Analytics: The science that analyze crude data to extract useful knowledge (patterns) from them.

This process can also include:

- data collection,
- organization,
- pre-processing,
- transformation,
- modeling and
- interpretation.

• **Descriptive Analytics:** Summarize or condense data to extract patterns.

• **Predictive Analytics:** Extract models from data to be used for future predictions.

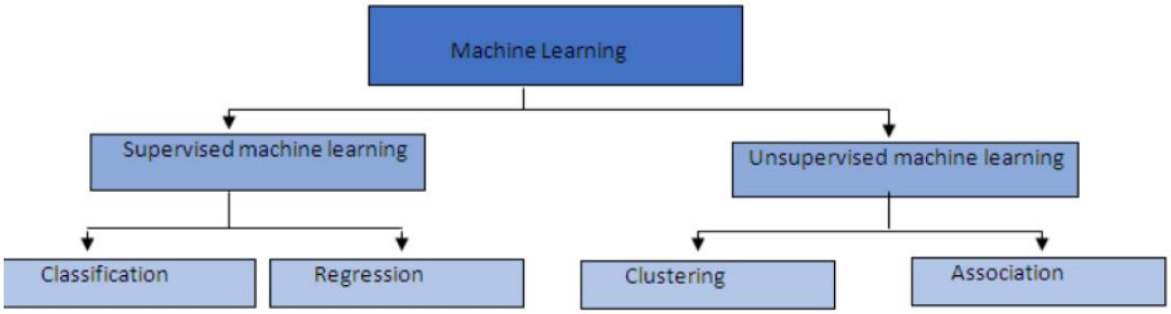
Simple Random Sampling

Researchers adopt a variety of sampling strategies.

The most straightforward is simple random sampling.

Such sampling ***requires every member of the population to have an equal chance of being selected into the sample***.

In addition, ***the selection of one member must be independent of the selection of every other member***.



Regression: the output variable takes continuous values.

Classification: the output variable takes class labels.

Regression involves estimating or predicting a response.

Classification is identifying group membership.

Regression and classification are both related to prediction, where regression predicts a value from a continuous set, whereas classification predicts the 'belonging' to the class.

For example, the price of a house depending on the 'size' (in some unit) and say 'location' of the house, can be some 'numerical value' (which can be continuous): this relates to regression.

Similarly, the prediction of price can be in words, viz., 'very costly', 'costly', 'affordable', 'cheap', and 'very cheap': this relates to classification.

Each class may correspond to some range of values.

Regression: given a set of data, find the best relationship that represents the set of data.

Classification: given a known relationship, identify the class that the data belongs to.

We can see that regression and classification start from opposing ends: to find a pattern or to find the pattern that it belongs to.

Modeling refers to the development of mathematical expressions that describe in some sense the behavior of a random variable of interest.

This variable may be the price of wheat in the world market, the number of deaths from lung cancer, the rate of growth of a particular type of tumor, or the tensile strength of metal wire.

In all cases, this variable is called the **dependent variable** and denoted with Y.

A subscript on Y identifies the particular unit from which the observation was taken, the time at which the price was recorded, the county in which the deaths were recorded, the experimental unit on which the tumor growth was recorded, and so forth.

Most commonly the modeling is aimed at describing how the **mean** of the dependent variable $\varepsilon(Y)$ changes with changing conditions; the variance of the dependent variable is assumed to be unaffected by the changing conditions.

Other variables which are thought to provide information on the behavior of the dependent variable are incorporated into the model as **predictor or explanatory variables**.

These variables are called the **independent variables** and are denoted by X with subscripts as needed to identify different independent variables.

Additional subscripts denote the observational unit from which the data were taken.

The Xs are assumed to be known constants.

In addition to the Xs, all models involve unknown constants, called ***parameters, which control the behavior of the model.***

These **parameters are denoted by Greek letters** and are to be estimated from the data.

