Classifiers based on Bayesian methods utilize training data to calculate an observed probability of each outcome based on the evidence provided by feature values. When the classifier is later applied to unlabeled data, it uses the observed probabilities to predict the most likely class for the new features.

- In fact, Bayesian classifiers have been used for:
  - Text classification, such as junk e-mail (spam) filtering
  - Intrusion or anomaly detection in computer networks
  - Diagnosing medical conditions given a set of observed symptom

Basic Concepts of Bayesian Method

- Bayesian probability theory is rooted in the idea that the estimated likelihood of an event, or a potential outcome, should be based on the evidence at hand across multiple trials, or opportunities for the event to occur.

The following table illustrates events and trials for several real-world outcomes:



Event	Trial
Heads result	Coin flip
Rainy weather	A single day
Message is spam	Incoming e-mail message
Candidate becomes president	Presidential election
Win the lottery	Lottery ticket

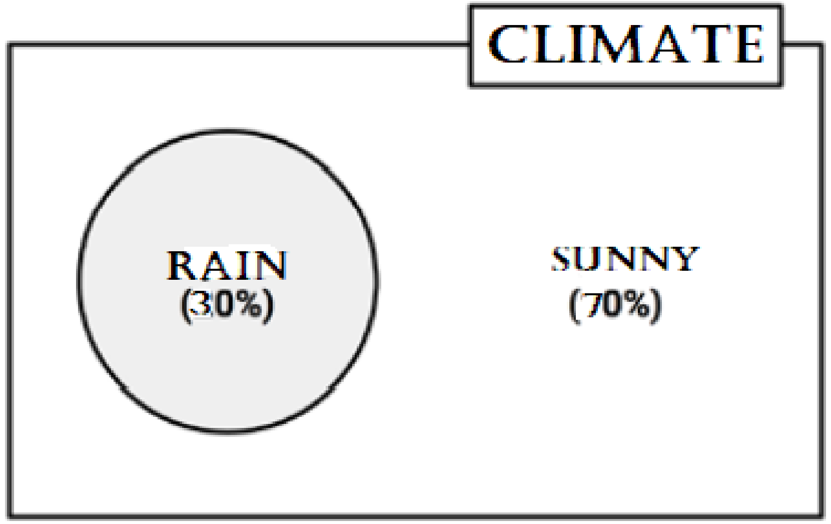
Bayesian methods provide insights into how the probability of these events can be estimated from the observed data.

Understanding probability

- **The probability of an event is estimated from the observed data by dividing the number of trials in which the event occurred by the total number of trials.**
- For example, if it rained 3 out of 10 days with similar conditions as today, the probability of rain today can be estimated as 3 / 10 = 0.30 or 30%.
- To denote these probabilities, we use notation in the form **P(A)**, which signifies the probability of event A. For example, P(rain) = 0.30
- The probability of all the possible outcomes of a trial must always sum to 1. Thus, if the trial has two outcomes that cannot occur simultaneously, such as rainy versus sunny, then knowing the probability of either outcome reveals the probability of the other.
- For example, given the value P(rain) = 0.30, we can calculate P(sunny) = 1 – 0.30 = 0.70. **This concludes that rain and sunny are mutually exclusive and exhaustive events**, which implies that they cannot occur at the same time and are the only possible outcomes.

Understanding probability

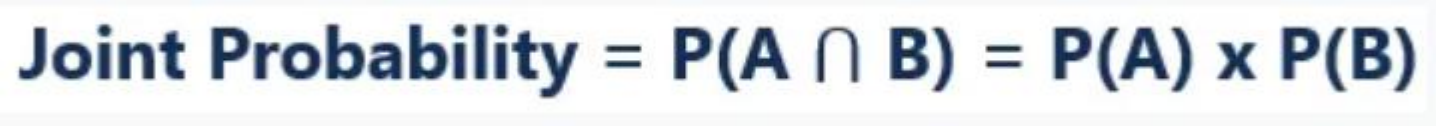
- Because an event cannot simultaneously happen and not happen, an event is always mutually exclusive and exhaustive **with its complement**.
- The complement of event A is typically denoted A^c or A'. P(¬A) can also used to denote the probability of event A not occurring. As P(¬rain) = 0.70. This notation is equivalent to P(A^c).



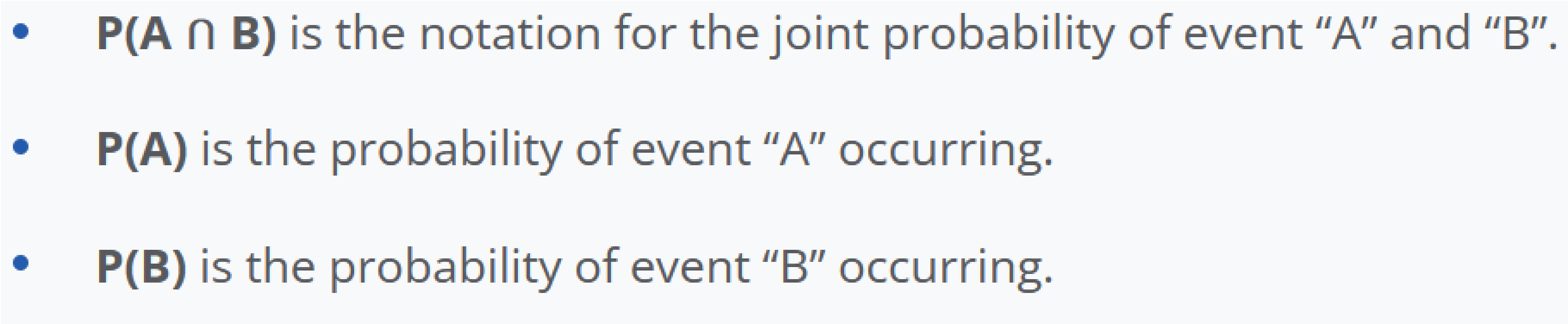
In the following diagram, the rectangle represents the possible outcomes for climate. The circle represents the 30 percent probability that the climate is rain. The remaining 70 percent represents the complement P(¬rain) = sunny.

