**第1章 大数据介绍**

**1. Three attributes defining Big Data**

- Huge volume of data

- Complexity of data types and structures

- Speed of new data creation and growth

**2. 3V:** Volume, Variety, and Velocity

**3. 5V:** 3V + Value and Veracity

**第2章 大数据循环**

**1. Six phases**

(1) Discovery

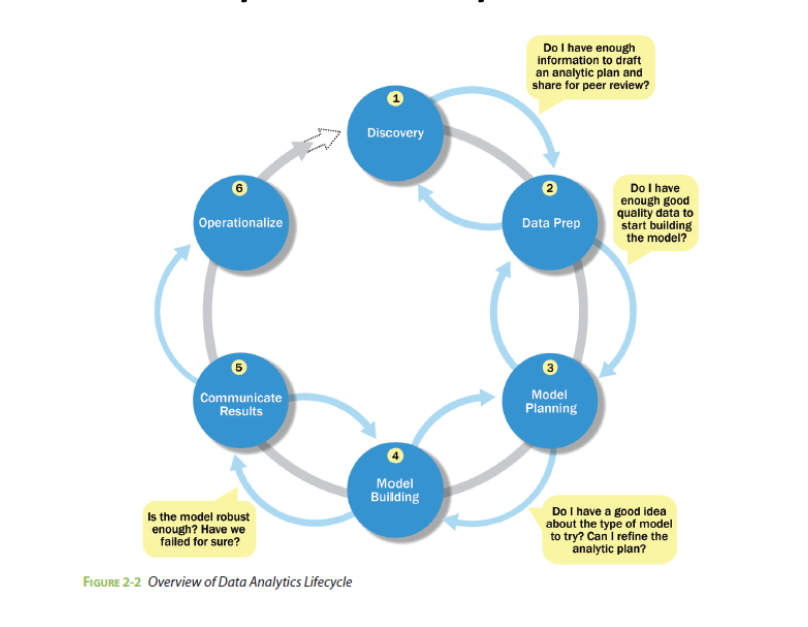
(2) Data preparation (analytic sandbox)

(3) Model planning (methods, techniques, workflow, variables, relationships, models)

(4) Model building (training and test datasets, software, and hardware)

(5) Communication results (identify key findings)

(6) Operationalize (delivery, pilot project)



**第4章 K-means聚类**

**K-means Clustering:**

**1. Four steps:**

(1) Choose the value of k and the k initial guess for the centroids.

(2) Compute the distance from each data point to each centroid. Assign each point to the closest centroid.

(3) Update the centroid of each cluster.

(4) Repeat Steps 2 and 3 until convergence.

**2. Compute the Euclidean distance**

**3. Compute the centroid for a cluster**

x代表某一维坐标

**4. Within Sum of Squares (WSS)**

Sum of the squares of the distances between each data point and the closest centroid.

**第5章 关联规则**

**1. Association Rules:**

- An unsupervised learning method

- Descriptive, not predictive

- Discover interesting, hidden relationship

Represented as rules or frequent itemsets

- Commonly used for mining transactions in databases

**2. Apriori Algorithm:**

**(1) Apriori property (downward closure property)**

- If an itemset is frequent, then any subset of this itemset must also be frequent

- It provides the basis for the Apriori algorithm

**(2) It takes a bottom-up iterative approach to uncovering frequent itemsets:**

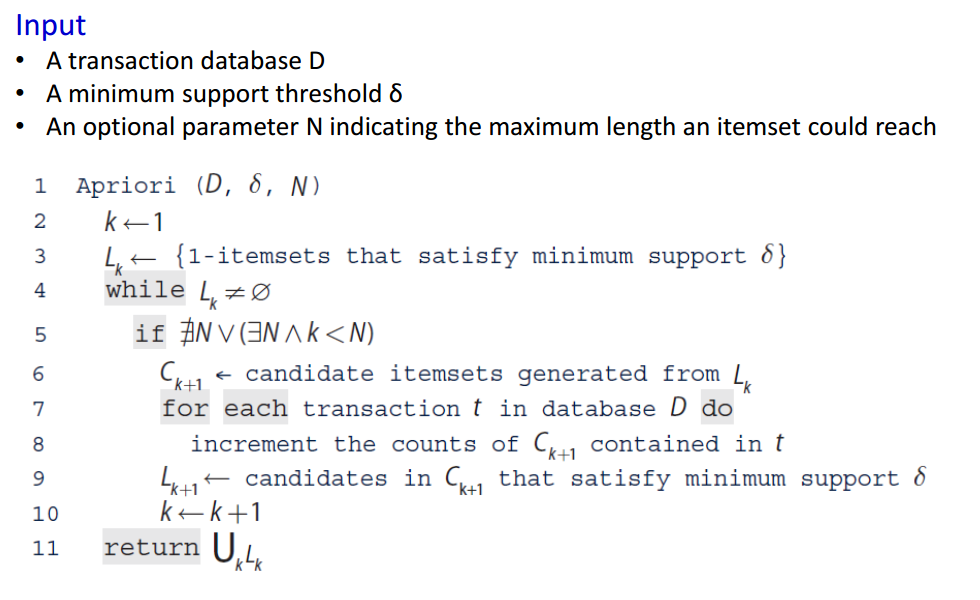
- First, identify all frequent items (or 1-itemsets)

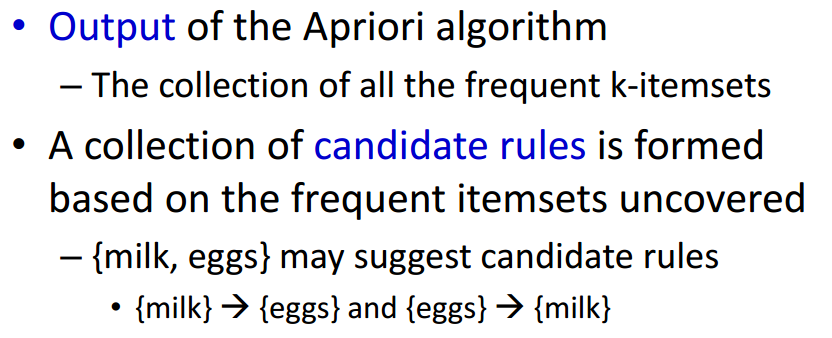
- The identified frequent 1-itemsets are paired into 2-itemsets to identify frequent 2-itemsets

- Grow the size of identified frequent itemsets and identify again

- Repeat this process until it runs out of support or the itemsets reach a predefined length

**(3) Pseudocode:**





**(4) Evaluation of Candidate Rules**

Four measures: Support, Confidence, Lift and Leverage

- Support: Given an item X, the support of X is the percentage of transactions that contain X. Denoted by support(X)

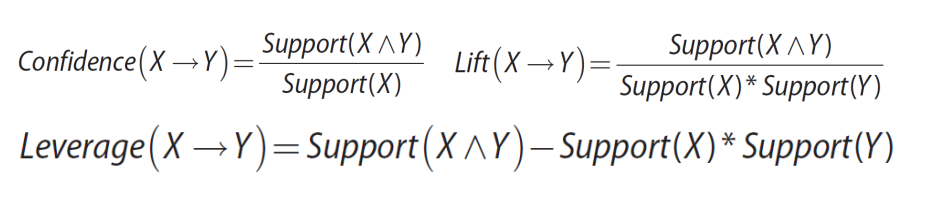
- Confidence: The measure of certainty or trustworthiness associated with each rule.

