机器学习介绍

www.ilinuxkernel.com

目录

机器学习是什么？

机器学习：计算机系统**分析历史数据**，来**预测未来趋势和行为**。

重点是**预测**，若能预期事情将如何发展，企业或个人即能早期投资以开创新商机，或者避免重大风险的发生。

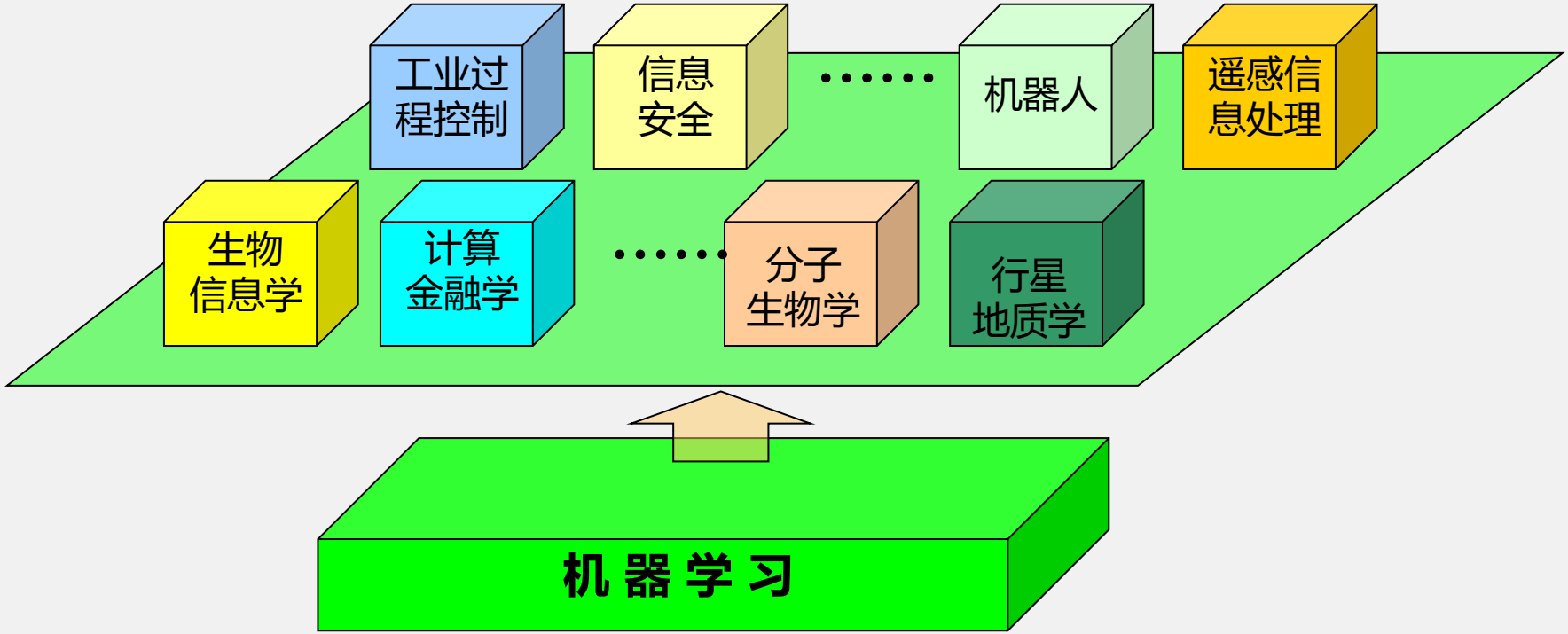
预测使用者行为来调整商业行为
-根据用户喜好推荐商品
-预测机器损坏的时间

分类
-判断信件是否为垃圾邮件
-判断客户是否续约或违约

分群
-社交网络划分性质相近的会员
-精准广告

- 机器学习是人工智能的核心研究领域之一
- **经典定义：**利用经验改善系统自身的性能
- 随着该领域的发展，主要做**智能数据分析**
- **典型任务：**根据现有数据建立预测模型

机器学习的用途



机器学习应用场景

Telemetry data analysis

Buyer propensity models

Social network analysis

Predictive maintenance

Web app optimization

Churn analysis

Natural resource exploration

Weather forecasting

Healthcare outcomes

Fraud detection

Life sciences research

Targeted advertising

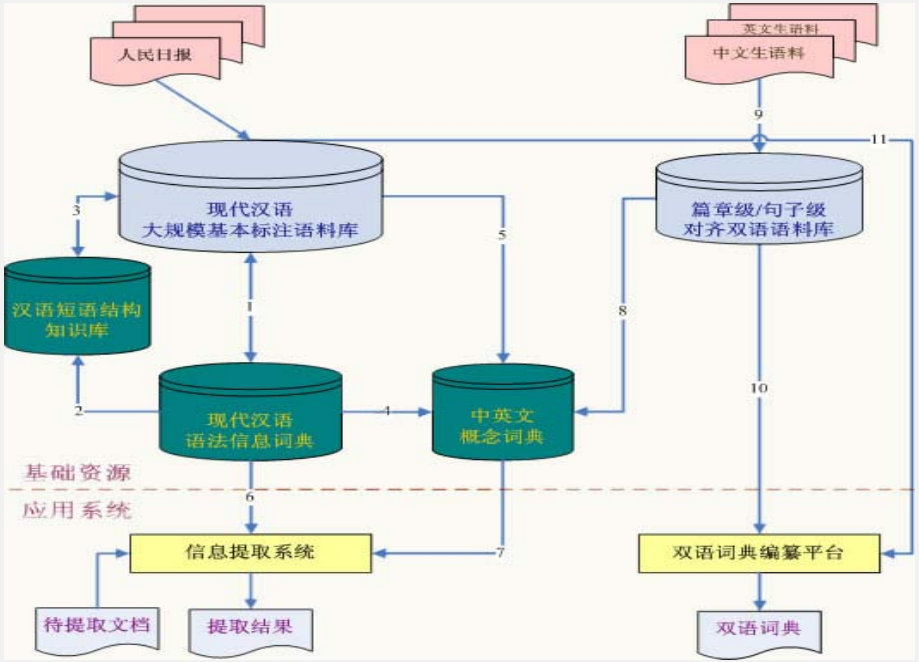
Network intrusion detection

Smart meter monitoring

机器学习应用 - *机器语言学*

主要应用：

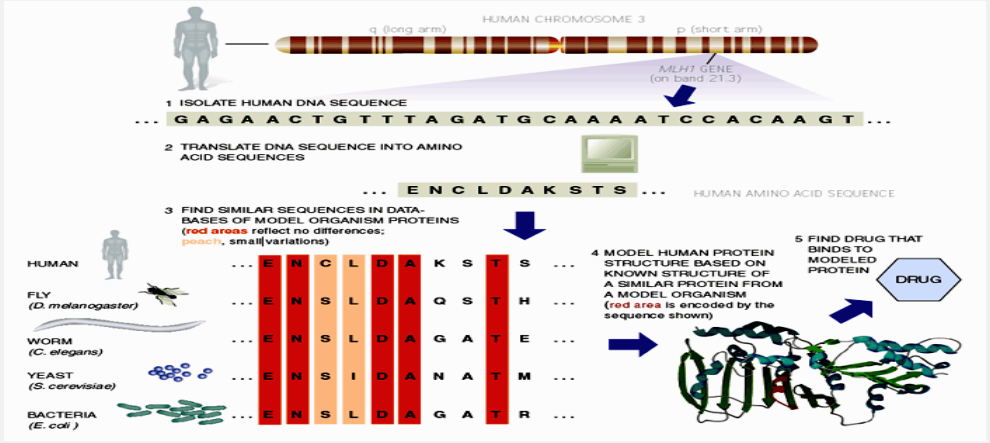
- 语音合成
- 语音识别
- 机器翻译
- 信息检索
- 信息抽取
- 问答系统