Joint Probability

- Joint probability is a statistical measure that calculates the likelihood of two events occurring together and at the same point in time. Joint probability is the probability of event Y occurring at the same time that event X occurs.



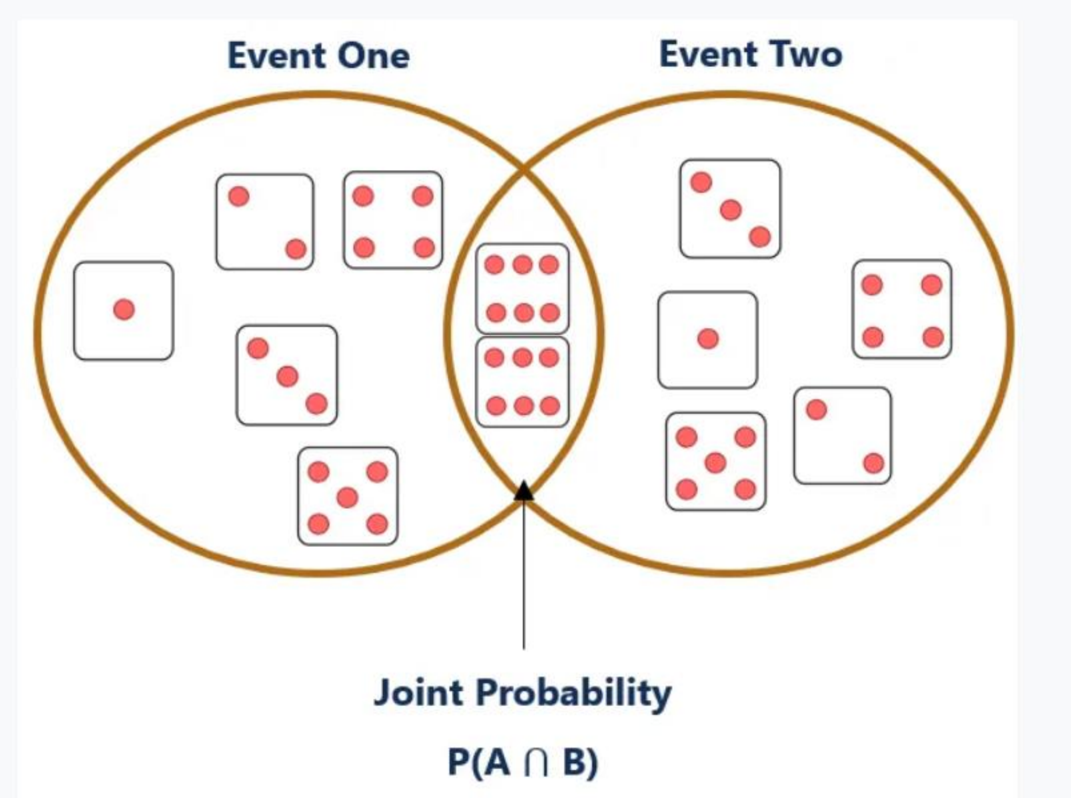
Joint Probability = $P(A \cap B) = P(A) \times P(B)$