- Lift: Measures how many times more often X and Y occur together than expected if they are statistically independent of each other. Measures how X and Y are really related rather than coincidentally happening together.

- Leverage: Measures the difference in the probability of X and Y appearing together compared to what would be expected if X and Y were statistically independent of each other.

- A high-confidence rule can sometimes be misleading.

- Lift and leverage not only ensure interesting rules but also filter out coincidental rules.



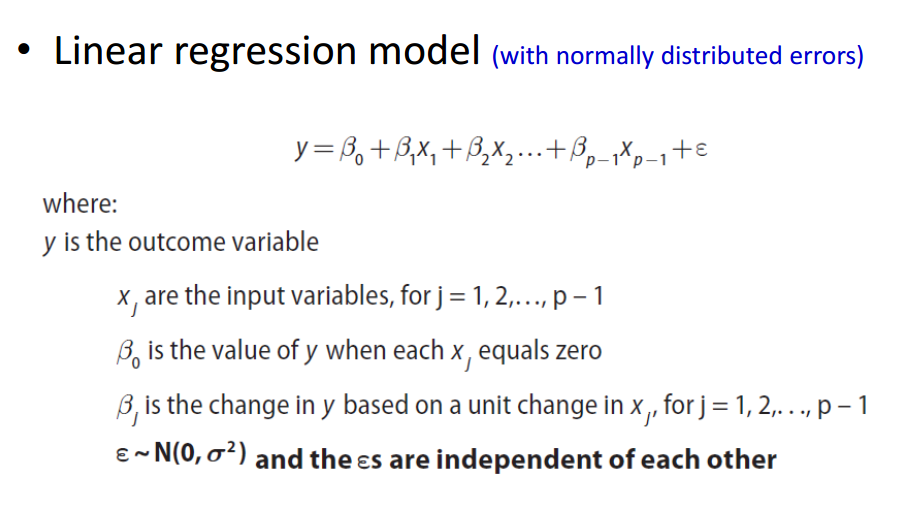
**第6章 回归**

**1. Linear Regression**

- An analytical technique used to model the relationship between several input variables and a continuous outcome variable.

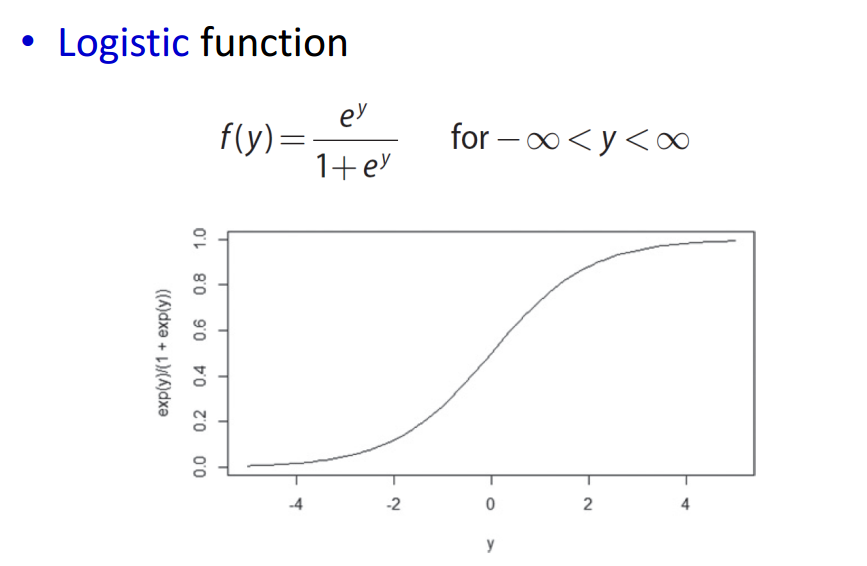
- A key assumption: the relationship is linear.

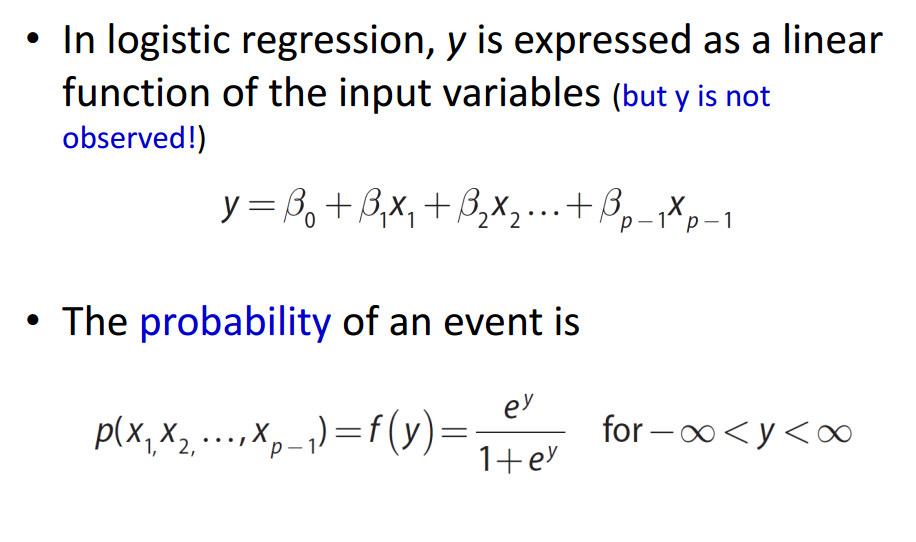
- Non-deterministic nature: Accounts for the randomness in an outcome. Provides the expected value of the outcome.