常用技术：

神经网络　隐马尔可夫模型　贝叶斯分类器　决策树
序列分析　聚类

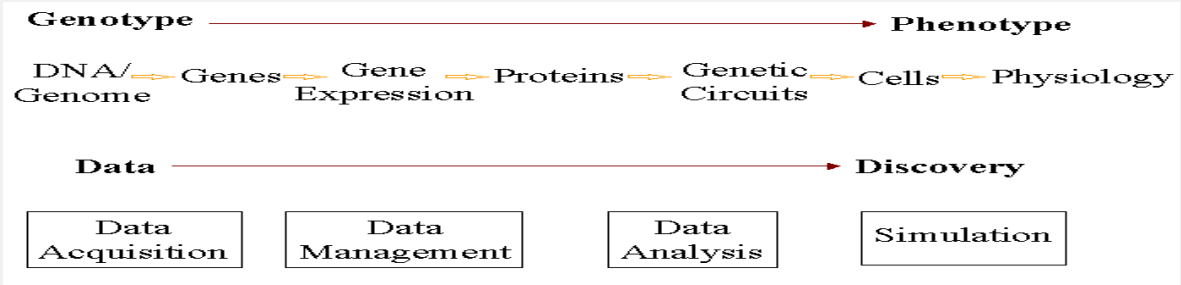
机器学习应用 - *生物信息学*



主要应用：

- 序列比对
- 基因识别
- 基因重组
- 蛋白质结构预测
- 基因表达
- 蛋白质反应预测

… …



常用技术：

神经网络　支持向量机

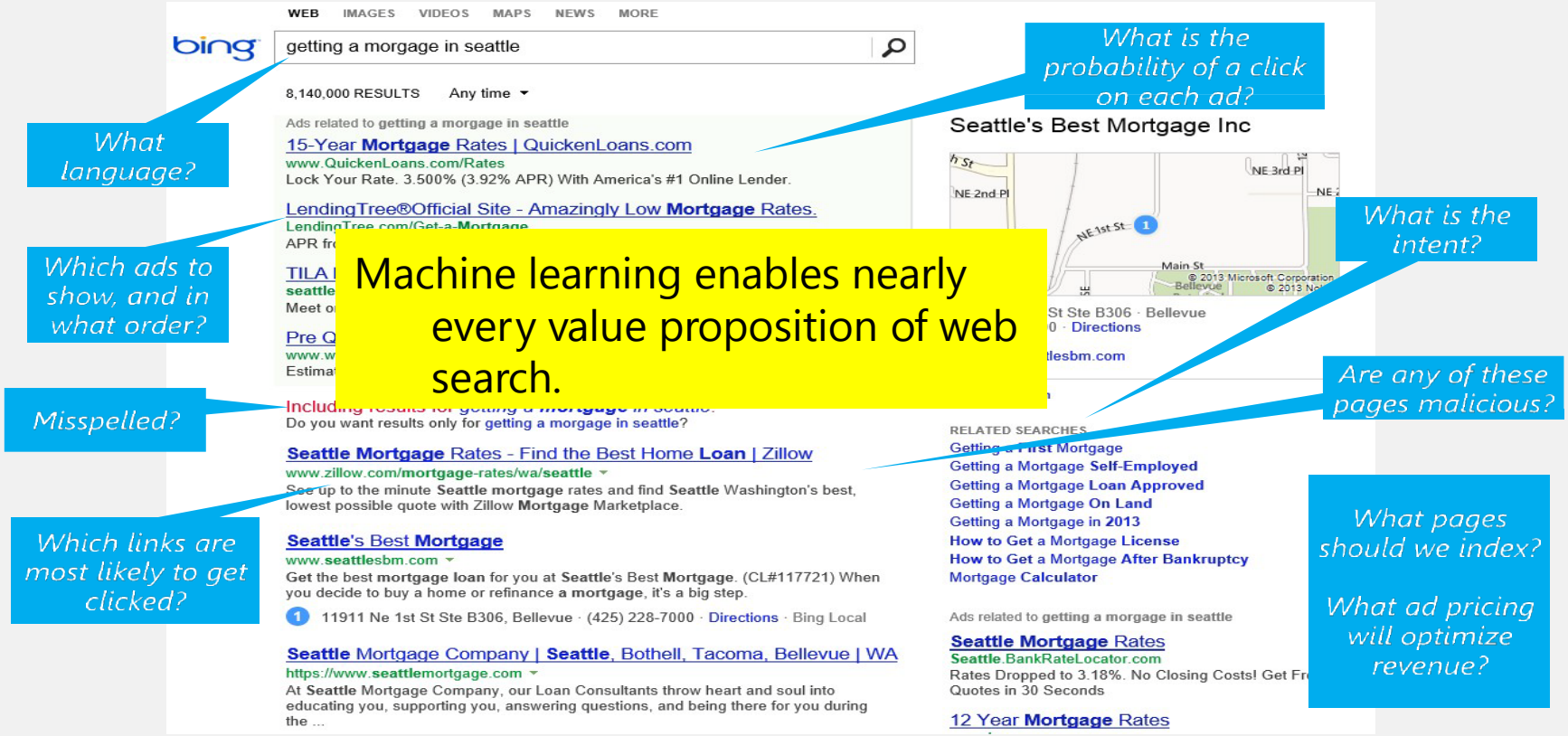
隐马尔可夫模型

贝叶斯分类器　k近邻

决策树　序列分析　聚类

…… ……

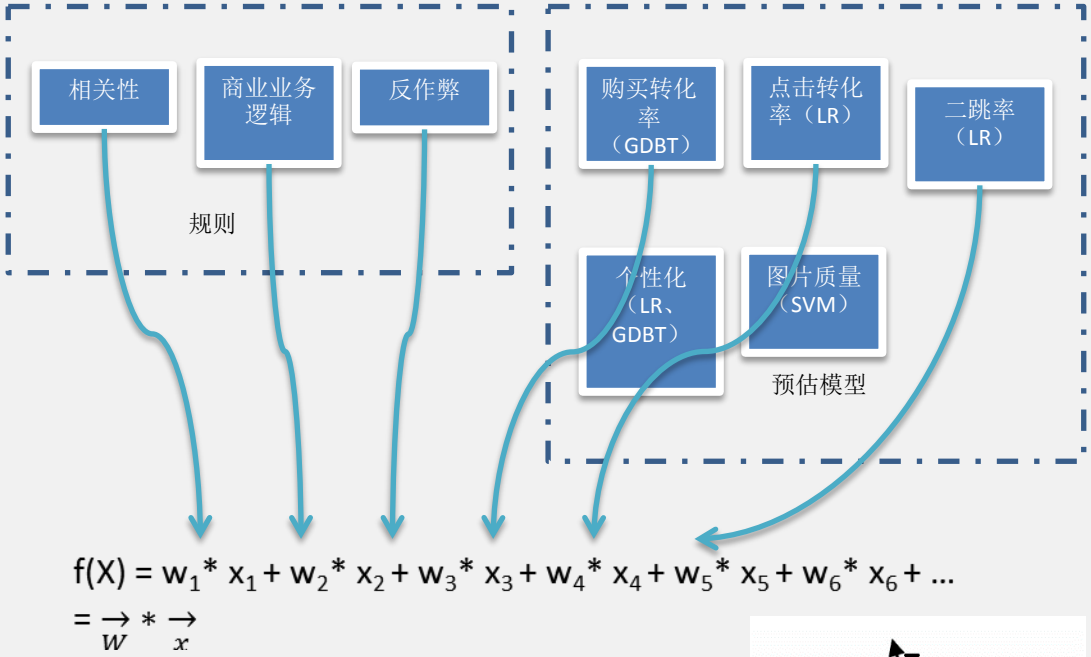
机器学习应用 - *Web搜索*



机器学习应用 - *购物推荐*

哪个商品更好？

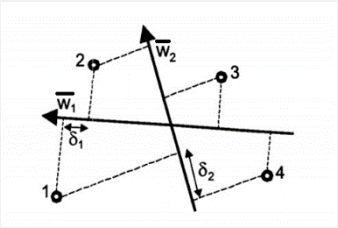
- *CTR*
- *CVR*
- *价格*



$$f(X) = w_1 * x_1 + w_2 * x_2 + w_3 * x_3 + w_4 * x_4 + w_5 * x_5 + w_6 * x_6 + \ldots$$

