南京大学
NANJING UNIVERSITY

机器学习导学

南京大学软件学院
任桐炜 李传艺
{rentw,lcy}@nju.edu.cn

什么是机器学习？

- T. Mitchell (米切尔)
  - Carnegie Mellon University Machine Learning
  - – Any computer algorithm that lets the system perform a task more effectively or more efficiently than before.

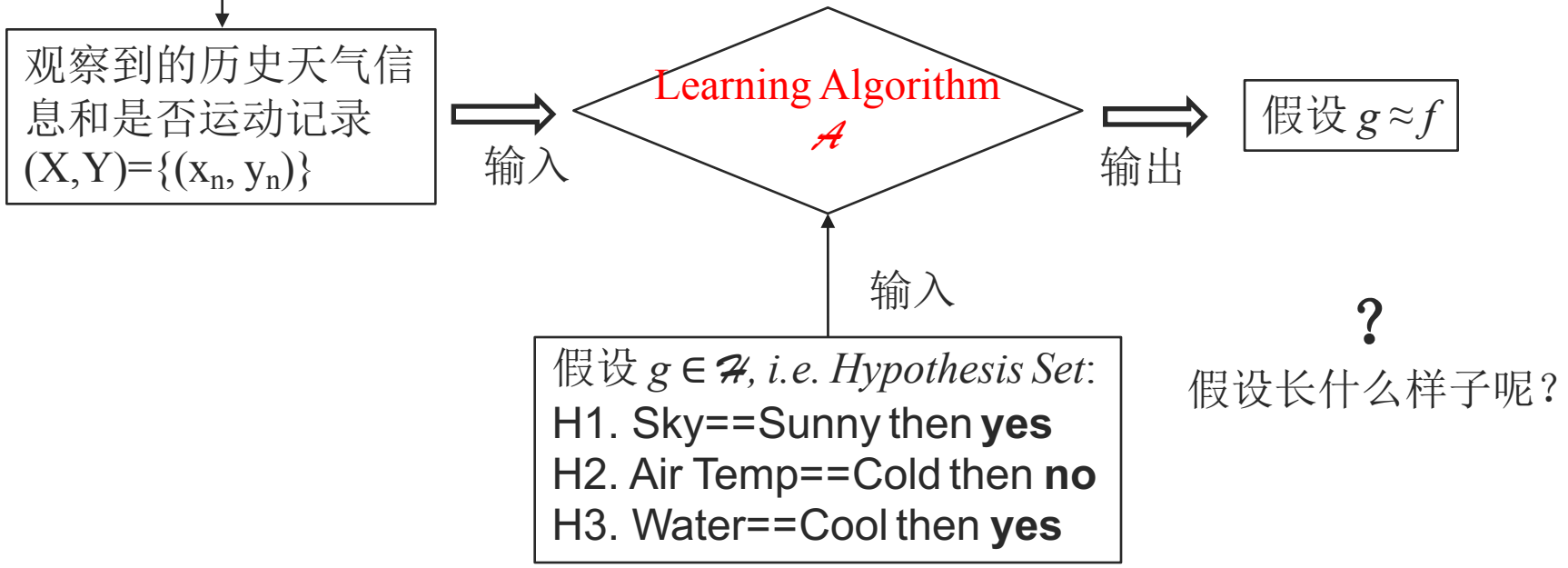
- H. Simon (西蒙)
  - Professor of Computer Science, Carnegie Mellon
  - – Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the task or tasks drawn from the same population more efficiently and more effectively the next time.
- **The ability to perform a task in a situation which has never been encountered before (Learning = Generalization)**

今天适合运动吗？

- 告诉你一些关于当前天气的状况，让你判断一下今天是否适合运动　　x = getInput()　　compute(x) → y
- 人：老夫夜观天象，见紫微星东移，掐指一算，得之……
- 计算机？
  - 给一些例子作为参考
  - 例子：ID+天气相关信息+是否适合运动的标记
  - 目标：根据新的天气信息（x）计算出标记（y） $\Longrightarrow$ f: x → y 目标函数，最理想的情况
  - f是老天爷（上帝）的杰作，人能得到的只是一个$g \approx f$　　Machine Learning
    假设hypothesis