The mathematical complexity of the model and the degree to which it is a realistic model depend on how much is known about the process being studied and on the purpose of the modeling exercise.

In preliminary studies of a process or ***in cases where prediction is the primary objective, the models usually fall into the class of models that are linear in the parameters***.

That is, **the parameters enter the model as simple coefficients on the independent variables or functions of the independent variables**.

Such models are referred to loosely as **Linear Models**.

The more realistic models, on the other hand, are often **nonlinear in the parameters**.

Most growth models, for example, are nonlinear models.

Most growth models, for example, are nonlinear models.

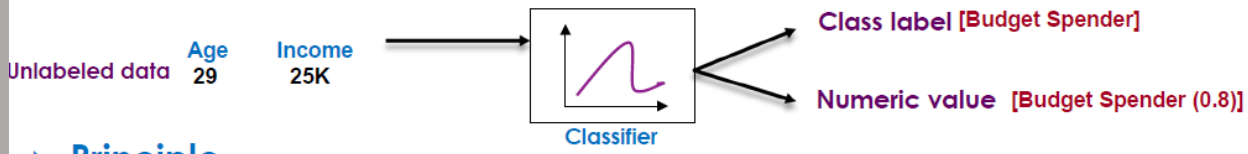
Nonlinear models fall into two categories: **intrinsically linear models**, which can be linearized by an appropriate transformation on the dependent variable, and those that cannot be so transformed.

4.1.1 Definition

- Classification is also called **Supervised Learning**
- **Supervision**
  - → The training data (observations, measurements, etc) are used to learn a classifier
  - → The training data are **labeled** data
  - → New data (**unlabeled**) are classified Using the training data

Training data

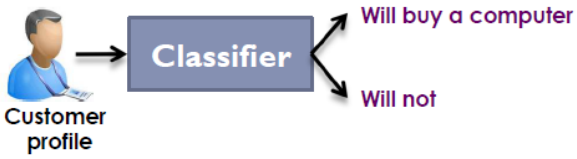
Age	Income	Class label
27	28K	Budget-Spenders
35	36K	Big-Spenders
65	45K	Budget-Spenders



- **Principle**
  - → Construct models (functions) based on some training examples
  - → Describe and distinguish classes or concepts for future prediction
  - → Predict some unknown class labels

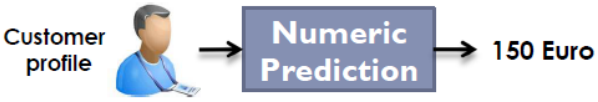
4.1.2 Classification vs. Prediction

- **Classification**
  - → Predicts categorical class labels (discrete or nominal)
  - → Use labels of the training data to classify new data
- **Example**



- A model or classifier is contsructed to predict **categorical labels** such as "safe" or "risky" for a loan application data.

- **Prediction**
  - → Models continuous-valued functions, i.e., predicts unknown or missing values
- **Example**
  - → A marketing manager would like to predict how much a given costumer will spend during a sale



  - → Unlike classification, it provides ordered values
  - → **Regression** analysis is used for prediction
  - → Prediction is a short name for **numeric prediction**

4.1.3 Classification Steps (1/2)

There are two main steps in classification

- **Step1: Model Construction (learning step, or training step)**
  - → Construct a classification model based on **training data**
  - → **Training data**
    - A set of tuples
    - Each tuple is assumed to belong to a predefined class
    - Labeled data (ground truth)
  - → **How a classification model looks like?**

    A classification model can be represented by one of the following forms:
    - Classification rules
    - Decision trees
    - Mathematical formulae