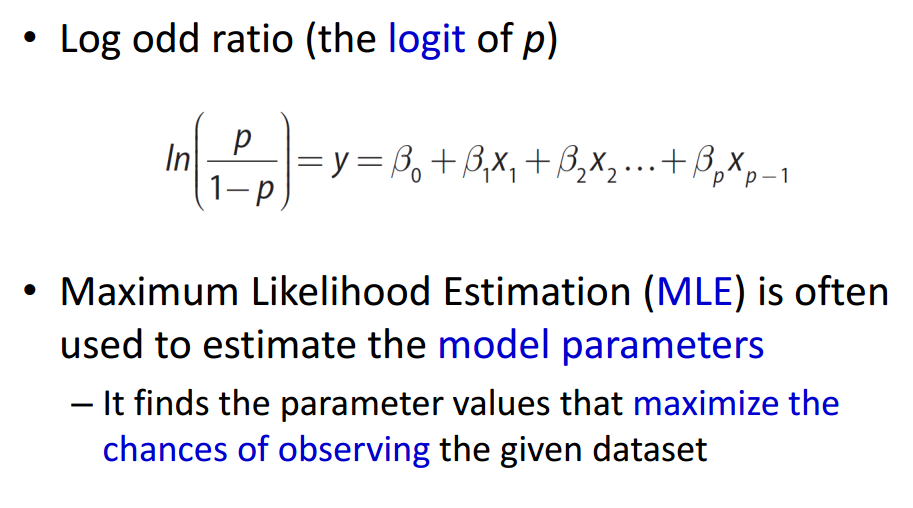


**2. Logistic Regression**

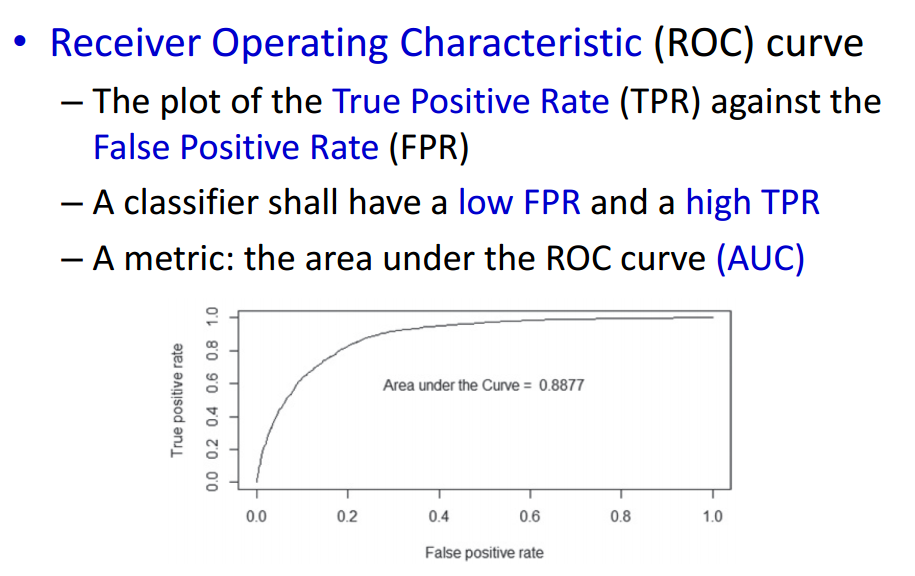
When the outcome variable is categorical in nature, logistic regression can be used to predict the probability of an outcome based on the input variables.





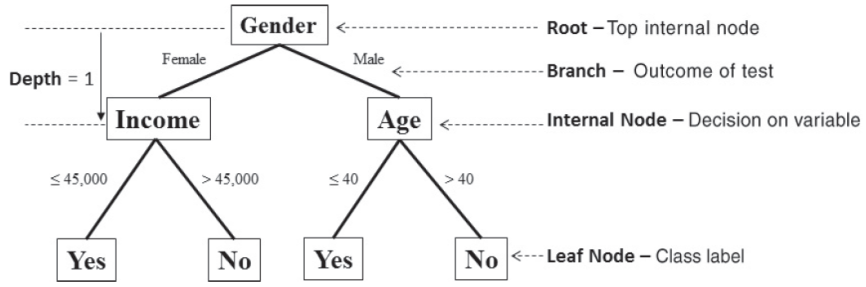


**3. ROC Curve**



**第7章 分类**

**1. Decision Tree**



- A decision tree uses a tree structure to specify sequences of decisions and consequences.

- Given input variable X = {x1, x2, …… , xn}, the goal is to predict an output variable Y.

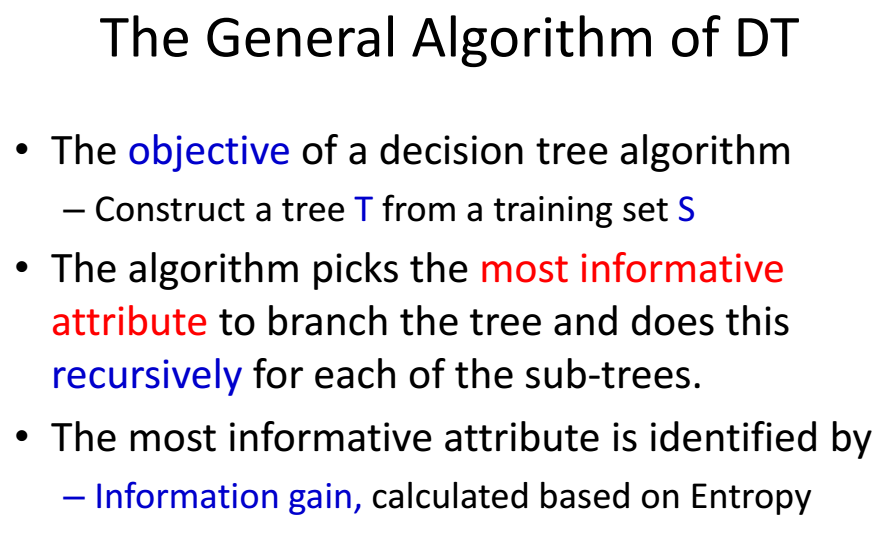
- Each node tests a particular input variable.