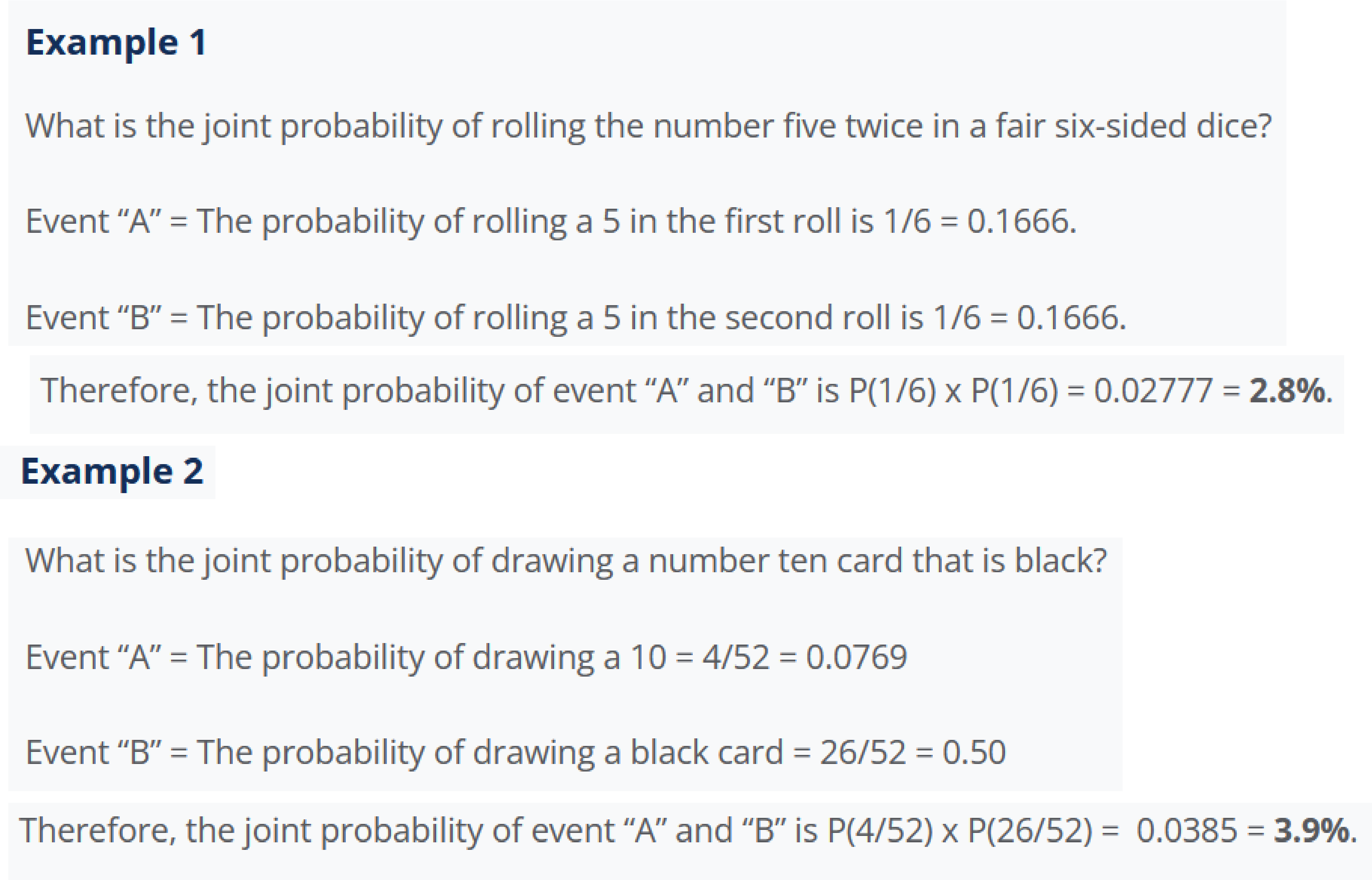
- **$P(A \cap B)$** is the notation for the joint probability of event “A” and “B”.
- **$P(A)$** is the probability of event “A” occurring.
- **$P(B)$** is the probability of event “B” occurring.

Joint Probability….

- Visual Representation of Joint Probability



Joint Probability….



Example 1

What is the joint probability of rolling the number five twice in a fair six-sided dice?

Event “A” = The probability of rolling a 5 in the first roll is 1/6 = 0.1666.

Event “B” = The probability of rolling a 5 in the second roll is 1/6 = 0.1666.

Therefore, the joint probability of event “A” and “B” is P(1/6) x P(1/6) = 0.02777 = **2.8%**.

Example 2

What is the joint probability of drawing a number ten card that is black?

Event “A” = The probability of drawing a 10 = 4/52 = 0.0769

Event “B” = The probability of drawing a black card = 26/52 = 0.50

Therefore, the joint probability of event “A” and “B” is P(4/52) x P(26/52) = 0.0385 = **3.9%**.

Conditional Probability with Bayes Theorm

- The relationships between dependent events can be described using Bayes' theorem, as shown in the following formula. This formulation provides a way of thinking about how to revise an estimate of the probability of one event in light of the evidence provided by another event:

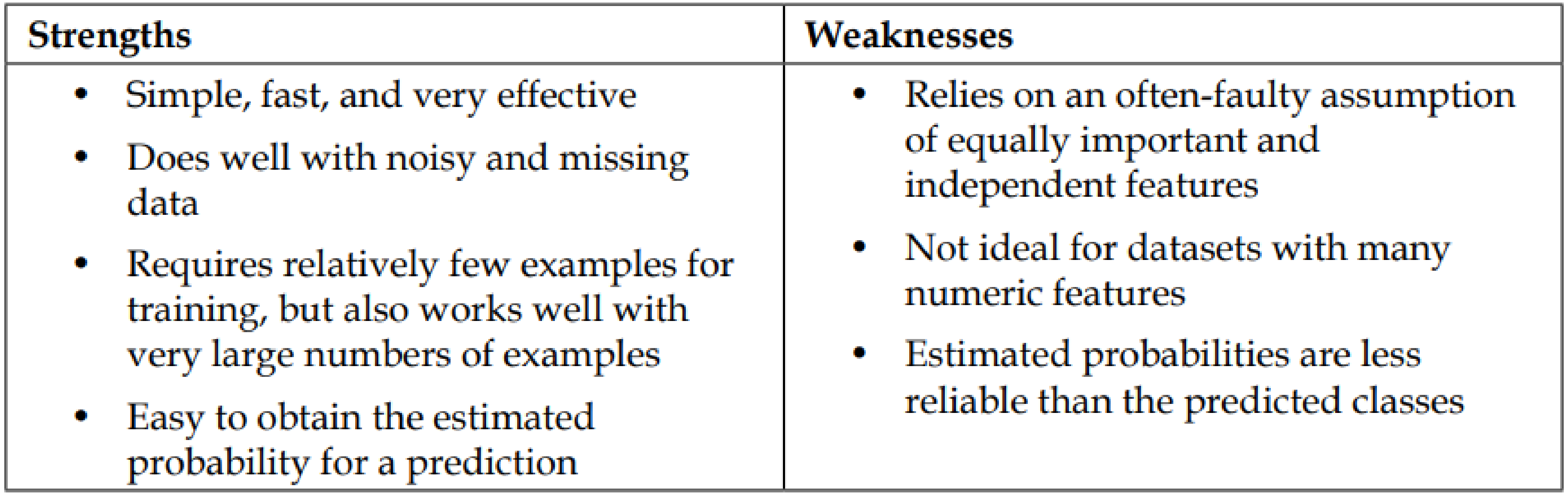
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

The notation P(A|B) is read as the probability of event A, given that event B occurred. This is known as conditional probability.