4.1.3 Classification Steps (2/2)

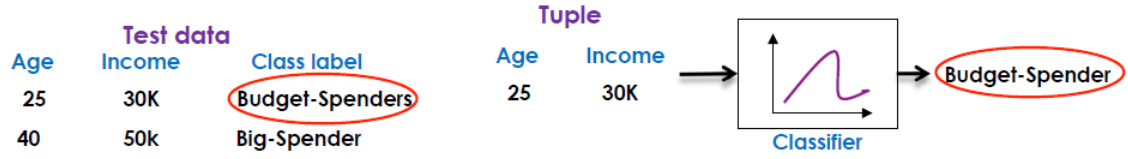
- **Step2: Model Usage**

  Before using the model, we first need to test its accuracy

  → **Measuring model accuracy**

  - To measure the accuracy of a model we need **test data**
  - Test data is similar in its structure to training data (labeled data)
  - **How to test?**

    The known label of test sample is compared with the classified result from the model



  - **Accuracy rate** is the percentage of test set samples that are correctly classified by the model
  - **Important**: test data should be independent of training set, otherwise over-fitting will occur

  → **Using the model:** If the accuracy is acceptable, use the model to **classify data** tuples whose class labels are not known

Classification steps

1. **Training** phase: a model is constructed from the training instances.
   - → classification algorithm finds relationships between predictors and targets
   - → relationships are summarised in a *model*
2. **Testing** phase: test the model on a test sample whose class labels are known but not used for training the model
3. **Usage** phase: use the model for classification on new data whose class labels are unknown

Instance-based Classification

Main idea:

Similar instances have similar classification

- no clear separation between the three phases of classification
- also called *lazy* classification, as opposed to *eager* classification

Eager vs Lazy Classification

Eager

- Model is computed before classification
- Model is independent of the test instance
- Test instance is not included in the training data
- Avoids too much work at classification time
- Model is not accurate for each instance

Lazy

- Model is computed during classification
- Model is dependent on the test instance
- Test instance is included in the training data
- High accuracy for models at each instance level

k-Nearest Neighbor (kNN)

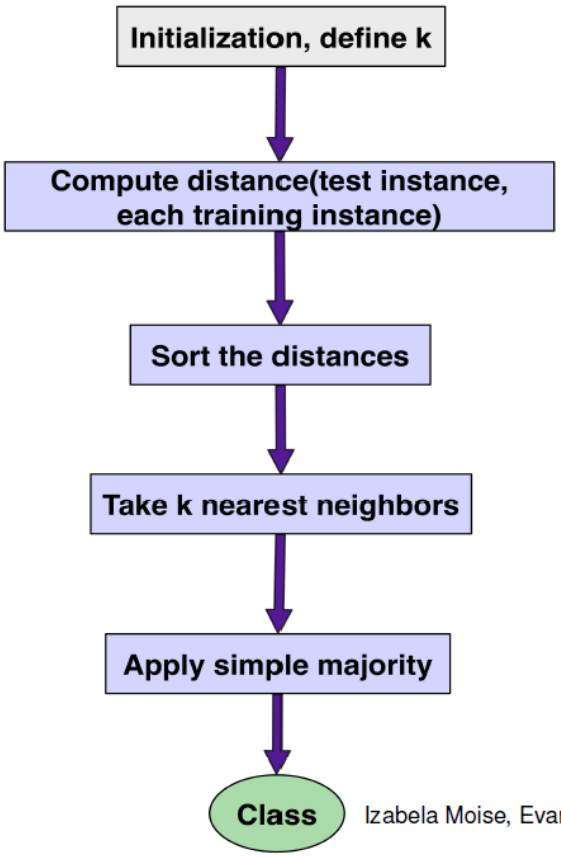
Learning by analogy:

Tell me who your friends are and I'll tell you who you are

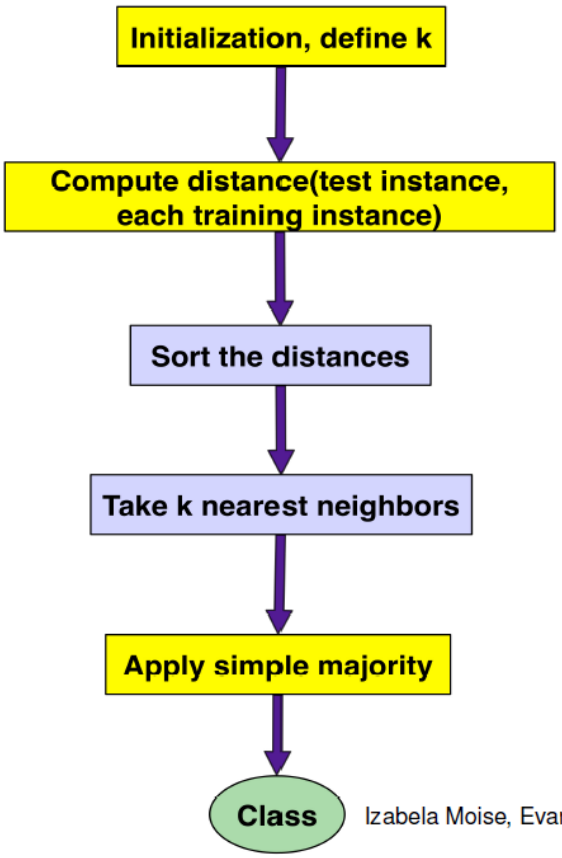
$\rightarrow$ an instance is assigned to the **most common** class among the instances **similar** to it

1. how to measure similarity between instances
2. how to choose the most common class

How does it work?



How does it work?

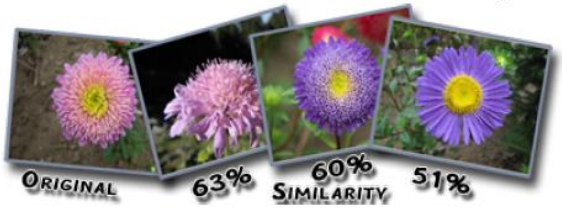


ESS

Izabela Moise, Evangelos Pournaras, Dirk Helbing

Comparing Objects

→ *Problem :* measure similarity between instances



vs. text similarity

- different types of data: numbers colours, geolocation, booleans etc.

✓ *Solution :* convert all features of the instances into numerical values

- represent instances as vectors of features in an n-dimensional space

An Example:

John:
Age=35
Income=95K
No. of credit cards=3

Rachel:
Age=41
Income=215K
No. of credit cards=2

- Closeness" is defined in terms of the *Euclidean* distance between two examples.
  - The Euclidean distance between $X=(x_1, x_2, x_3,...x_n)$ and $Y=(y_1,y_2, y_3,...y_n)$ is defined as:

$$D(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Distance (John, Rachel)=sqrt $[(35-41)^2+(95K-215K)^2+(3-2)^2]$

Distance Metrics

1. Euclidean Distance

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

2. Manhattan Distance

$$D = \sum_{i=1}^{n} |x_i - y_i|$$

3. Minkowski Distance

$$D = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

Choosing k

- Classification is sensitive to the correct selection of k
- if k is too small ⇒ overfitting
  → algorithm performs too good on the training set, compared to its true performance on unseen test data

small k?

larger k?

Choosing k

- Classification is sensitive to the correct selection of k
- if k is too small $\Rightarrow$ overfitting
  - $\rightarrow$ algorithm performs too good on the training set, compared to its true performance on unseen test data

small k? $\rightarrow$ less stable, influenced by noise

larger k? $\rightarrow$ less precise, higher bias

Choosing k

- Classification is sensitive to the correct selection of k
- if k is too small $\Rightarrow$ overfitting
  - $\rightarrow$ algorithm performs too good on the training set, compared to its true performance on unseen test data

small k? $\rightarrow$ less stable, influenced by noise

larger k? $\rightarrow$ less precise, higher bias

k= $\sqrt[2]{n}$

Pros and Cons

Pros:

- ✓ simple to implement and use
- ✓ robust to noisy data by averaging k-nearest neighbours
- ✓ kNN classification is based solely on local information
- ✓ the decision boundaries can be of arbitrary shapes

Supervised Learning

Asks the machine to learn from our data when we specify a target variable.