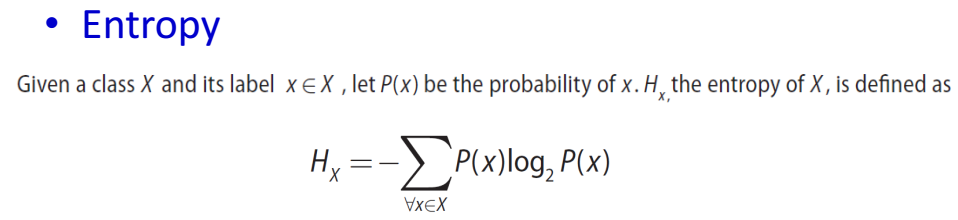
- Each branch represents the decision made

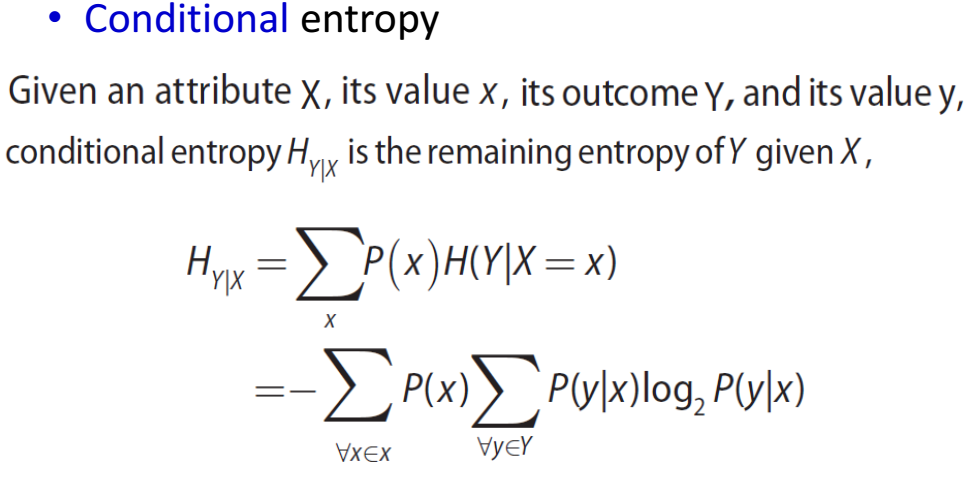
- Classifying a new observation is to traverse this decision tree.

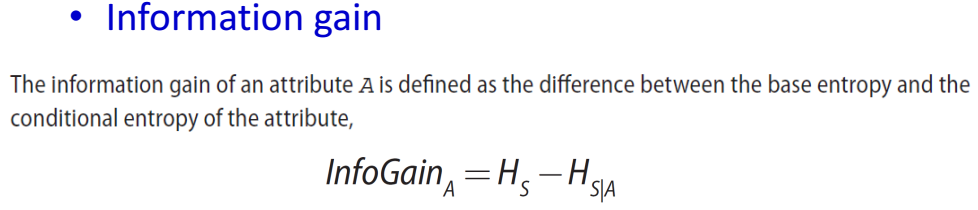
- The depth of a node is the minimum number of steps required to reach the node from root.

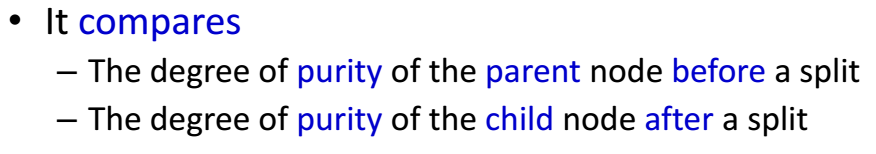
- Leaf nodes are at the end of the last branches on the tree, representing class labels.