Conditional probability is defined as the likelihood of an event or outcome occurring, based on the occurrence of a previous event or outcome.

The Naive Bayes algorithm

- The Naive Bayes algorithm describes a simple method to apply Bayes' theorem to classification problems. It is the most common one and is particularly true for text classification. The strengths and weaknesses of this algorithm are as follows:



Strengths	Weaknesses
• Simple, fast, and very effective • Does well with noisy and missing data • Requires relatively few examples for training, but also works well with very large numbers of examples • Easy to obtain the estimated probability for a prediction	• Relies on an often-faulty assumption of equally important and independent features • Not ideal for datasets with many numeric features • Estimated probabilities are less reliable than the predicted classes

The Naive Bayes algorithm..

- In particular, Naive Bayes assumes that all of the features in the dataset are equally important and independent. These assumptions are rarely true in most real-world applications.

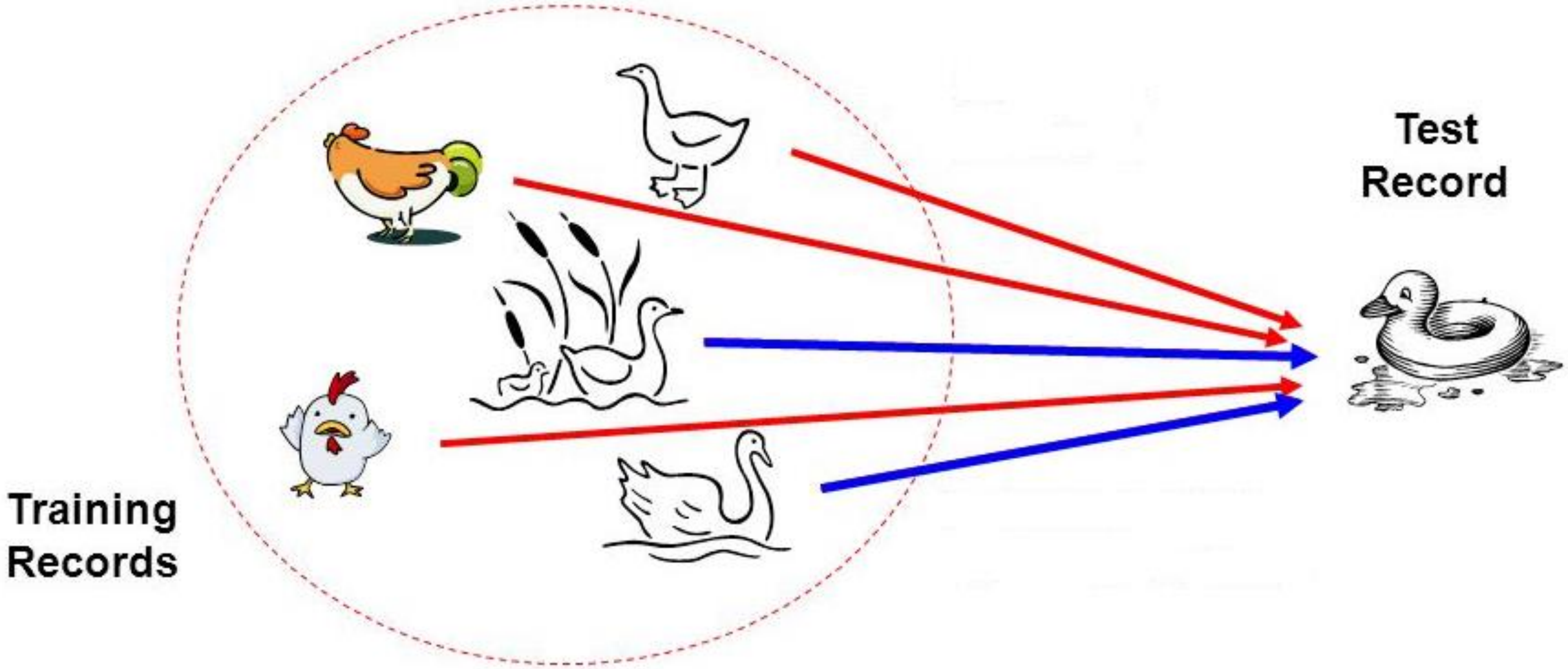
- However, in most cases when these assumptions are violated, Naive Bayes still performs fairly well. This is true even in extreme circumstances where strong dependencies are found among the features. Due to the algorithm's versatility and accuracy across many types of conditions, Naive Bayes is often a strong first candidate for classification learning tasks

Bayesian Classifier

Bayesian Classifier

Principle

If it walks like a duck, 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
  1. The classes are mutually exclusive and exhaustive.
  2. The attributes are independent given the class.
- Called "Naïve" classifier because of these assumptions.
  - Empirically proven to be useful.
  - Scales very well.

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 $P(X = x_2 | Y = A) = \frac{2}{5} = 0.4$.

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* = $\{X_1, X_2, \ldots\ldots, X_n\}$ is a set of attributes, with instance of *Y*.

INPUT (X)	CLASS(Y)
… … …	
… … …	…
$x_1, x_2, \ldots, x_n$	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 calculate 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)*$\cdot$ $P(Y)$.