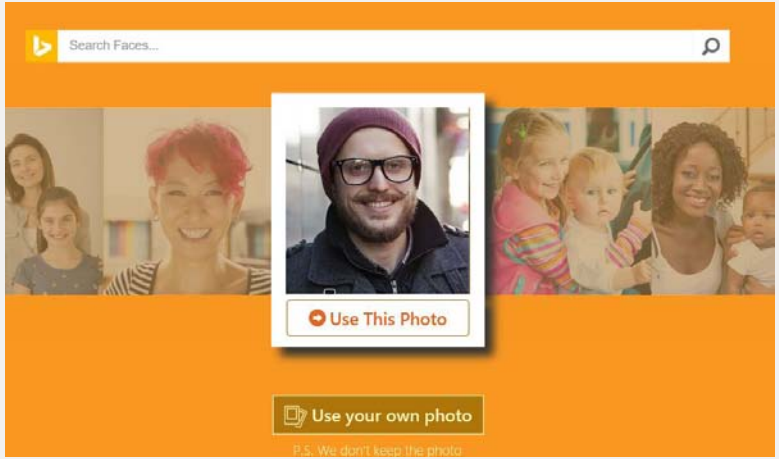
$$= \overrightarrow{W} * \overrightarrow{x}$$

- 通过线性模型来组合非线性的特征
- 计算效率高
- 可解释性好



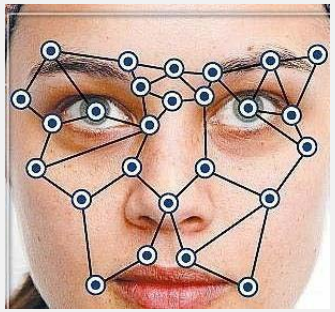
机器学习应用 – *其他*

您多大了？ http://www.how-old.net



在线/机器翻译

人脸识别

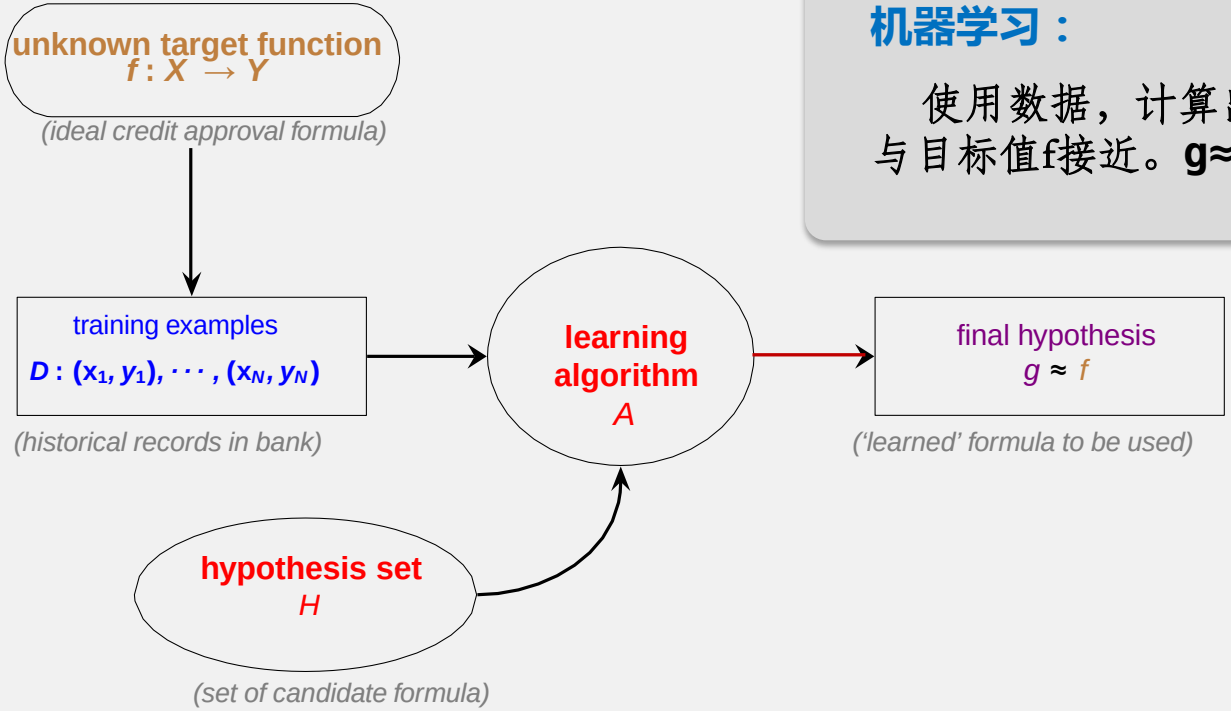


语音识别

股市预测

目录

机器学习定义

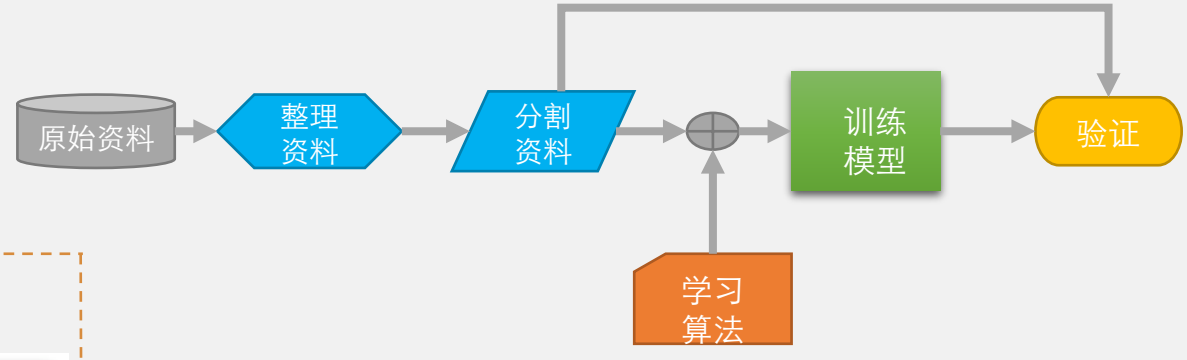


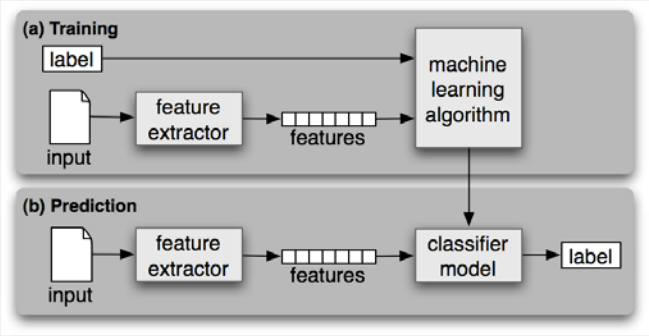
机器学习：

使用数据，计算出一个假设值g，使其尽可能与目标值f接近。**g≈f**

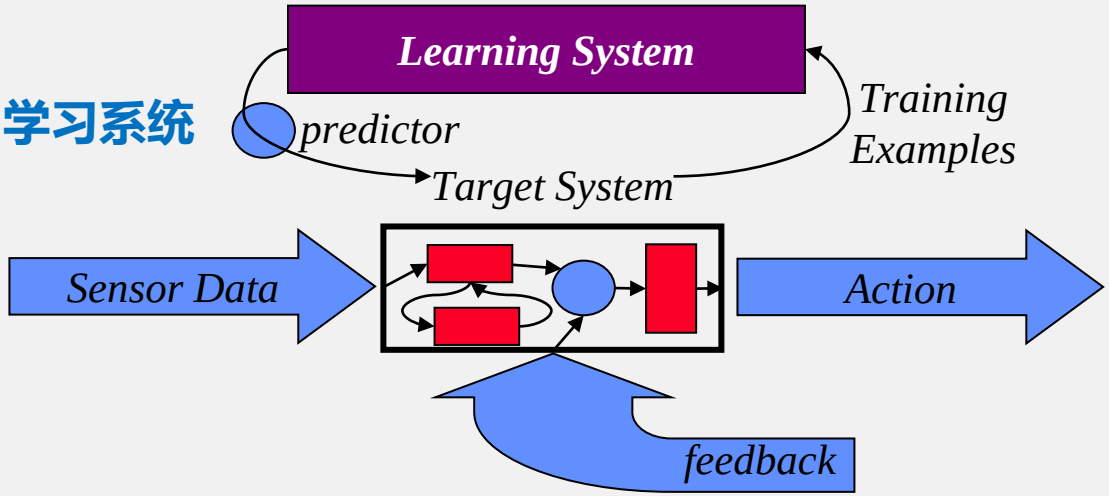
机器学习流程

学习流程





学习系统



机器学习类型

按照学习策略从简单到复杂的次序分为六种基本类型：

1）**机械学习**(Rote learning)