Example	Sky	Air Temp	Humidity	Wind	Water	Forecast	Enjoy Sport
0	Sunny	Warm	Normal	Strong	Warm	Same	**Yes**
1	Sunny	Warm	High	Strong	Warm	Same	**Yes**
2	Rainy	Cold	High	Strong	Warm	Change	**No**
3	Sunny	Warm	High	Strong	Cool	Change	**Yes**

Example	Sky	Air Temp	Humidity	Wind	Water	Forecast	Enjoy Sport
0	Sunny	Warm	Normal	Strong	Warm	Same	**Yes**
1	Sunny	Warm	High	Strong	Warm	Same	**Yes**
2	Rainy	Cold	High	Strong	Warm	Change	**No**
3	Sunny	Warm	High	Strong	Cool	Change	**Yes**



以观察到的数据为样本从假设空间中选择一个与目标函数最像的假设

$\mathcal{A}[(X,Y)] \rightarrow g \in \mathcal{H}\ (g \approx f)$

学习算法的评估

- 通过评估输出的g判断学习算法是否优秀
  - 判断g和f的接近程度
  - 切记：
    - f是不可知的
    - 只有观察到的数据（如天气信息和是否运动标记）
  - 通过更多的观察数据判断g和f的接近程度
    - 数据由f产生，一定程度上代表了f

- 训练集
- 测试集　g：假设类型+表达假设的参数W
- 验证集　$\mathcal{A}$：使用到一些配置参数，记为C
- 交叉验证-cross validation

此处应该有一例子

机器学习和相关领域

- 基础知识
  - 线性代数
  - 统计、概率论
- 领域
  - 生物学
  - 心理学
  - 营养学
  - 语言学
- 应用
  - 人工智能：自然语言处理、计算机视觉、智能决策支持、高性能计算等
  - 计算机科学：编译器、软件系统、数据库等

机器学习和数据挖掘、人工智能

- 机器学习：$\mathcal{A}[(X,Y)] \rightarrow g \in \mathcal{H}\ (g \approx f)$

- 数据挖掘：使用数据发现有趣、有用的性质——相互交融，难以分割
  - 如果有趣、有用的性质就是g
  - 如果有趣、有用的性质能够帮助构造更好的g
  - 如果g能够帮助发现更有趣、有用的性质

- 人工智能：让机器智能地完成某些任务
  - 如果机器能够学到有用的g，就能够智能地完成某些任务
  - 还有其它方式让机器智能地完成任务

——机器学习只是实现人工智能的其中一种途径

机器学习的发展（人工智能）【摘自西瓜书】

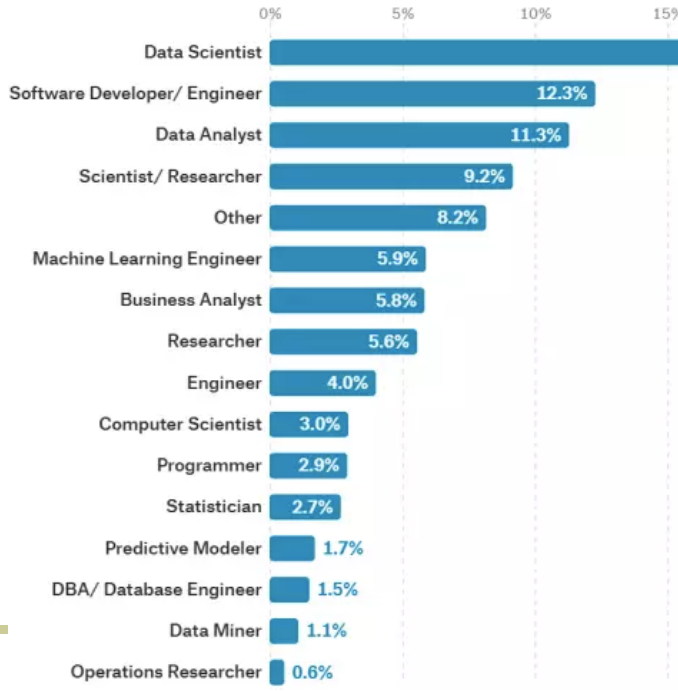
- 推理期（1950-1970）
  - 逻辑理论家：赋予机器逻辑推理的能力
  - 只有推理能力远远不能实现智能
- 知识期（1975-）
  - 专家系统-基于知识库：要让机器拥有知识→知识工程
  - 人工总结知识交给机器太难
- 学习期：让机器自己学习知识
  - 1980机器学习独立学科：从样例中学习、问题求解和规划中学习等
  - 符号主义（蓬勃发展1960-1970）
    - 决策树、基于逻辑的学习（1980流行）
  - 连接主义（1950-）
    - 神经网络→深度学习（CNN，RNN）（2000卷土重来）
  - 统计学习（1960开始有基础，1990闪亮登场）
    - SVM