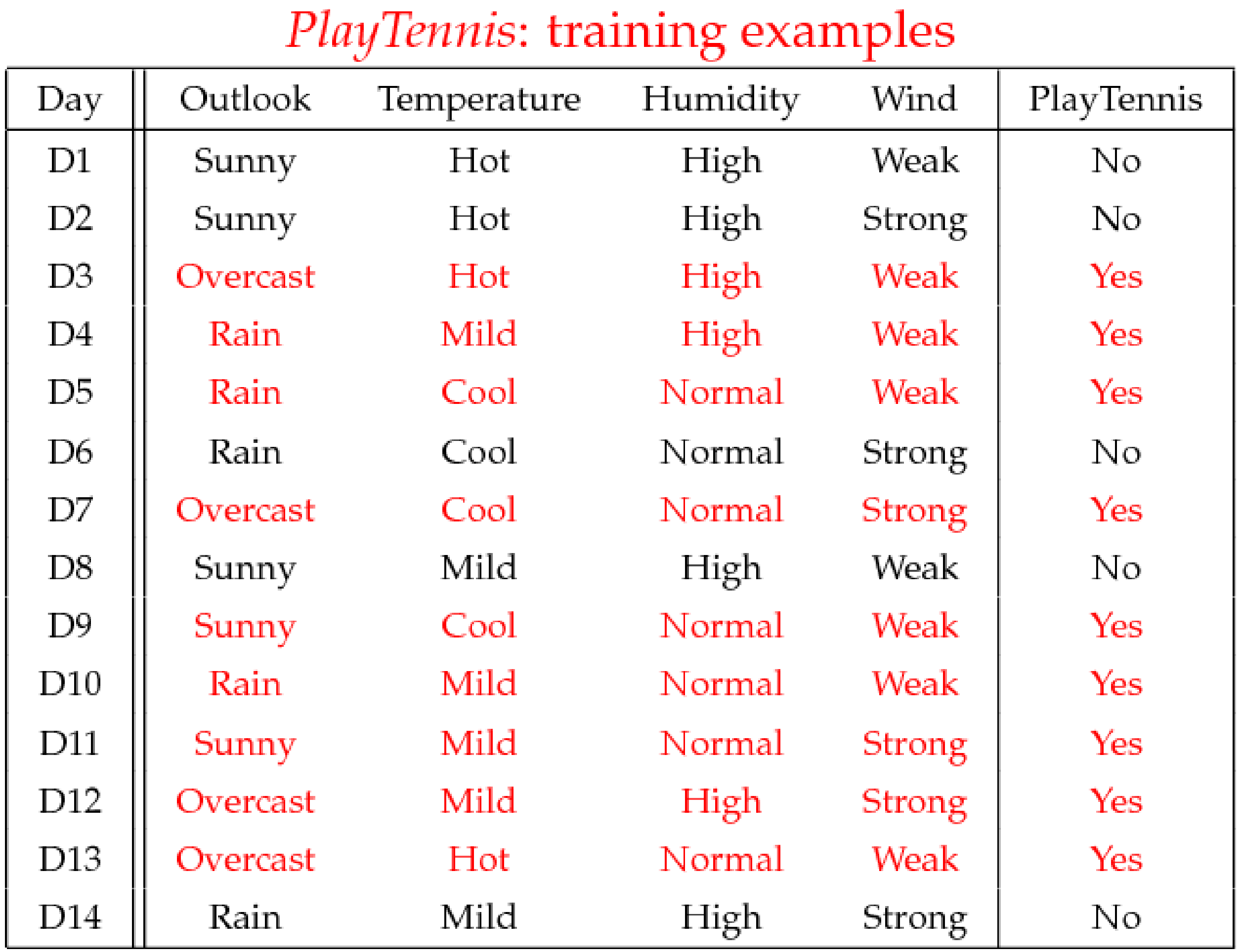
- Thus, *P(Y|X)* can be taken as a measure of *Y* given that *X*.

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

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*.

- Example: Play Tennis



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

- Learning Phase

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

$P(\text{Play}=\mathit{Yes}) = 9/14$ $P(\text{Play}=\mathit{No}) = 5/14$

- Test Phase
  - Given a new instance,

    $\mathbf{x}'$=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)