This reduces the machine's task to only divining some pattern from the input data to get the target variable.

We address two cases of the target variable.
The first case occurs when the target variable can take only nominal values: true or false; reptile, fish, mammal, amphibian, plant, fungi.

The second case of classification occurs when the target variable can take an infinite number of numeric values, such as 0.100, 42.001, 1000.743, ... This case is called **Regression**.

Classifying with distance measurements - k-Nearest Neighbors

Pros: High accuracy, insensitive to outliers, no assumptions about data
Cons: Computationally expensive, requires a lot of memory
Works with: Numeric values, nominal values.

k-NN Working Principles:

- We have an existing set of example data, our training set.
- We have labels for all of this data - We know what class each piece of the data should fall into.
- When we're given a new piece of data without a label, we compare that new piece of data to the existing data, every piece of existing data.
- We then take the most similar pieces of data (the nearest neighbors) and look at their labels.
- We look at the top k most similar pieces of data from our known dataset; this is where the *k* comes from. (*k* is an integer and it's usually less than 20.)
- Lastly, we take a majority vote from the k most similar pieces of data, and the majority is the new class we assign to the data we were asked to classify.

General approach to kNN

1. Collect: Any method.
2. Prepare: Numeric values are needed for a distance calculation. A structured data format is best.
3. Analyze: Any method.
4. Train: Does not apply to the kNN algorithm.
5. Test: Calculate the error rate.
6. Use: This application needs to get some input data and output structured numeric values. Next, the application runs the kNN algorithm on this input data and determines which class the input data should belong to. The application then takes some action on the calculated class.

The data mining tasks can be broadly classified in two categories: descriptive and predictive.

Descriptive mining tasks characterize the general properties of the data in the database.

Predictive mining tasks perform inference on the current data in order to make predictions.

According to different goals, the mining task can be mainly divided into four types:

- Class/Concept Description,
- Association Analysis,
- Classification or Prediction and
- Clustering Analysis.

Feature Selection

Many irrelevant attributes may be present in data to be mined.

So they need to be removed.

Also many mining algorithms don't perform well with large amounts of features or attributes. Therefore **Feature Selection Techniques needs to be applied before any kind of mining algorithm is applied**.

The main objectives of feature selection are to **avoid overfitting** and **improve model performance** and to provide faster and more cost-effective models.

Overfitting: In statistics and machine learning, one of the most common tasks is to fit a "model" to a set of training data, so as to be able to make reliable predictions on general untrained data.

In overfitting, **a statistical model describes random error or noise instead of the underlying relationship**.

Overfitting occurs when *a model is excessively complex*, such as **having too many parameters relative to the number of observations**.

A model that has been overfit ***has poor predictive performance***, **as it overreacts to minor fluctuations in the training data**.

The possibility of overfitting exists because the criterion used for training the model is not the same as the criterion used to judge the efficacy of a model.

In particular, a model is typically trained by maximizing its performance on some set of training data.

However, its efficacy is determined not by its performance on the training data but by its ability to perform well on unseen data.

Overfitting occurs when a model begins to "memorize" training data rather than "learning" to generalize from trend.

An extreme example: If the number of parameters is the same as or greater than the number of observations, a simple model or learning process can perfectly predict the training data simply by memorizing the training data in its entirety, but such a model will typically fail drastically when making predictions about new or unseen data, since the simple model has not learned to generalize at all.

The **potential for overfitting depends on**:

- Number of parameters and data
- The conformability of the model structure with the data shape, and
- The magnitude of model error compared to the expected level of noise or error in the data.