2）**示教学习**(Learning from instruction)

3）**演绎学习**(Learning by deduction)

4）**类比学习**(Learning by analogy)

5）**基于解释的学习**(Explanation-based learning)

6）**归纳学习**(Learning from induction)

机器学习方法

- **有监督/无监督学习**

  有监督(Supervised)：分类、回归

  无监督(Unsupervised)：概率密度估计、聚类、降维

  半监督(Semi-supervised)：EM、Co-training

- **其他学习方法**

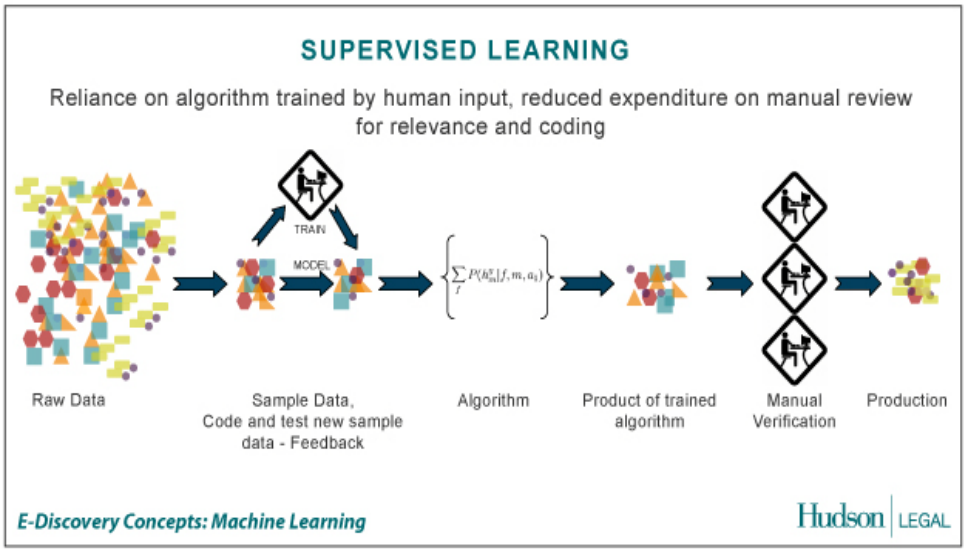
  增强学习(Reinforcement Learning)

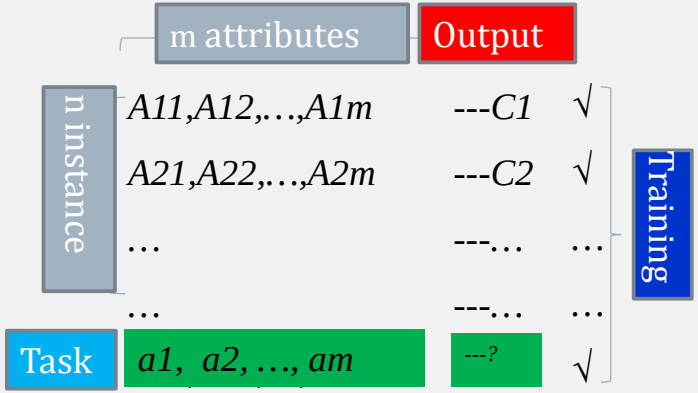
  多任务学习(Multi-task learning)

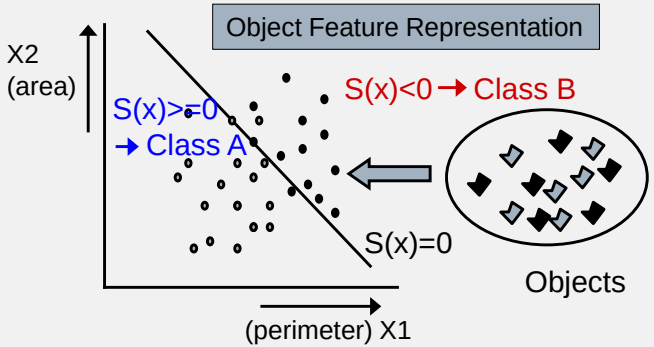
  主动学习（Active learning)

机器学习方法 - *有监督学习*

- 训练数据有正确输出
- 使用训练数据，准备算法，并根据目标输出与实际输出的误差信号来调节参数
- 典型方法
  - 全局：BN, NN,SVM, Decision Tree
  - 局部：KNN、CBR(Case-base reasoning)



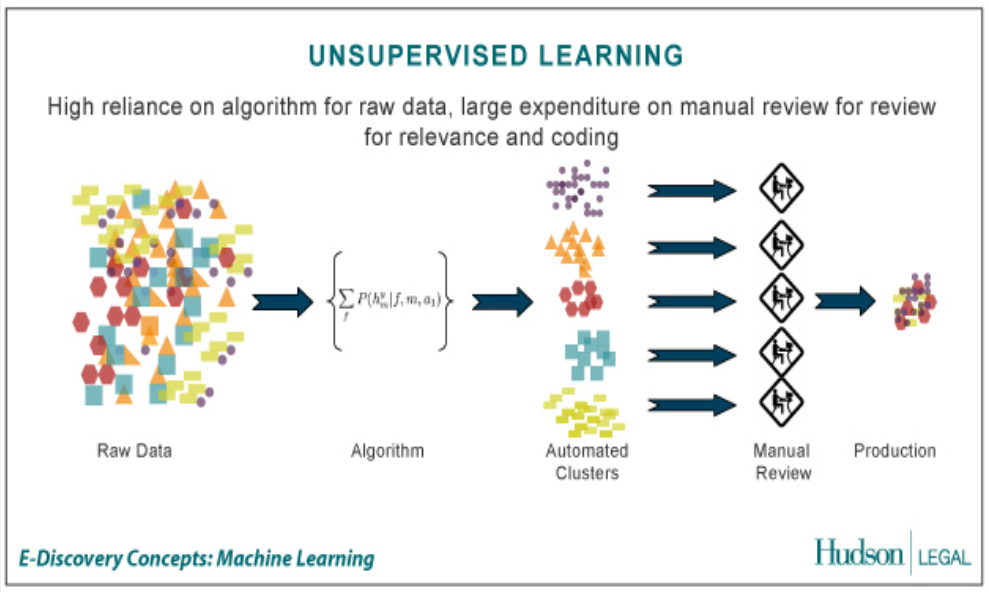




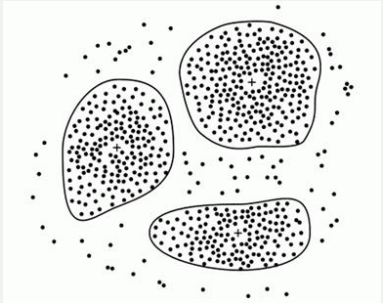
机器学习方法 - *无监督学习*

- 训练数据的分类正确性未知
- 学习机根据外部数据的统计规律（如Cohension&divergence）来调节系统参数，以使输出能反映数据的某种特性
- 典型方法

  K-means、SOM….



	m attributes	Output
n instance	*A11,A12,…,A1m*	*---C1* X
	A21,A22,…,A2m	*---C2* X
	…	*---…* …
	…	*---…* …
	An1,An2,…,Anm	Task X