  - Look up tables

$P(\text{Outlook}=\mathit{Sunny}|\text{Play}=\mathit{Yes}) = 2/9$

$P(\text{Temperature}=\mathit{Cool}|\text{Play}=\mathit{Yes}) = 3/9$

$P(\text{Huminity}=\mathit{High}|\text{Play}=\mathit{Yes}) = 3/9$

$P(\text{Wind}=\mathit{Strong}|\text{Play}=\mathit{Yes}) = 3/9$

$P(\text{Play}=\mathit{Yes}) = 9/14$

$P(\text{Outlook}=\mathit{Sunny}|\text{Play}=\mathit{No}) = 3/5$

$P(\text{Temperature}=\mathit{Cool}|\text{Play}==\mathit{No}) = 1/5$

$P(\text{Huminity}=\mathit{High}|\text{Play}=\mathit{No}) = 4/5$

$P(\text{Wind}=\mathit{Strong}|\text{Play}=\mathit{No}) = 3/5$

$P(\text{Play}=\mathit{No}) = 5/14$

  - MAP rule

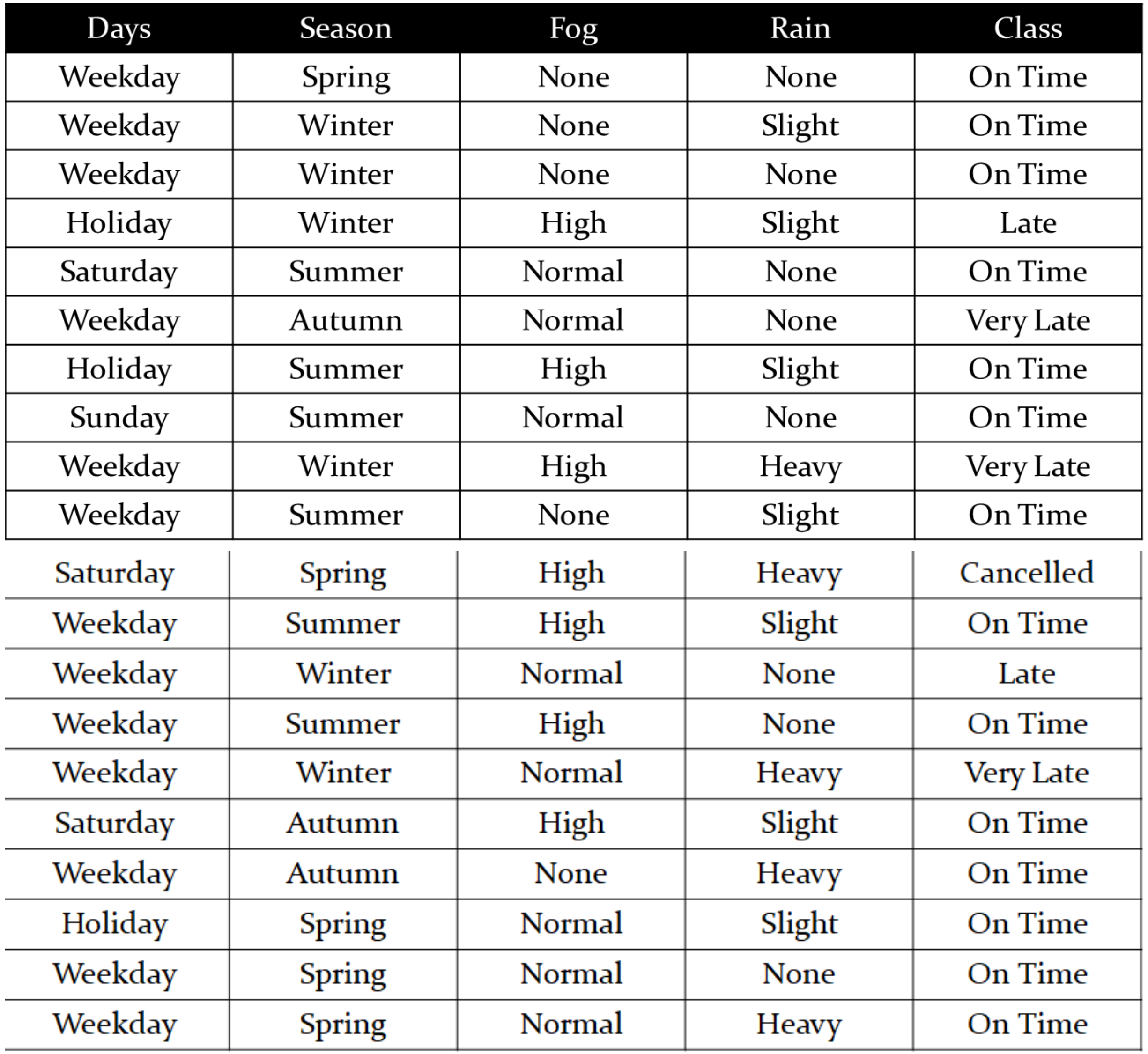
$P(\mathit{Yes}|\mathbf{x}')$: $[P(\mathit{Sunny}|\text{Yes})P(\mathit{Cool}|\text{Yes})P(\mathit{High}|\text{Yes})P(\mathit{Strong}|\text{Yes})]P(\text{Play}=\mathit{Yes}) = 0.0053$

$P(\mathit{No}|\mathbf{x}')$: $[P(\mathit{Sunny}|\text{No})\, P(\mathit{Cool}|\text{No})P(\mathit{High}|\mathit{No})P(\mathit{Strong}|\mathit{No})]P(\text{Play}=\mathit{No}) = 0.0206$

Given the fact $P(\mathit{Yes}|\mathbf{x}') < P(\mathit{No}|\mathbf{x}')$, we label $\mathbf{x}'$ to be "*No*".

- Naïve Bayes based on the independence assumption
  - Training is very easy and fast;
  - Test is straightforward; just looking up tables or calculating conditional probabilities with normal distributions
- A popular generative model
  - Performance competitive to most of state-of-the-art classifiers
  - Many successful applications, e.g., spam mail filtering
  - Apart from classification, naïve Bayes can do more…

Example 2



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

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 most likely classification for any other unseen instance, for example:

Week Day	**Winter**	**High**	**None**	**???**

- Classification technique maps this tuple into an accurate class.

Naïve Bayesian Classifier

- **Example:** With reference to the Air Traffic Dataset mentioned earlier, let us tabulate all the posterior and prior probabilities as shown below.

		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

Naïve Bayesian Classifier

		Class			
Attribute		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

Naïve Bayesian Classifier

Instance:

Week Day	Winter	High	Heavy	???