In order **to avoid overfitting**, it is necessary to use additional techniques like:

- Cross-Validation,
- Regularization,
- Early Stopping,
- Pruning,
- Bayesian priors on parameters or model comparison,

that can indicate when further training is not resulting in better generalization.

A **good analogy for the overfitting problem**:

- Imagine a baby trying to learn what is a window or what is not a window,

- We start to show him windows and he detects at an initial phase that all windows have glasses, and a frame and you can look outside, some of them may be opened.
- If we keep showing the same windows the baby may also falsely deduce that all windows are green, and that all green frames are windows.
- Thus overfitting the problem.

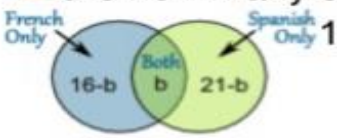
KNN CLASSIFIER SOLVED EXAMPLE - 30 APR 2021

NAME	ACID DURABILITY (XI)	STRENGTH (YI)	CLASS	XT - X1	SQUARE (XT - XI)	YT - Y1	SQUARE (YT - YI)	SUM (SQUARES)	SQRT	RANK
TYPE-1	7	7	BAD	-4	16	0	0	16	4	3
TYPE-2	7	4	BAD	-4	16	3	9	25	5	4
TYPE-3	3	4	GOOD	0	0	3	9	9	3	1
TYPE-4	1	4	GOOD	2	4	3	9	13	3.606	2

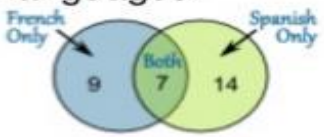
TEST INSTANCE VALUES		
ACID DURABILITY	3	XT
STRENGTH	7	YT
CLASS LABEL = ?		

Example

- 16 people study French, 21 study Spanish and there are 30 altogether. Work out the probabilities.
  - This is definitely a case of not Mutually Exclusive (you can study French AND Spanish).
    - b is how many study both languages.



1−b) = 30



Multiplication rule for Independent event

$$P(A \text{ and } B) = P(A \cap B) = P(A) * P(B)$$

Example : Suppose we roll one die followed by another and want to find the probability of rolling a 4 on the first die and rolling an even number on the second die.

$$P(A \cap B) = 1/12$$

Definition 8.3: **Joint Probability**

If *P(A)* and *P(B)* are the probability of two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If A and B are mutually exclusive, then $P(A \cap B) = 0$
If A and B are independent events, then $P(A \cap B) = P(A).P(B)$

Thus, for mutually exclusive events

$$P(A \cup B) = P(A) + P(B)$$

Definition 8.2: **Conditional Probability**

If events are dependent, then their probability is expressed by conditional probability. The probability that A occurs given that B is denoted by $P(A|B)$.

Suppose, A and B are two events associated with a random experiment. The probability of A under the condition that B has already occurred and $P(B) \neq 0$ is given by

$$P(A|B) = \frac{\text{Number of events in } B \text{ which are favourable to } A}{\text{Number of events in } B}$$

$$= \frac{\text{Number of events favourable to } A \cap B}{\text{Number of events favourable to } B}$$

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

Conditional Probability

- Generalization of Conditional Probability:

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

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

By the law of total probability : $P(B) = P[(B \cap A) \cup (B \cap \bar{A})]$, where $\bar{A}$ denotes the compliment of event A. Thus,

$$P(A|B) = \frac{P(B|A)\cdot P(A)}{P[(B \cap A) \cup (B \cap \bar{A})]}$$

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

Definition 8.3: **Total Probability**

Let $E_1, E_2, \ldots \ldots E_n$ be n mutually exclusive and exhaustive events associated with a random experiment. If A is any event which occurs with E_1 *or* E_2 *or* $\ldots \ldots E_n$, then

$$P(A) = P(E_1).P(A|E_1) + P(E_2).P(A|E_2) + \cdots \ldots \ldots \ldots . + P(E_n).P(A|E_n)$$

Theorem 8.4: **Bayes' Theorem**

Let $E_1, E_2, \ldots \ldots E_n$ be n mutually exclusive and exhaustive events associated with a random experiment. If A is any event which occurs with E_1 *or* E_2 *or* $\ldots \ldots E_n$, then

$$P(E_i|A) = \frac{P(E_i).P(A|E_i)}{\sum_{i=1}^{n} P(E_i).P(A|E_i)}$$

Prior and Posterior Probabilities

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

Example 8.6: 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

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

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.