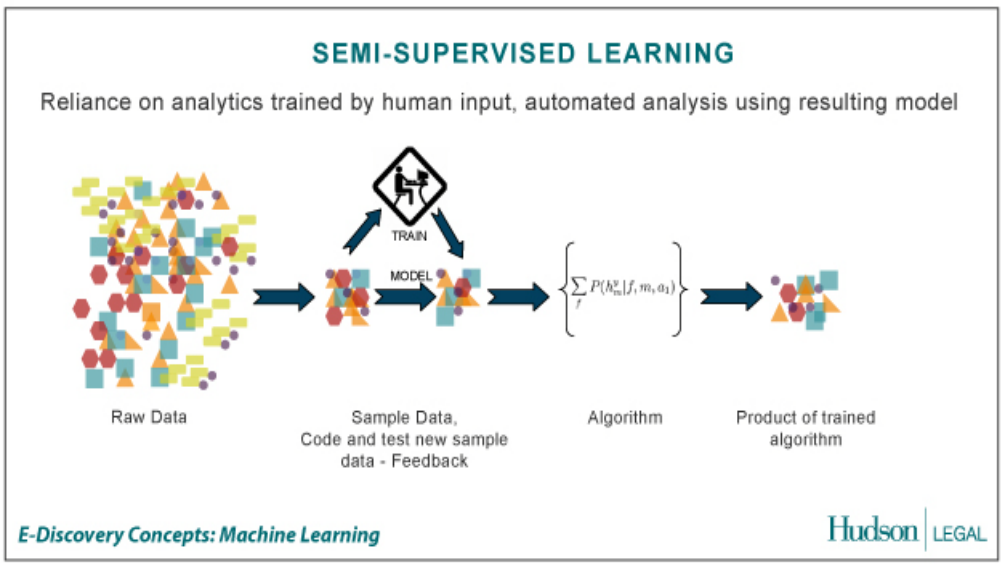


示例：聚类

机器学习方法 - *半监督学习*

- 结合（少量的）有确定输出训练数据和（大量的）未确定输出训练数据来进行学习
- 典型方法

  Co-training、EM、Latent variables….

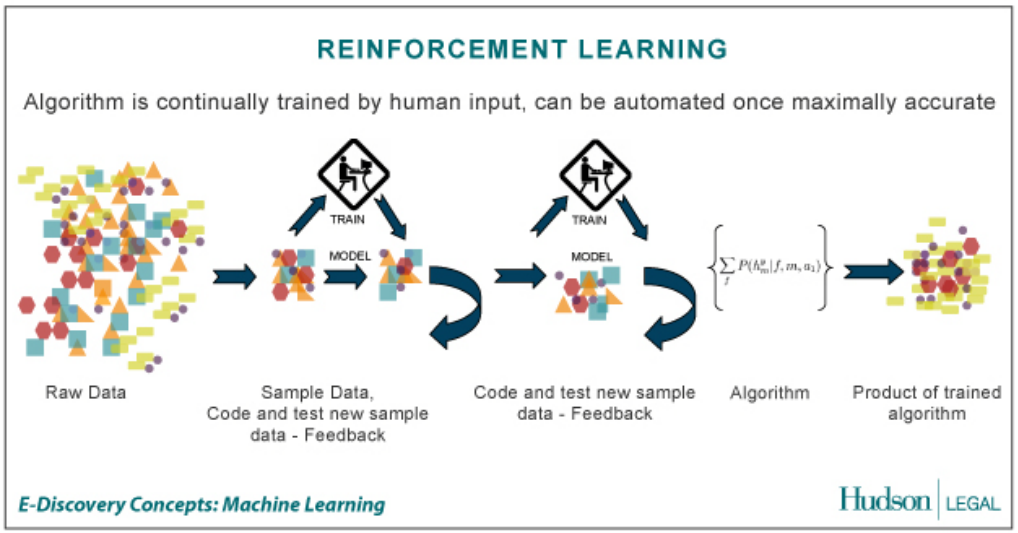


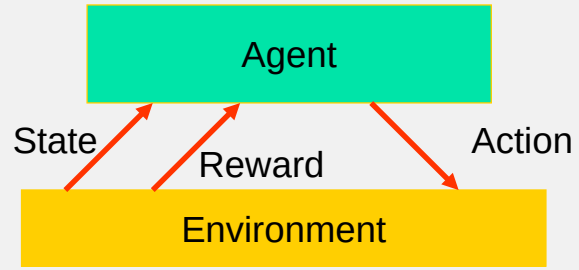
监督学习与未监督学习结合体

	m attributes	Output	
n instance	*A11,A12,…,A1m*	*---C1*	√
	A21,A22,…,A2m	*---?*	*X*
	…	*---…*	…
	…	*---…*	…
	An1,An2,…,Anm	*---Cn*	√
Task	*a1, a2, …, am*	*---?*	

机器学习方法 - *增强学习*

- 外部环境对输出只给出评价信息而非正确答案，学习机通过强化受奖励的动作来改善自身的性能
- 训练数据包含部分学习目标信息





$s_0 \xrightarrow[r_0]{a_0} s_1 \xrightarrow[r1]{a1} s2 \xrightarrow[r2]{a2} s3$

- S– set of states
- A– set of actions
- T(s,a,s') = P(s'|s,a)– the probability of transition from *s* to *s'* given action *a*
- R(s,a)– the expected reward for taking action *a* in state *s*

$$R(s,a) = \sum_{s'} P(s'|s,a) r(s,a,s')$$

$$R(s,a) = \sum_{s'} T(s,a,s') r(s,a,s')$$

机器学习的输出

- Classification（分类）

  如果你的问题能够以 Yes/No 来回答 => Classification (分类)

- Regression（回归分析）

  如果您期望的解答是一個數值的話 => Regression(回归分析)

- Clustering (分群)

  如果你想将具相同特性的数据群集分类=>Clustering (分群)

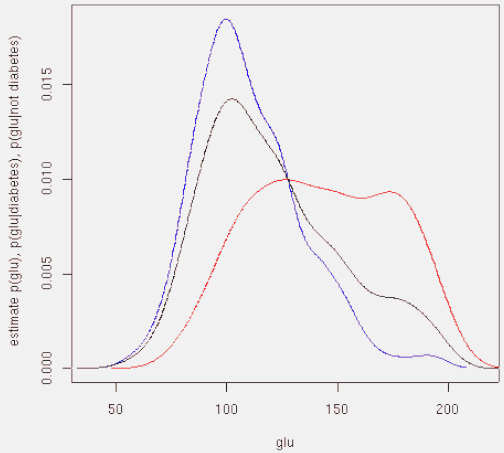
- Density Estimation

  如果你想知道某个问题的概率分布 => Density Estimation

  基于概率统计论，可用于异常检测

- Dimensionality Reduction

  如果你想降低输入数据的维数 => Dimensionality Reduction

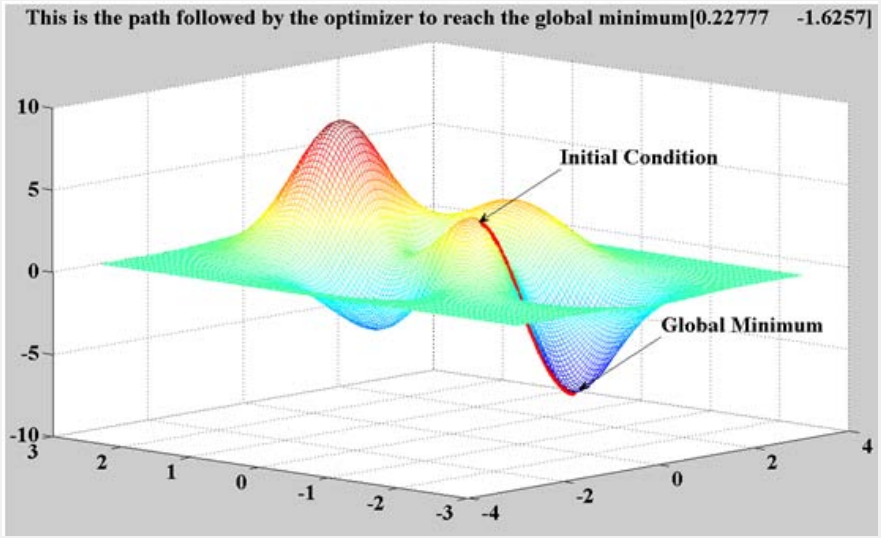


Estimated density of
p (glu | diabetes=1) (red),p (glu | diabetes=0) (blue), and p (glu) (black)