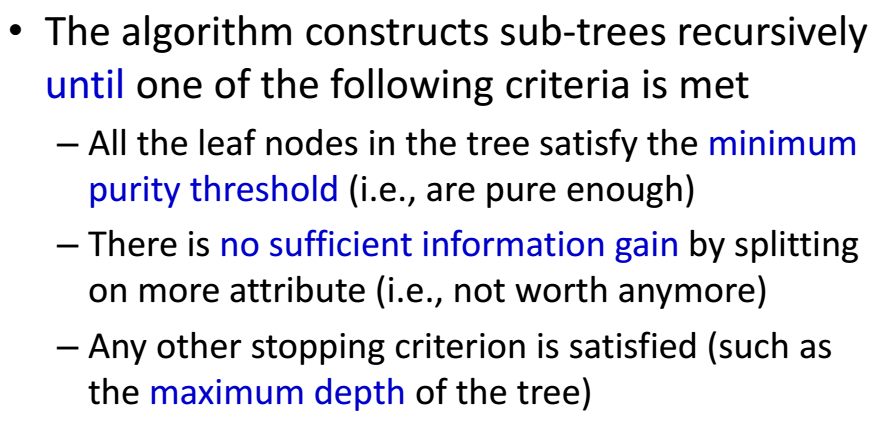


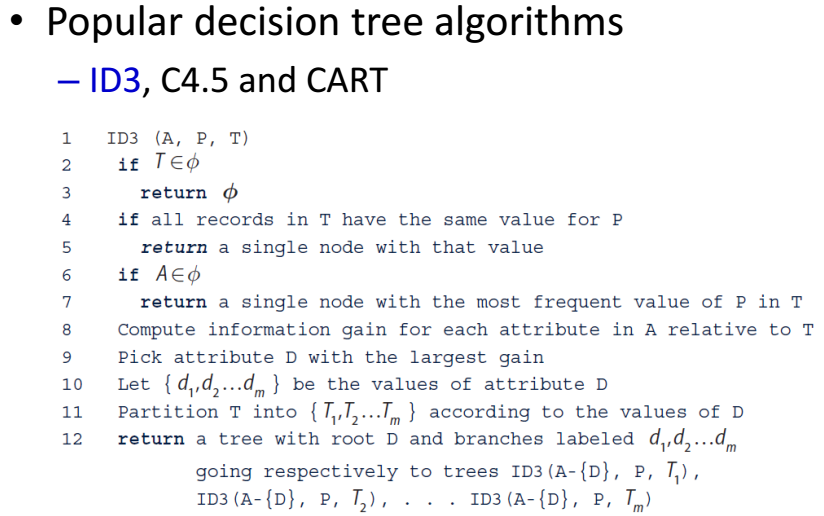


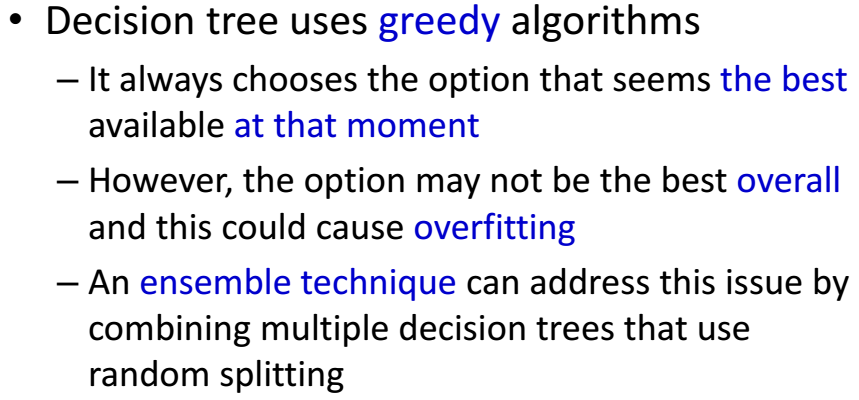


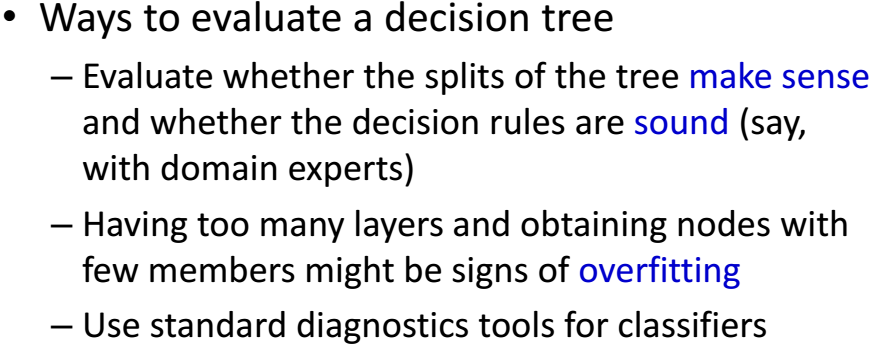


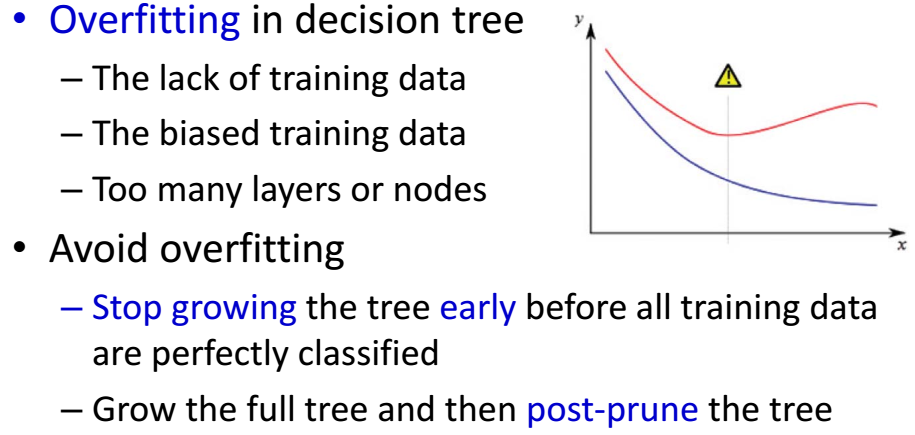
- The algorithm splits on the attribute with the largest information gain at each round.