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

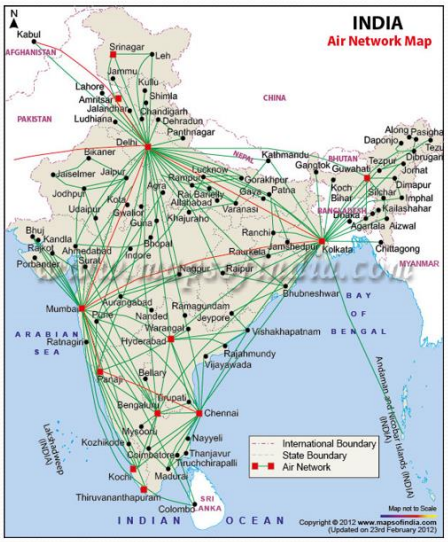
Naïve Bayesian Classifier

- 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 $X = x$.

BAYES THEOREM EXAMPLE

Example: Bayesian Classification

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



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 eventually to map this tuple into an accurate class.

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

- 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 eventually to map 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.

Attribute		Class: 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

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

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**

A Practice Example

Example 8.4

Class:
C1:buys_computer = 'yes'
C2:buys_computer = 'no'

Data instance
X = (age <=30,
Income = medium,
Student = yes
Credit_rating = fair)

age	income	student	credit_rating	com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

- $P(C_i)$: P(buys_computer = "yes") = 9/14 = 0.643
  P(buys_computer = "no") = 5/14= 0.357

- Compute $P(X|C_i)$ for each class
  P(age = "<=30" | buys_computer = "yes") = 2/9 = 0.222
  P(age = "<= 30" | buys_computer = "no") = 3/5 = 0.6
  P(income = "medium" | buys_computer = "yes") = 4/9 = 0.444
  P(income = "medium" | buys_computer = "no") = 2/5 = 0.4
  P(student = "yes" | buys_computer = "yes) = 6/9 = 0.667
  P(student = "yes" | buys_computer = "no") = 1/5 = 0.2
  P(credit_rating = "fair" | buys_computer = "yes") = 6/9 = 0.667
  P(credit_rating = "fair" | buys_computer = "no") = 2/5 = 0.4

- **X = (age <= 30 , income = medium, student = yes, credit_rating = fair)**

$P(X|C_i)$: P(X|buys_computer = "yes") = 0.222 x 0.444 x 0.667 x 0.667 = 0.044
P(X|buys_computer = "no") = 0.6 x 0.4 x 0.2 x 0.4 = 0.019

$P(X|C_i)*P(C_i)$: P(X|buys_computer = "yes") * P(buys_computer = "yes") = 0.028
P(X|buys_computer = "no") * P(buys_computer = "no") = 0.007

Therefore, X belongs to class ("buys_computer = yes")

Bayes' Theorem

The probability of event A, given that event B has subsequently occurred, is

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{[P(A) \cdot P(B|A)] + [P(\overline{A}) \cdot P(B|\overline{A})]}$$

That's a formidable expression, but we will simplify its calculation. See the following example, which illustrates use of the above expression, but also see the alternative method based on a more intuitive application of Bayes' theorem.

Example 2

In Orange County, 51% of the adults are males. (It doesn't take too much advanced mathematics to deduce that the other 49% are females.) One adult is randomly selected for a survey involving credit card usage.

a. Find the prior probability that the selected person is a male.

b. It is later learned that the selected survey subject was smoking a cigar. Also, 9.5% of males smoke cigars, whereas 1.7% of females smoke cigars (based on data from the Substance Abuse and Mental Health Services Administration). Use this additional information to find the probability that the selected subject is a male.