目录

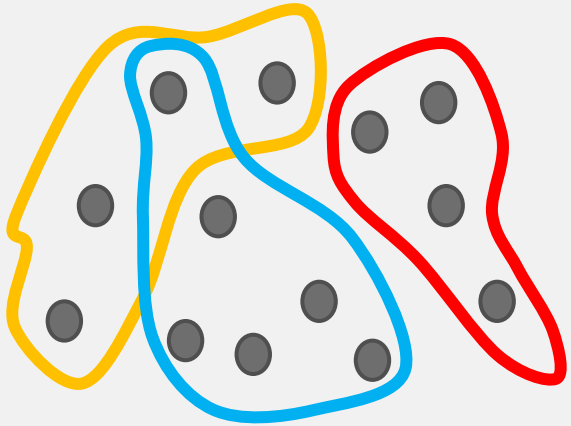
机器学习算法类型

- Clustering
- Association learning
- Parameter estimation
- Recommendation engines
- Classification
- Similarity matching
- Neural networks
- Bayesian networks
- Genetic algorithms



Clustering方法

- 按照预先定义的特性，对相似的数据进行**分组**，但没有任何标记。
- Clustering属于非监督学习。

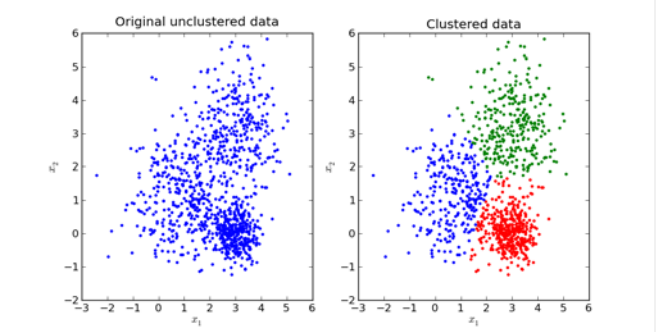


Clustering能将有相同特征者聚集在一起，通常用来处理没有正确答案的问题。

比如腾讯微信辨别使用者不同的群体（运动爱好者、自拍爱好者、美食爱好者等），以实现精准广告。

典型算法：

k-means、... ...



Classification方法

按照预先定义的类型，对数据进行分类。

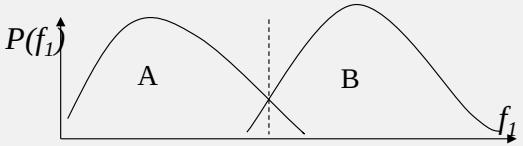
给定：m个类，训练样本和未知数据

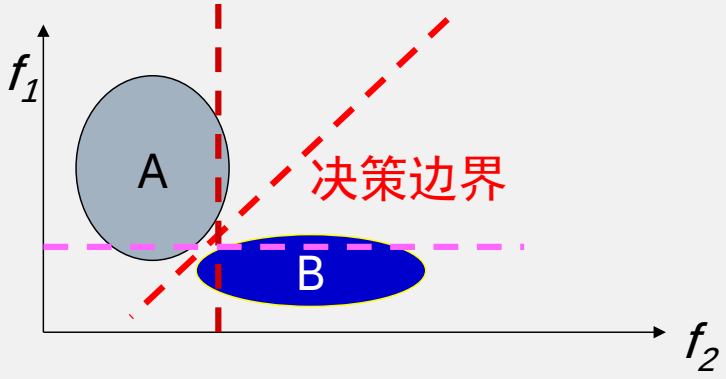
目标：给每个输入数据标记一个类属性

两个阶段：

建模/学习：基于训练样本学习分类规则.

分类/测试：对输入数据应用分类规则





应用：

信用分类

目标市场

医学诊断

欺诈检测

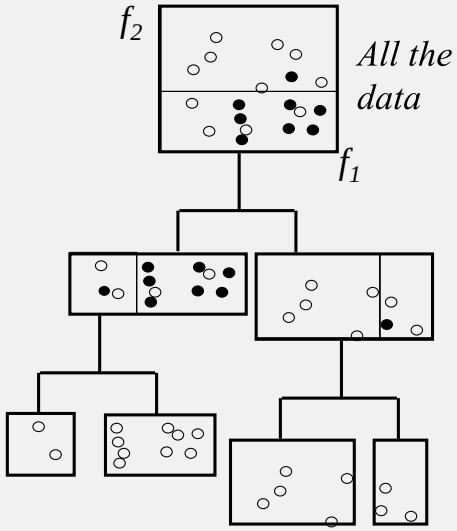
……

Decision Tree方法

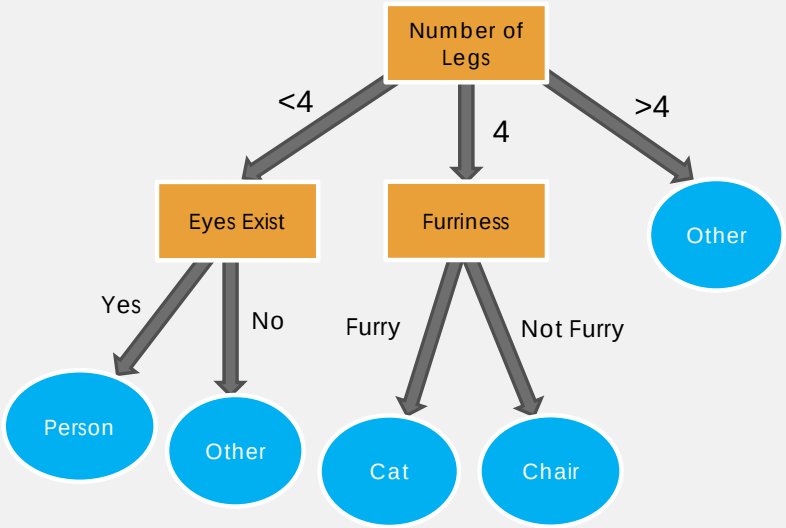
- 决策树算法是根据数据的值和属性建立决策模型。
- 对于一条记录，建立树状结构。
- 使用数据，对决策树进行训练，用于分类和回归问题。

典型算法：

CART
ID3
C4.5
MARS
… …

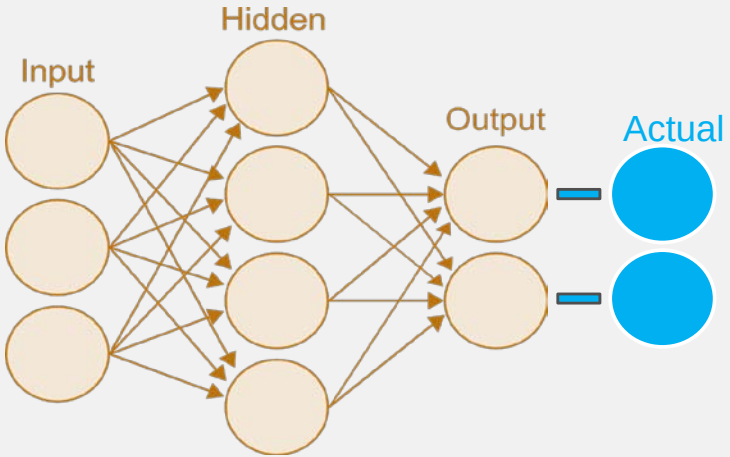


At each step, choose the feature that "reduces entropy" most. Work towards "node purity".