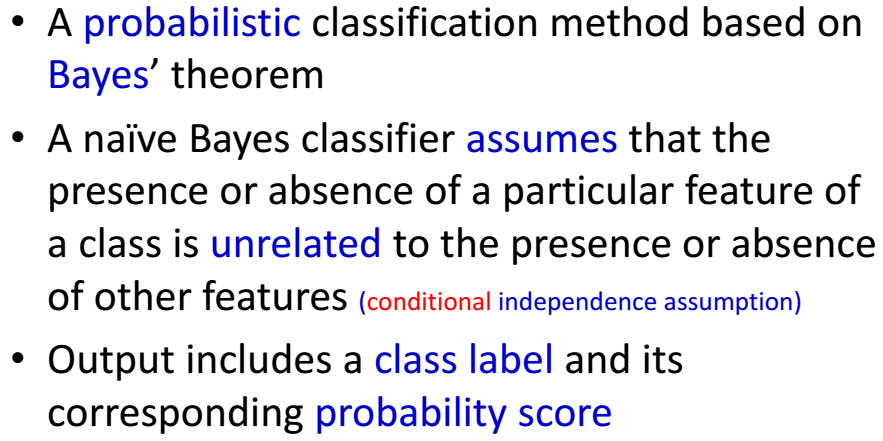


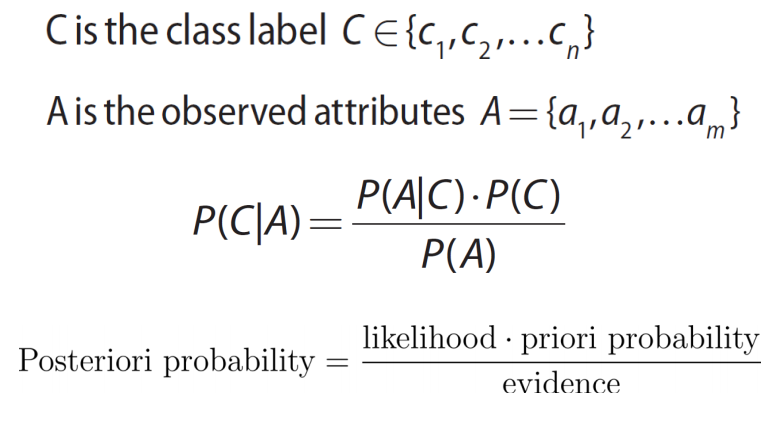


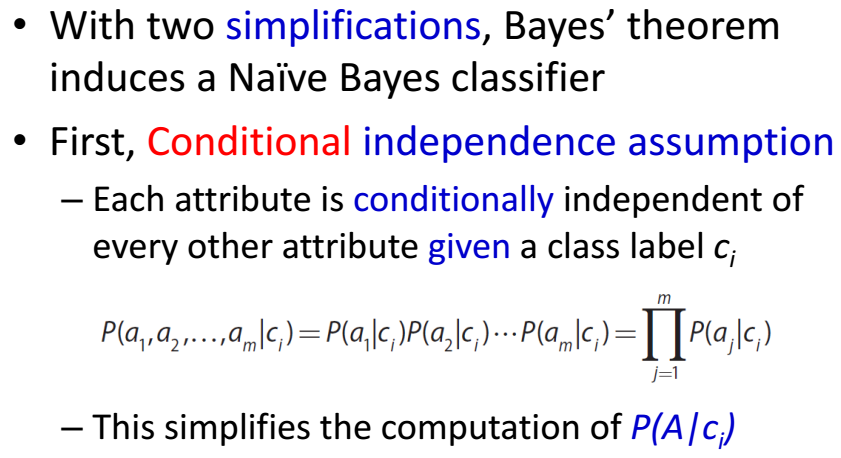


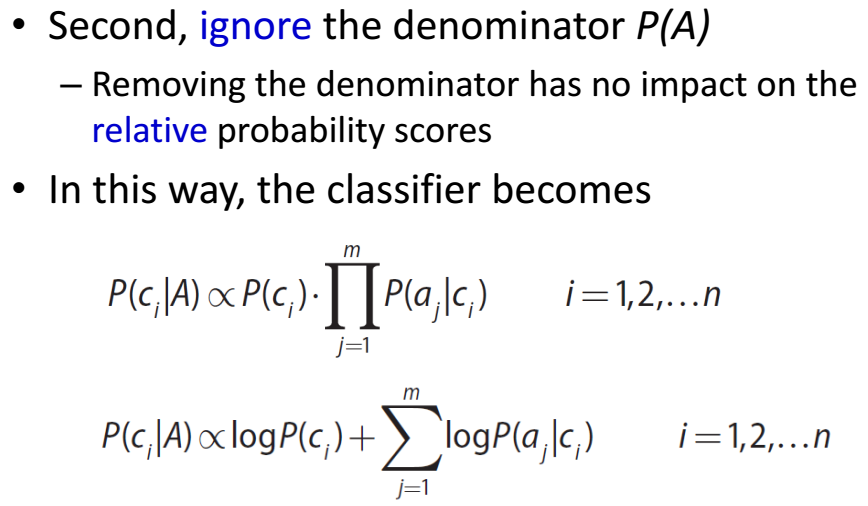


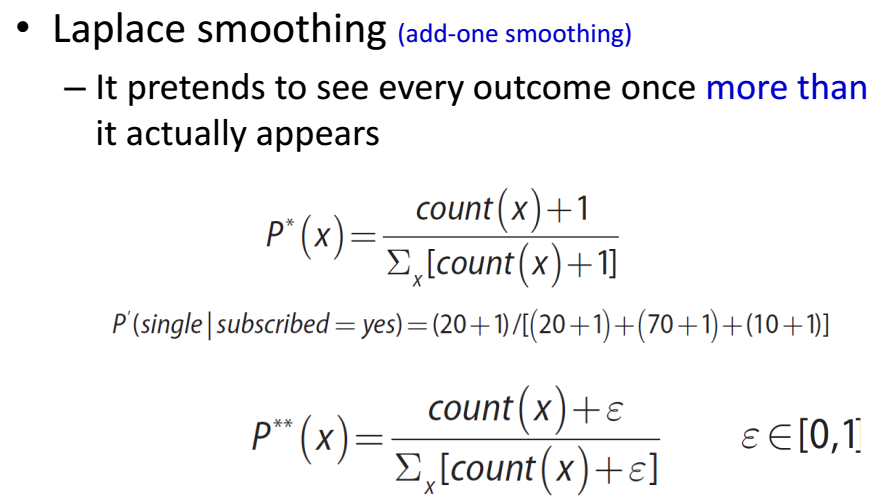
**2. Naive Bayes**











**第8章 文本分析**

**1. Text Analysis Steps**

(1) Parsing: Takes unstructured text and imposes a structure for further analysis.

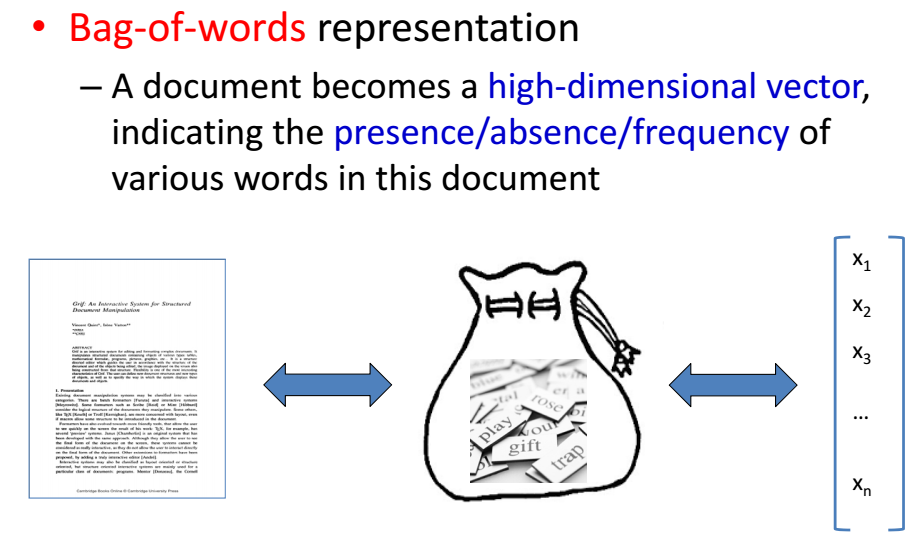
(2) Search and retrieval: Identification of the documents in a corpus that contain search items.

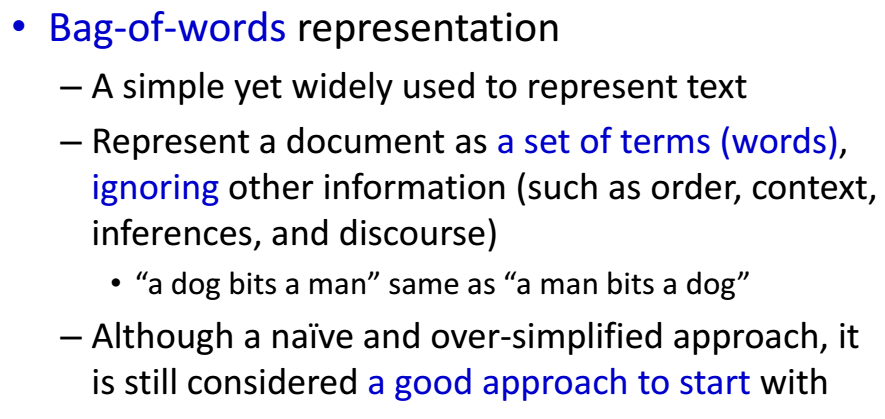
(3) Text mining: Use the results of the prior steps to discover meaningful insights.

- Clustering and classification techniques can be adapted to text mining.