Solution

Let's use the following notation:

M = male $\quad\quad$ $\overline{\text{M}}$ = female (or not male)

C = cigar smoker $\quad\quad$ $\overline{\text{C}}$ = not a cigar smoker.

a. Before using the information given in part b, we know only that 51% of the adults in Orange County are males, so the probability of randomly selecting an adult and getting a male is given by $P(\text{M}) = 0.51$.

b. Based on the additional given information, we have the following:

$P(\text{M}) = 0.51$	because 51% of the adults are males
$P(\overline{\text{M}}) = 0.49$	because 49% of the adults are females (not males)
$P(\text{C}\|\text{M}) = 0.095$	because 9.5% of the males smoke cigars (That is, the probability of getting someone who smokes cigars, given that the person is a male, is 0.095.)
$P(\text{C}\|\overline{\text{M}}) = 0.017.$	because 1.7% of the females smoke cigars (That is, the probability of getting someone who smokes cigars, given that the person is a female, is 0.017.)

Let's now apply Bayes' theorem by using the preceding formula with M in place of A, and C in place of B. We get the following result:

$$P(M \mid C) = \frac{P(M) \cdot P(C|M)}{[P(M) \cdot P(C|M)] + [P(\overline{M}) \cdot P(C|\overline{M})]}$$

$$= \frac{0.51 \cdot 0.095}{[0.51 \cdot 0.095] + [0.49 \cdot 0.017]}$$

$$= 0.85329341$$

$$= 0.853 \text{ (rounded)}$$

Before we knew that the survey subject smoked a cigar, there is a 0.51 probability that the survey subject is male (because 51% of the adults in Orange County are males). However, after learning that the subject smoked a cigar, we revised the probability to 0.853. There is a 0.853 probability that the cigar–smoking respondent is a male. This makes sense, because the likelihood of a male increases dramatically with the additional information that the subject smokes cigars (because so many more males smoke cigars than females).

For the preceding example, simply assume some value for the adult population of Orange County, such as 100,000, then use the given information to construct a table, such as the one shown below.

Finding the number of males who smoke cigars: If 51% of the 100,000 adults are males, then there are 51,000 males. If 9.5% of the males smoke cigars, then the number of cigar–smoking males is 9.5% of 51,000, or 0.095 × 51,000 = 4845. See the entry of 4845 in the table. The other males who do *not* smoke cigars must be 51,000 – 4845 = 46,155. See the value of 46,155 in the table.

Finding the number of females who smoke cigars: Using similar reasoning, 49% of the 100,000 adults are females, so the number of females is 49,000. Given that 1.7% of the females smoke cigars, the number of cigar–smoking females is 0.017 × 49,000 = 833. The number of females who do *not* smoke cigars is 49,000 – 833 = 48,167. See the entries of 833 and 48,167 in the table.

	C (Cigar Smoker)	$\overline{C}$ (Not a Cigar Smoker)	**Total**
M (male)	4845	46,155	**51,000**
$\overline{M}$ (female)	833	48,167	**49,000**
Total	**5678**	**94,322**	**100,000**

The above table involves relatively simple arithmetic. Simply partition the assumed population into the different cell categories by finding suitable percentages.

Now we can easily address the key question as follows: To find the probability of getting a male subject, given that the subject smokes cigars, simply use the same conditional probability described in the textbook. To find the probability of getting a male given that the subject smokes, restrict the table to the column of cigar smokers, then find the probability of getting a male in that column. Among the 5678 cigar smokers, there are 4845 males, so the probability we seek is 4845/5678 = 0.85329341. That is, $P(M \mid C) = 4845/5678 = 0.85329341 = 0.853$ (rounded).