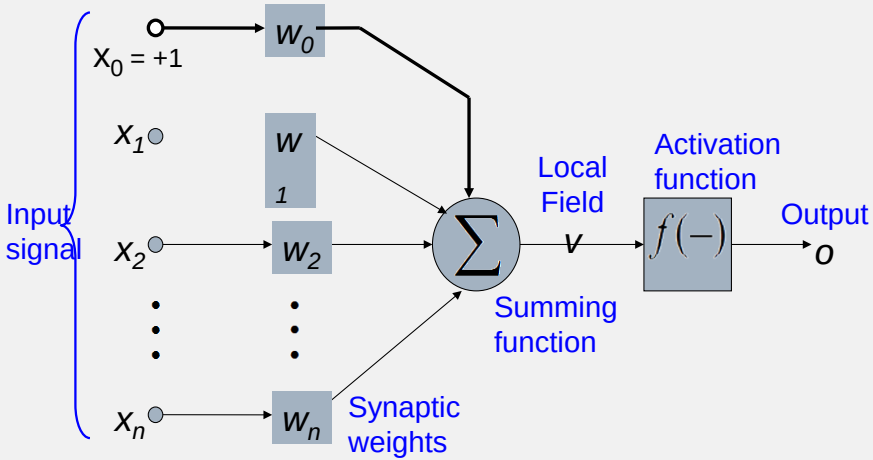
Neural networks方法

神经网络(Neural Networks)：模拟人脑的学习。

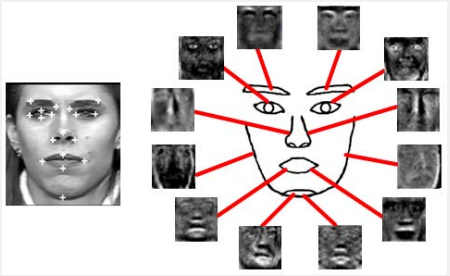


人工神经元模拟生物神经元的一阶特性。

- 输入： X=(x1, x2, …, xn)
- 联接权： W=(w1, w2, …,wn)T
- 网络输入： net=∑xiwi
- 向量形式： net=XW
- 激活函数： f
- 网络输出： o=f(net)

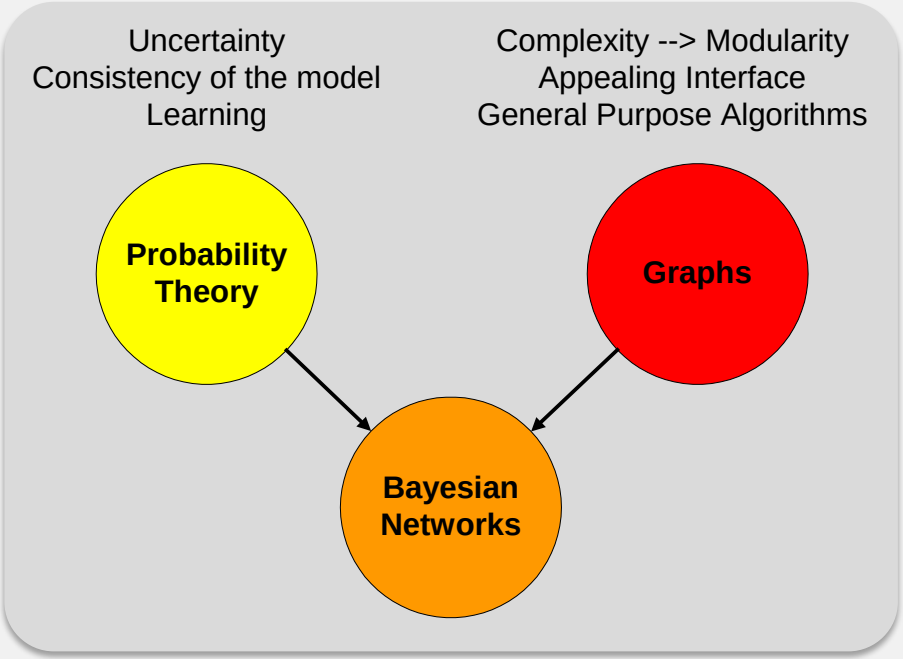


示例：特征检测



Bayesian networks方法

贝叶斯网络（ Bayesian networks ）是表示变量间概率依赖关系的有向无环图。

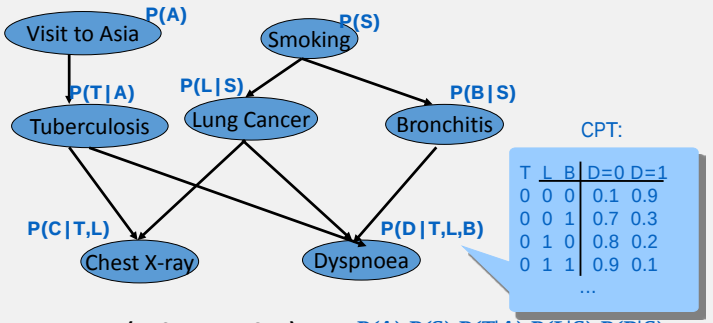


$$P(h_i \mid \mathbf{D}) = \frac{P(\mathbf{D}|h_i)P(h_i)}{P(\mathbf{D})} = \alpha P(\mathbf{D} \mid h_i)P(h_i)$$

$$P(x \mid \mathbf{D}) = \sum_i P(x \mid h_i)P(h_i \mid \mathbf{D})$$

贝叶斯网络作为一种**不确定性的因果推理模型**，主要**用于概率推理及决策**。

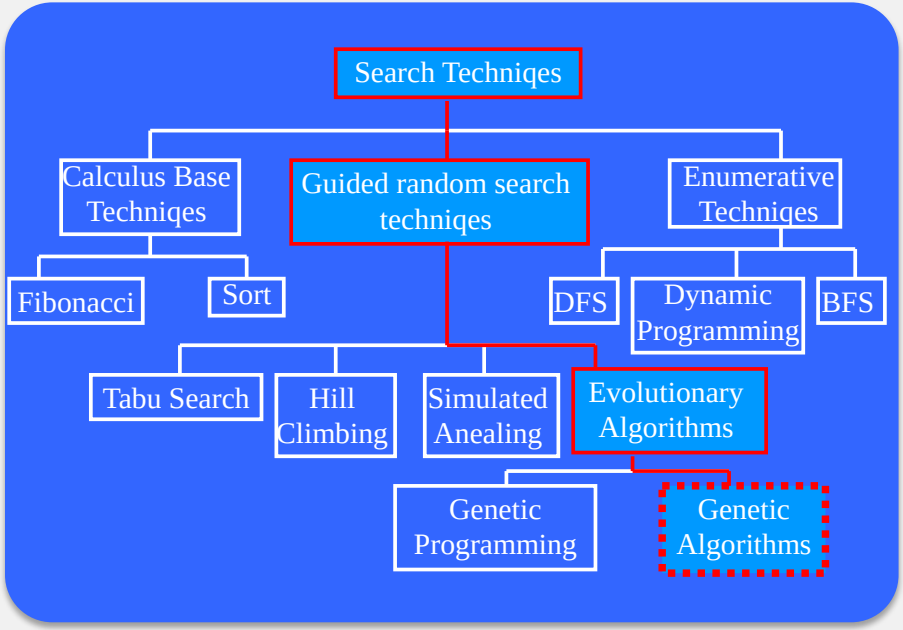
应用：医疗诊断、信息检索、电子技术与工业工程等。



P(A, S, T, L, B, C, D) = *P(A) P(S) P(T|A) P(L|S) P(B|S) P(C|T,L) P(D|T,L,B)*

Genetic Algorithms方法

- **遗传算法**（Genetic Algorithms）是一种**大致基于模拟进化**的学习方法。
- 不再是从一般到特殊或从简单到复杂地搜索假设，而是通过变异和重组当前已知的最好假设来生成后续的假设。
- 遗传算法研究的问题是搜索候选假设空间并确定最佳假设(最佳假设定义为使适应度最优的假设)。



遗传算法的广泛应用

- ✓ 电路布线
- ✓ 任务调度
- ✓ 函数逼近
- ✓ 选取人工神经网络的拓扑结构

机器学习TOP 10算法

1. C4.5
2. k-means clustering
3. Support vector machines(SVM)
4. Apriori
5. EM(Expectation Maximization)
6. PageRank
7. AdaBoost
8. k-Nearest Neighbours(kNN)
9. Naive Bayes
10. CART(Classification and Regression Tree)