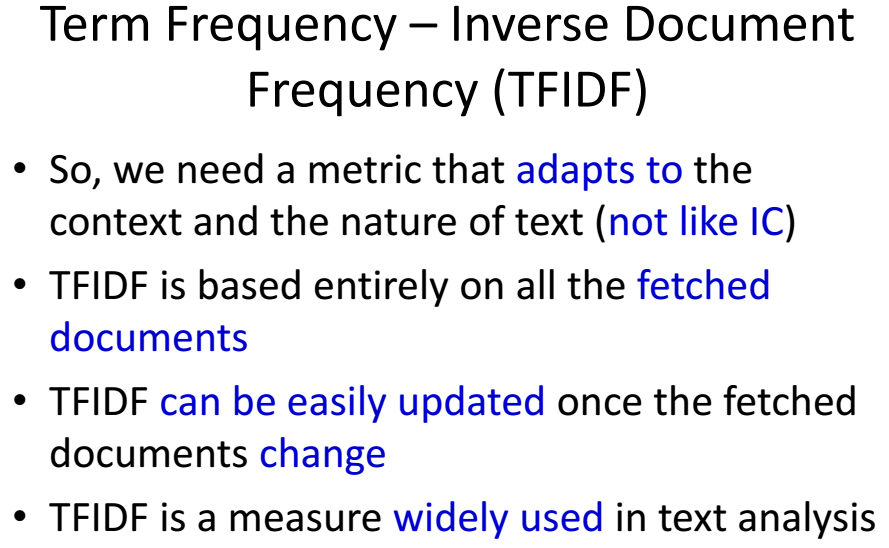
- Utilizes various methods and techniques: statistical analysis, information retrieval, data mining and NLP.