Case1: Class = On Time : 0.70 × 0.64 × 0.14 × 0.29 × 0.07 = 0.0013

Case2: Class = Late : 0.10 × 0.50 × 1.0 × 0.50 × 0.50 = 0.0125

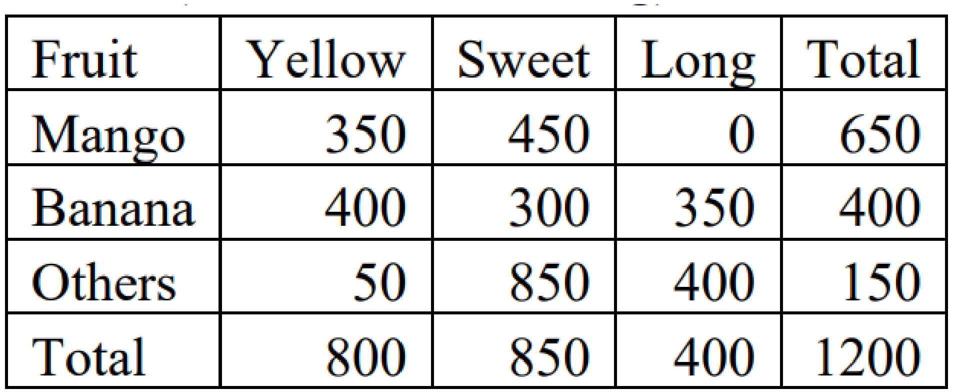
Case3: Class = Very Late : 0.15 × 1.0 × 0.67 × 0.33 × 0.67 = 0.0222

Case4: Class = Cancelled : 0.05 × 0.0 × 0.0 × 1.0 × 1.0 = 0.0000

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

Example for Dataset with numerical Features

Consider the given dataset. Apply Naïve Bayes algorithm and predict that if a fruit has following properties, then which fruit it is.
Fruit = (Yellow, Sweet, Long)



Fruit	Yellow	Sweet	Long	Total
Mango	350	450	0	650
Banana	400	300	350	400
Others	50	850	400	150
Total	800	850	400	1200

- Learning Phase

$P(\mathit{Yellow}|\text{Mango}) = 350/800 = 7/16$

$P(\mathit{Sweet}|\text{Mango}) = 450/850 = 9/17$

$P(\mathit{Long}|\text{Mango}) = 0/400 = 0$

$P(\mathit{Yellow}|\text{Banana}) = 400/800 = 1/2$

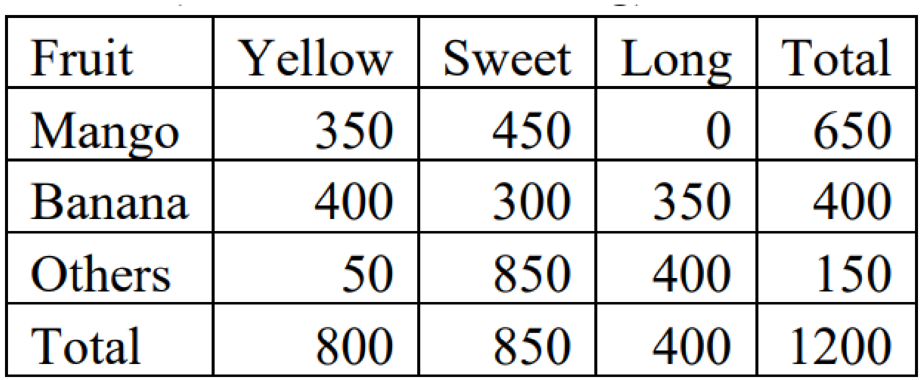
$P(\mathit{Sweet}|\text{Banana}) = 300/850 = 6/17$

$P(\mathit{Long}|\text{Banana}) = 350/400 = 7/8$

$P(\mathit{Yellow}|\text{Others}) = 50/800 = 1/16$

$P(\mathit{Sweet}|\text{Others}) = 850/850 = 1$

$P(\mathit{Long}|\text{Others}) = 400/400 = 1$



Fruit	Yellow	Sweet	Long	Total
Mango	350	450	0	650
Banana	400	300	350	400
Others	50	850	400	150
Total	800	850	400	1200

Prior Probability

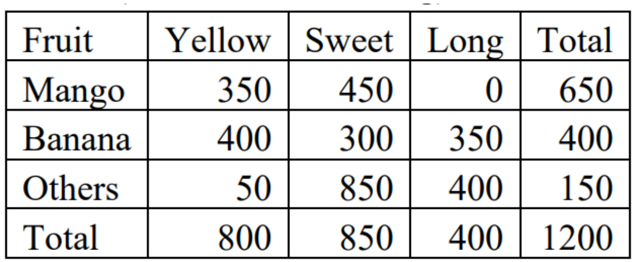
$P(\mathit{Mango}) = 650/1200 = 13/24$

$P(\text{Banana}) = 400/1200 = 1/3$

$P(\text{Others}) = 150/1200 = 3/24$

- Class Determination
  - Given a new instance,

    Fruit = (Yellow, Sweet, Long)



Fruit	Yellow	Sweet	Long	Total
Mango	350	450	0	650
Banana	400	300	350	400
Others	50	850	400	150
Total	800	850	400	1200

- Label Mapping

CASE 1: Mango

P(*Mango***|** Fruit = (Yellow, Sweet, Long)**):**

= [P(*Yellow*|Mango). P(*Sweet*|Mango). P(*Long*|Mango)]. P(Fruit=Mango)

= [(7/16). (9/17). 0] x (13/24)

= 0

CASE 2: Banana

P(*Banana***|** Fruit = (Yellow, Sweet, Long)**):**

= [P(*Yellow*|Banana). P(*Sweet*|Banana). P(*Long*|Banana)]. P(Fruit=Banana)

= [(1/2). (6/17). (7/8)] x (1/3)

= [0.5 x 0.353 x 0.875] x 0.333 = 0.514

CASE 3: Others

P(*Others*| Fruit = (Yellow, Sweet, Long)**):**

= [P(*Yellow*|Others). P(*Sweet*|Others). P(*Long*|Others)]. P(Fruit=Others)

= [(1/16). (1). (1)] x (3/24)

= 0.0625 x 0.125

= 0.0078

Case2 is the strongest; Hence the class label of Fruit= (Yellow, Sweet, Long) is **Banana**

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(C₁), P(C₂),….. P(Cₖ)*.