机器学习算法特点

- Linear algebra
- Calculus
- Probability theory
- Graph theory
- …

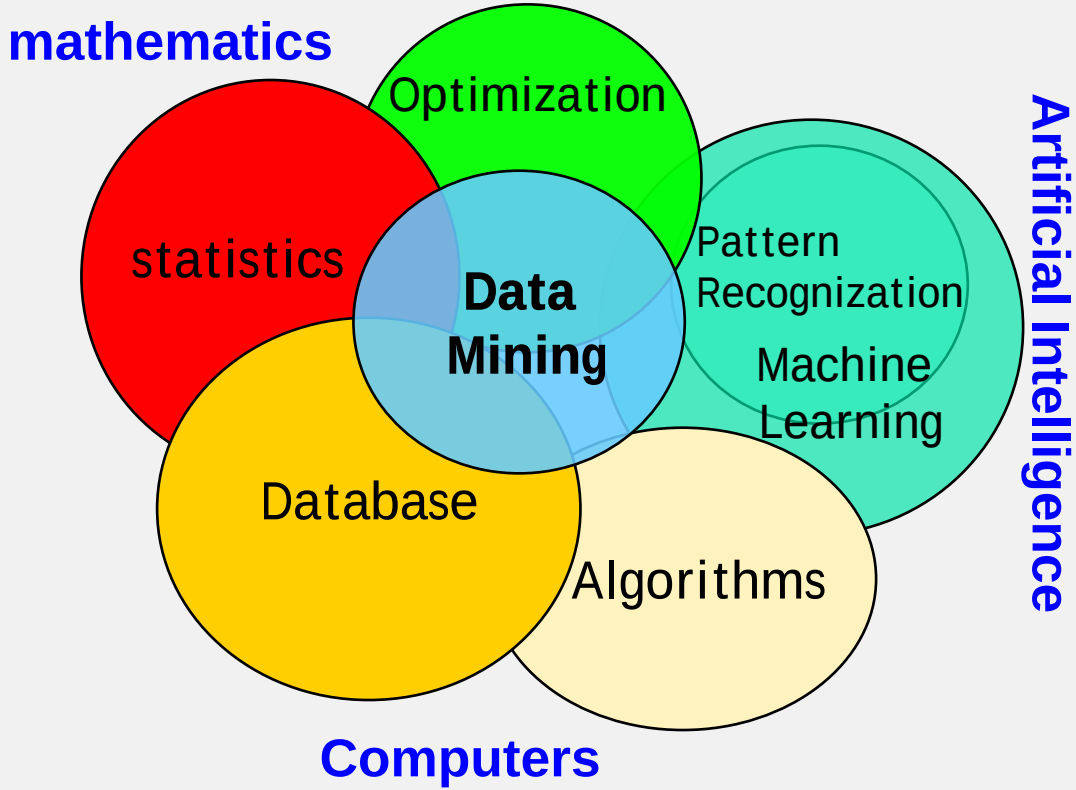
Basically, it's all maths…

Only 10% in devops are know how of work with Big Data. Only 1% are realize they are need 2 Big Data for fault tolerance

https://twitter.com/devops_borat

目录

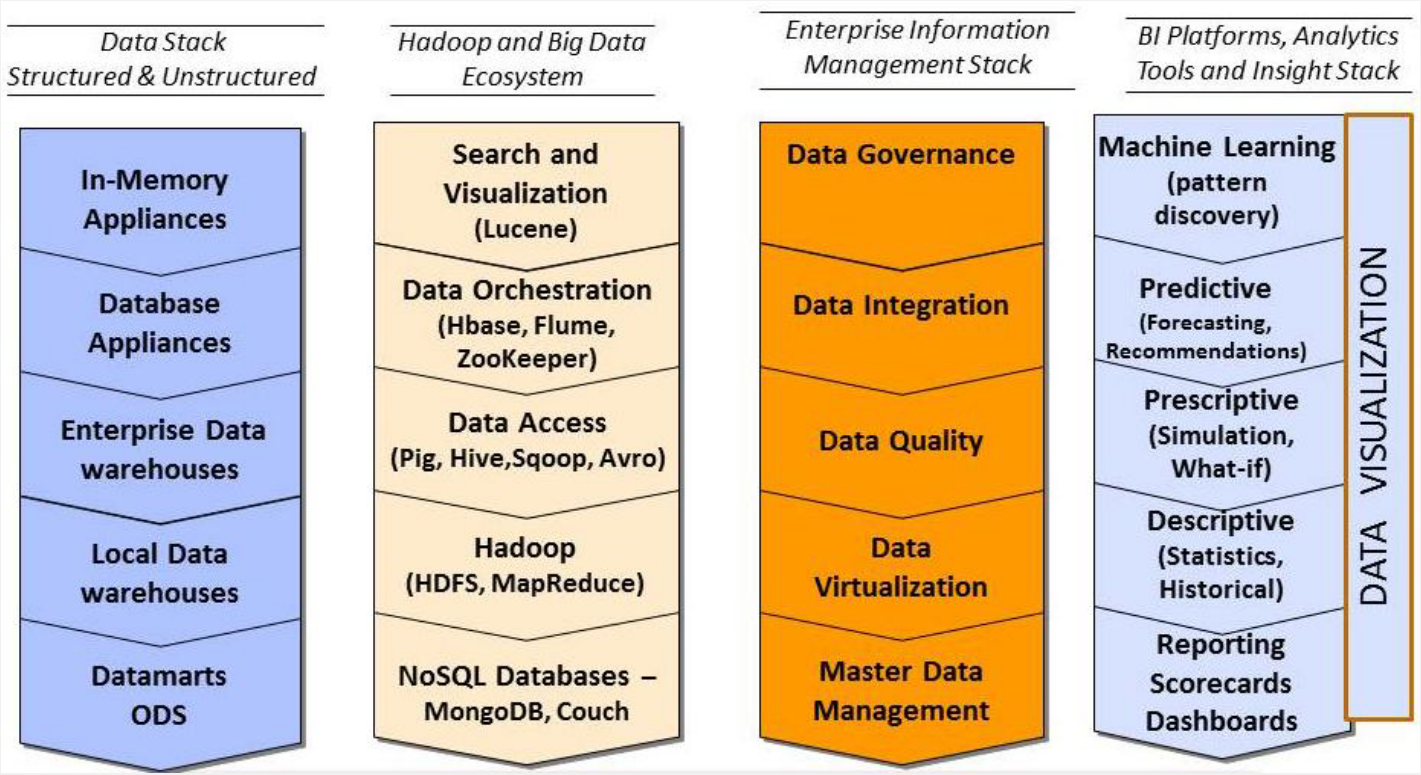
机器学习与相关学科关系



机器学习与相关学科

机器学习	数据挖掘	人工智能	统计学
Machine Learning: **根据已知数据，计算出一个假设值g**，使其尽可能接近真实值f $g \approx f$	Data Mining: **从大量的数据中找出有价值信息**。 1、若有价值的信息就是“假设值接近真实值”，那么ML=DM（KDDCup的工作就是如此）。 2、若有价值信息和“假设值接近真实值”有关，那么DM可以帮助ML，反之亦然。 3、传统的DM聚焦在高效计算大型数据库。	Artificial Intelligence：**计算的信息可以显示出智能行为。** g ≈ f通常情况下也是一种智能行为。ML可以实现AI。 *如下棋：* *传统AI：游戏树* *ML for AI：从实际游戏数据中学习。* ML是实现AI的一种可能途径。	Statistics：使用数据，**对一个未知的过程做出推断**。收集、处理、分析、解释数据并**从数据中得出结论**。 g就是一个推测结果，f的值通常是未知的。可以采用统计学方法实现ML。 传统的统计学关注可证明的数学假设，而不关注计算。 **统计学是ML的有用工具。**

机器学习与大数据关系



Machine learning in Big Data Infrastructure

机器学习与大数据比较

What	Why	How
Relational Data Warehouse	Data integrity, structure, fast, well-known, governance, fixed schemas	ETL, BIML, Index
Hadoop & HDInsight	Unstructured data, large volumes of text, flexible schemas	Hbase, Map Reduce, HDFS
Tabular	Fast analytics, agility, preserves types	In-memory
Multidimensional OLAP	Fast analytics, large data volumes	Preaggregated calculations
Data Mining & Machine Learning	Complex analytics, discovery, predictive models, forecasting	Estimations

业界主要工具与服务

业界流行框架及工具

- Weka
- Carrot2
- Gate
- OpenNLP
- LingPipe
- Stanford NLP
- Mallet – Topic Modelling
- Gensim – Topic Modelling (Python)
- Apache Mahout
- MLib – Apache Spark
- scikit-learn - Python
- LIBSVM : Support Vector Machines
- and many more…

业界主要服务提供商

- ✓ Google Prediction API
- ✓ Azure ML
- ✓ NeuroMine
- ✓ BigML
- ✓ Wise.io
- ✓ Algorithms.io
- ✓ Infer.com

谢谢！