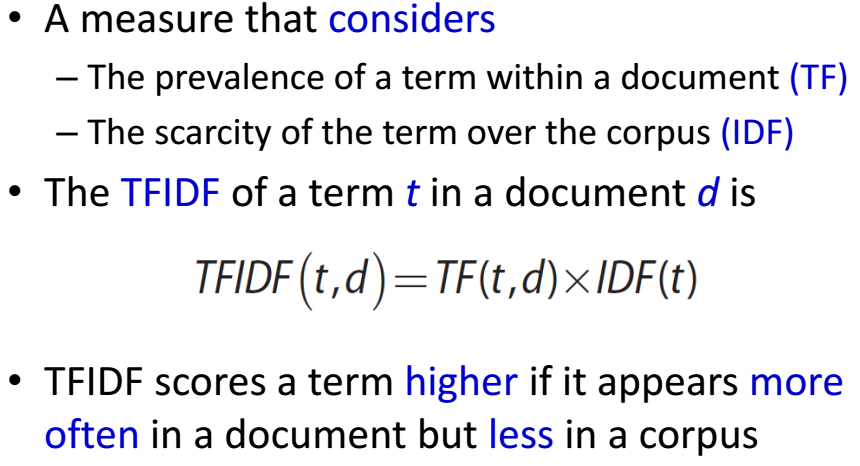
**2. Bag of words**

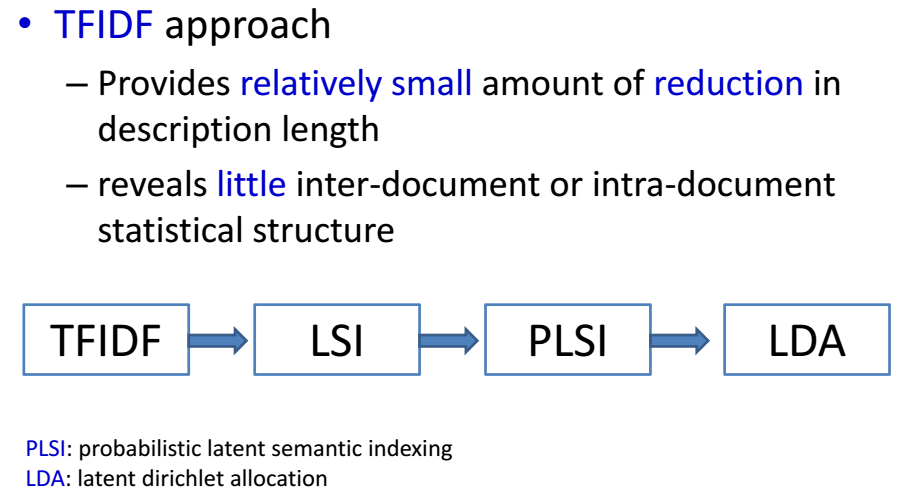




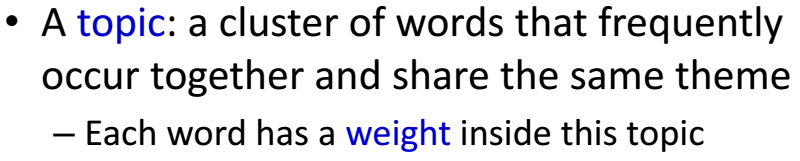
**3. TFIDF**

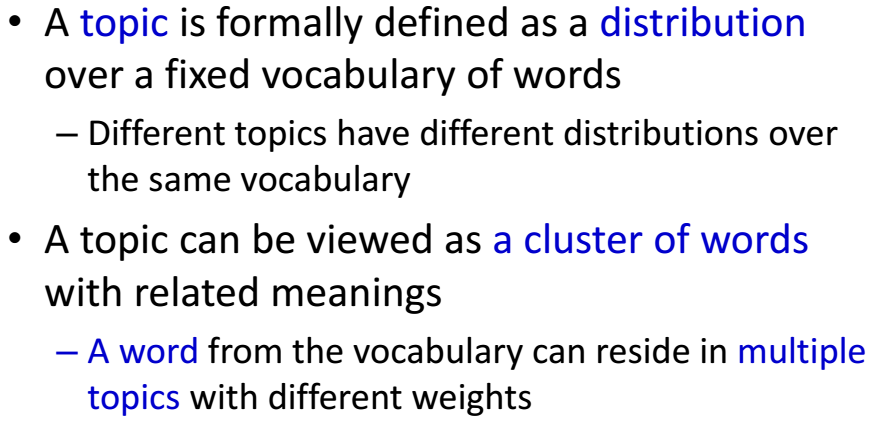


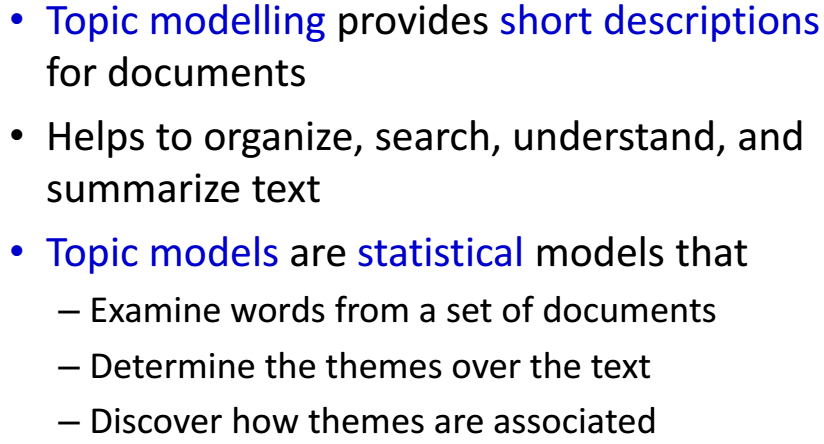


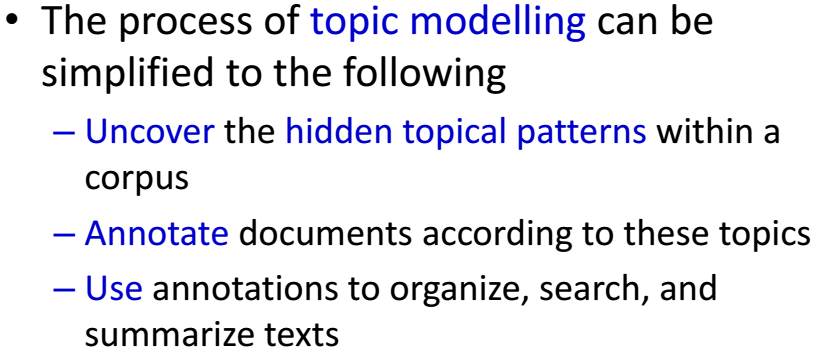


**4. Topic Model**



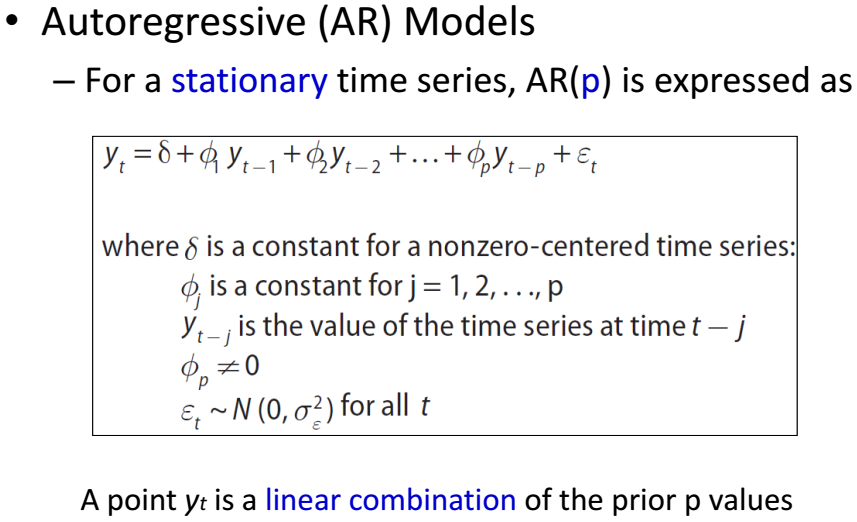




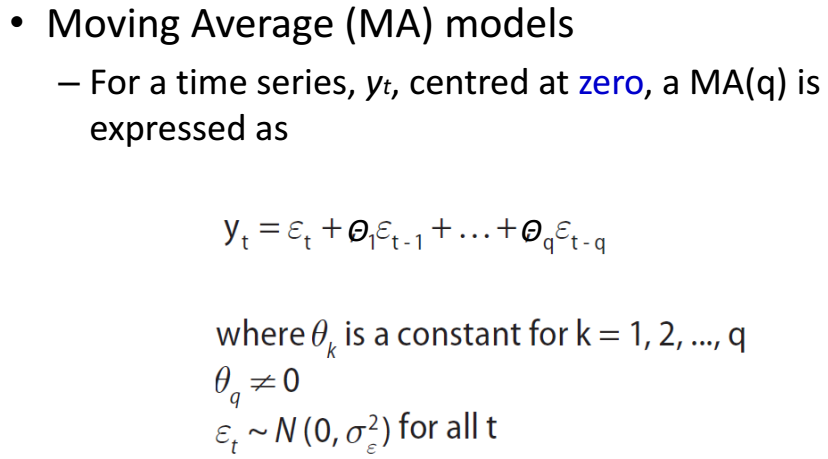


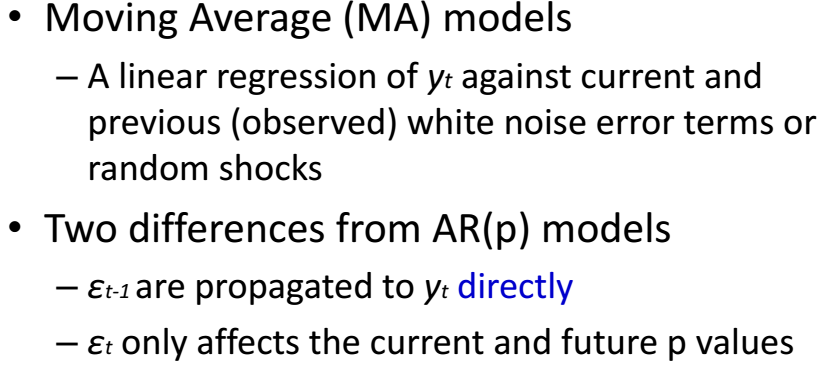
**第11章 时间序列**

**1. Autoregressive Model**

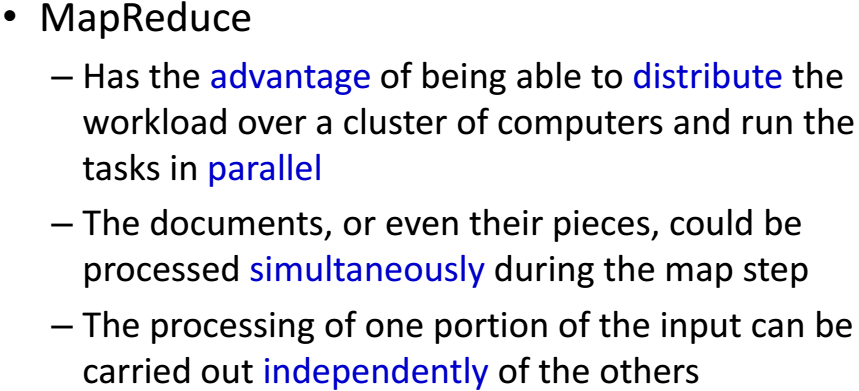


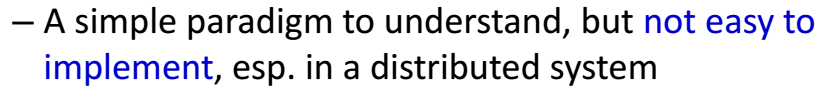
**2. Moving Average Model**





**第12章 MapReduce和Hadoop**





给定一个很大的文档集，如何用MapReduce统计每个单词出现的次数？