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 ∈ C, calculate the class condition probabilities, *i* = 1,2,…..,*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: $\sum \boldsymbol{p_i} \neq \mathbf{1}$, because they are not probabilities rather proportion values (to posterior probabilities)

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.

Naïve Bayesian Classifier

Approach to overcome the limitations in Naïve Bayesian Classification

- Estimating the posterior probabilities for continuous attributes
  - In real life situation, all attributes are not necessarily be 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 attributes and then replace the continuous values with its corresponding discrete intervals.

2. We can assume a certain form of 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.

Naïve Bayesian Classifier

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

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

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.

Naïve Bayesian Classifier

Laplace-estimater of Conditional Probability

- The Laplace-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 attribute 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 Laplace-estimate approach.

Naïve Bayesian Classifier

Laplace-estimater

- The Laplace estimator essentially adds a small number to each of the counts in the frequency table, which ensures that each feature has a nonzero probability of occurring with each class. Typically, the Laplace estimator is set to 1, which ensures that each class-feature combination is found in the data at least once.

Normal case

		Class			
Attribute		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

Applying Laplace Estimater

		Class			
Attribute		On Time	Late	Very Late	Cancelled
Fog	None	6/15 = 0.4	1/3 = 0.33	1/4 = 0.25	1/2 = 0.5
	High	5/15 = 0.33	2/3 = 0.667	2/4 = 0.5	2/2 = 1
	Normal	6/15 = 0.4	2/3 = 0.667	3/4 = 0.75	1/2 = 0.5

Naïve Bayesian Classifier

Using numeric features with Naive Bayes

Because Naive Bayes uses frequency tables to learn the data, each feature must be categorical in order to create the combinations of class and feature values comprising of the matrix. Since numeric features do not have categories of values, the preceding algorithm does not work directly with numeric data. There are, however, ways that this can be addressed.

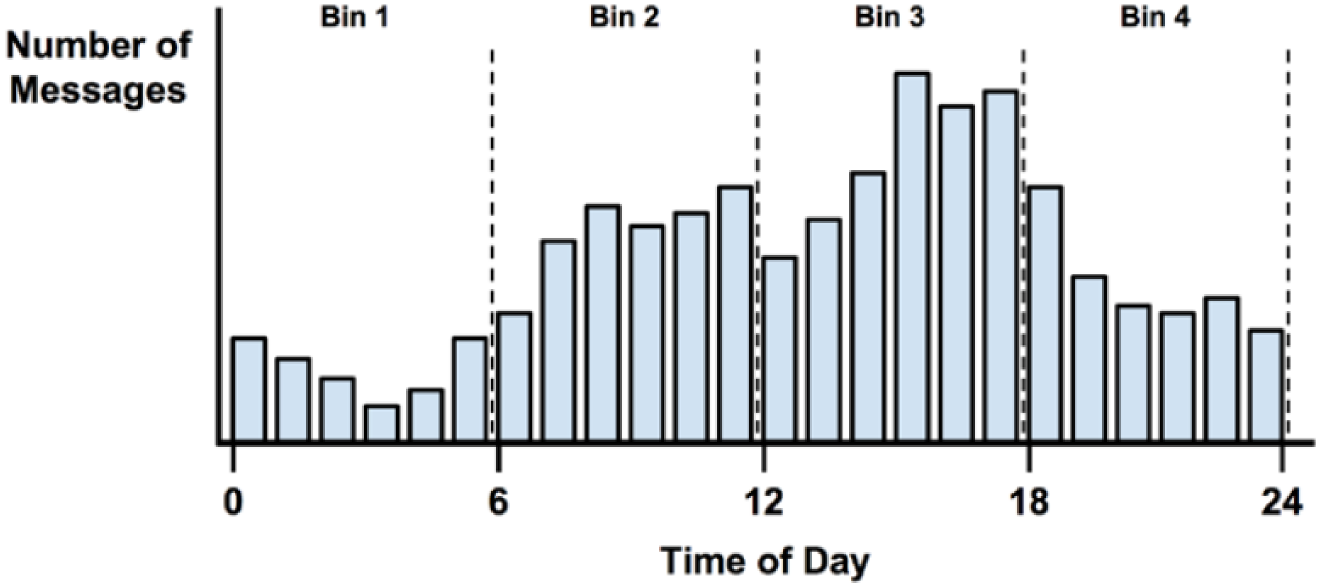
One easy and effective solution is to discretize numeric features, which simply means that the numbers are put into categories known as bins. For this reason, discretization is also sometimes called binning. This method is ideal when there are large amounts of training data, a common condition while working with Naive Bayes.

Using numeric features with Naive Bayes

There are several different ways to discretize a numeric feature. Perhaps the most common is to explore the data for natural categories or cut points in the distribution of data. For example, suppose that you added a feature to the spam dataset that recorded the time of night or day the e-mail was sent, from 0 to 24 hours past midnight.

Depicted using a histogram, the time data might look something like the following diagram. In the early hours of the morning, the message frequency is low. The activity picks up during business hours and tapers off in the evening. This seems to create four natural bins of activity, as partitioned by the dashed lines indicating places where the numeric data are divided into levels of a new nominal feature, which could then be used with Naive Bayes

Using numeric features with Naive Bayes



The choice of four bins was somewhat arbitrary based on the natural distribution of data and a hunch about how the proportion of spam might change throughout the day.