

# Two Main Branches of Learning



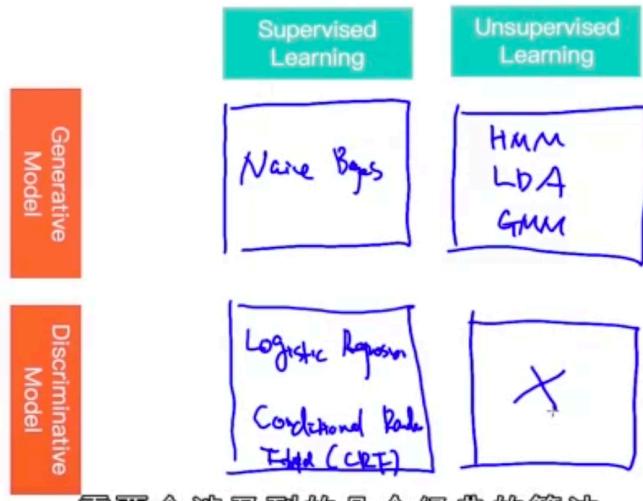
## 1. machine learning

### 定义

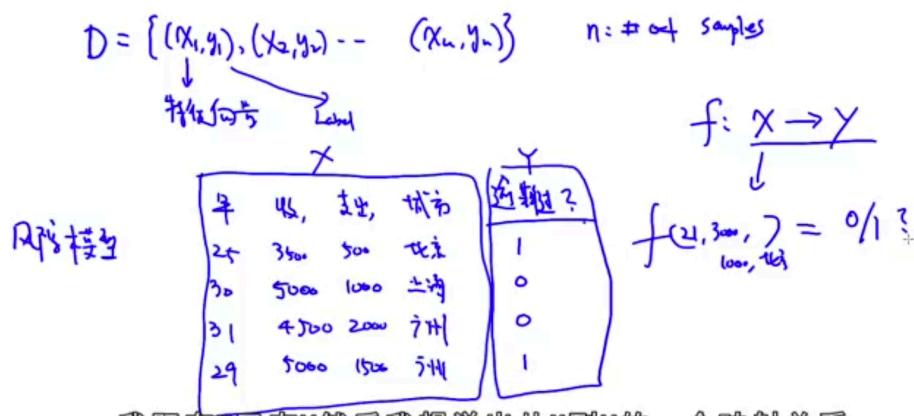
自动从已有的数据里找出一些规律，然后把学到的这些规律应用到对未来数据 (future data) 的预测中，或者在不确定环境下自动地做一些决策

核心：自动

(区别于，专家系统：人工)



### 1.1 supervised learning



e.g. sentiment classification

e.g. algorithms

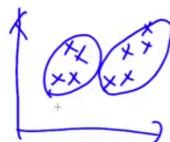
- 线性回归 (Linear Regression)
- 逻辑回归 (Logistic Regression)
- 朴素贝叶斯 (Naïve Bayes)
- 神经网络 (Neural Network)
- SVM (Support Vector Machine)
- 随机森林 (Random Forest)
- Adaboost
- CNN (Convolutional Neural Network)

## 1.2 unsupervised learning

$$D = \{x_1, x_2, \dots, x_n\}$$

↓  
聚类 (clustering)  
(exploration)

$$f: x \rightarrow \text{cluster}$$

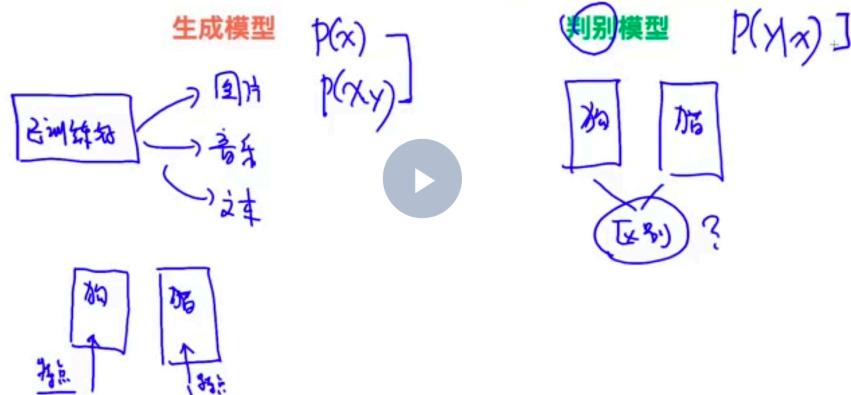


- K-means clustering
- PCA (Principal Component Analysis)
- ICA (Independent Component Analysis)
- MF (Matrix Factorization)
- LSA (Latent Semantic Analysis)
- LDA (Latent Dirichlet Allocation)

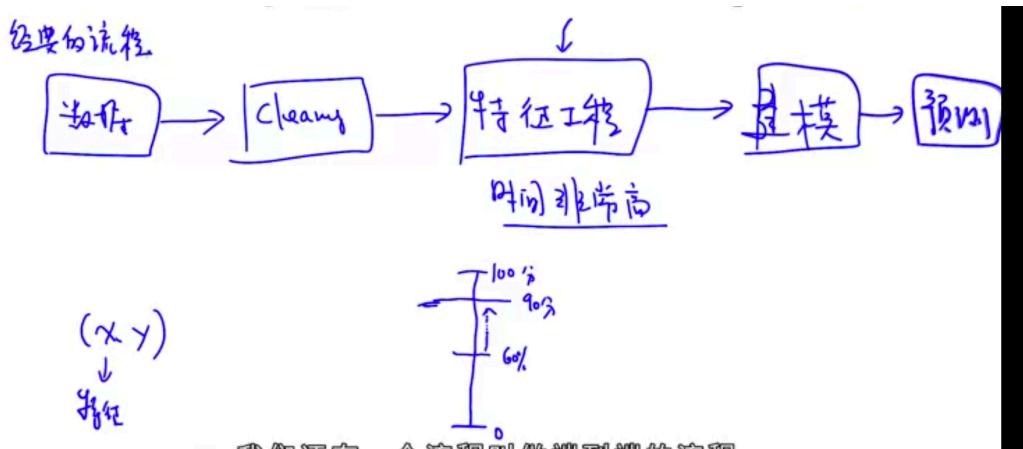
MF：推荐系统

LSA、LDA：文本分析

## 1.3 generative vs discriminative models



## 1.4 working pipelines



端到端：模型上集成了特征工程，所以节省了人力去选择特征。(多用于图像处理)

## 2. naïve Bayes

重点：文本分类、垃圾邮件分类



核心：统计一些敏感词，在正常邮件或垃圾邮件中，出现的概率。



正常邮件含有“购买”词的概率多少？

垃圾邮件含有“购买”词的概率多少？

## 2.1 prior information

我们需要一定的先验信息。例如，正常邮件在所有邮件中的概率。

## 2.2 条件独立 conditional independence

$$P(X, Y | z) = P(X|z) \cdot P(Y|z)$$

X和Y是条件独立于z

## 2.3 predictions



新邮件

$$\begin{aligned} & P(\text{正常} | \text{内容}) \geq ? \quad P(\text{垃圾} | \text{内容}) \\ = & \frac{P(\text{内容正常}) \cdot P(\text{正常})}{P(\text{内容})} \geq ? \quad \frac{P(\text{内容垃圾}) \cdot P(\text{垃圾})}{P(\text{内容})} \\ & \frac{P(\text{内容} | \text{正常}) P(\text{正常})}{P_{\text{new}}} \geq ? \quad \frac{P(\text{内容} | \text{垃圾}) P(\text{垃圾})}{P_{\text{new}}} \end{aligned}$$