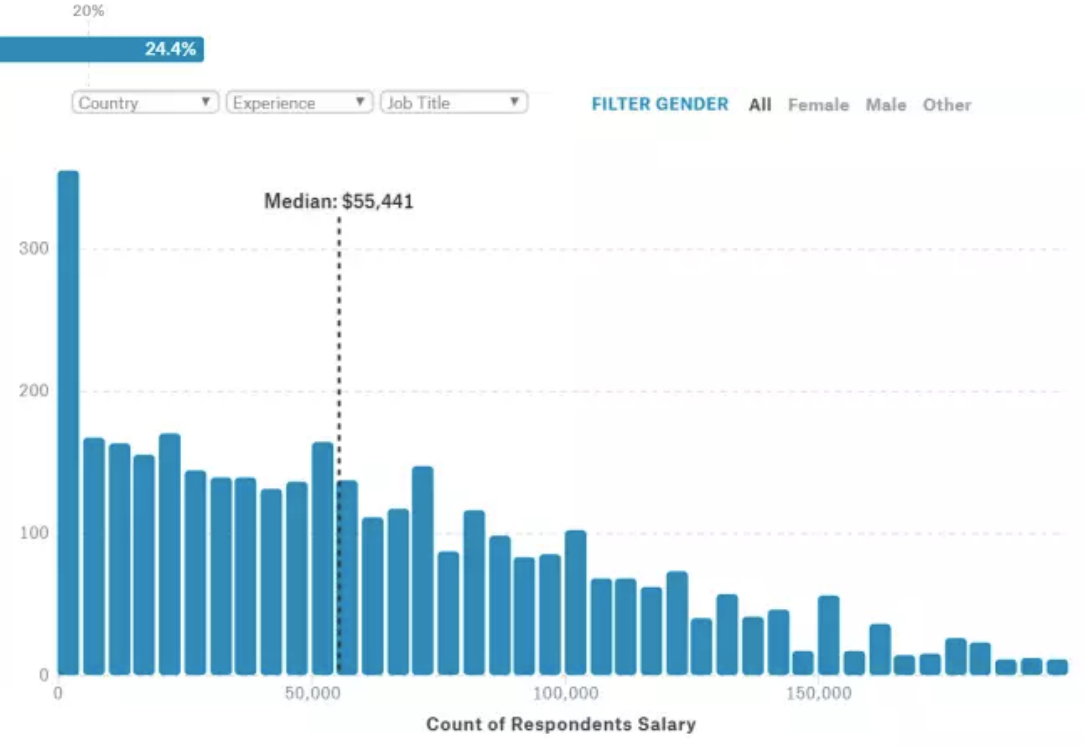
为什么要学机器学习

- 应用广泛
- 好找工作
- 驱动自动化、智能化
- 时机对了
  - 数据量不断变大、计算能力不断提高，使得机器学习成为可能

Kaggle首份机器学习大调查（1）

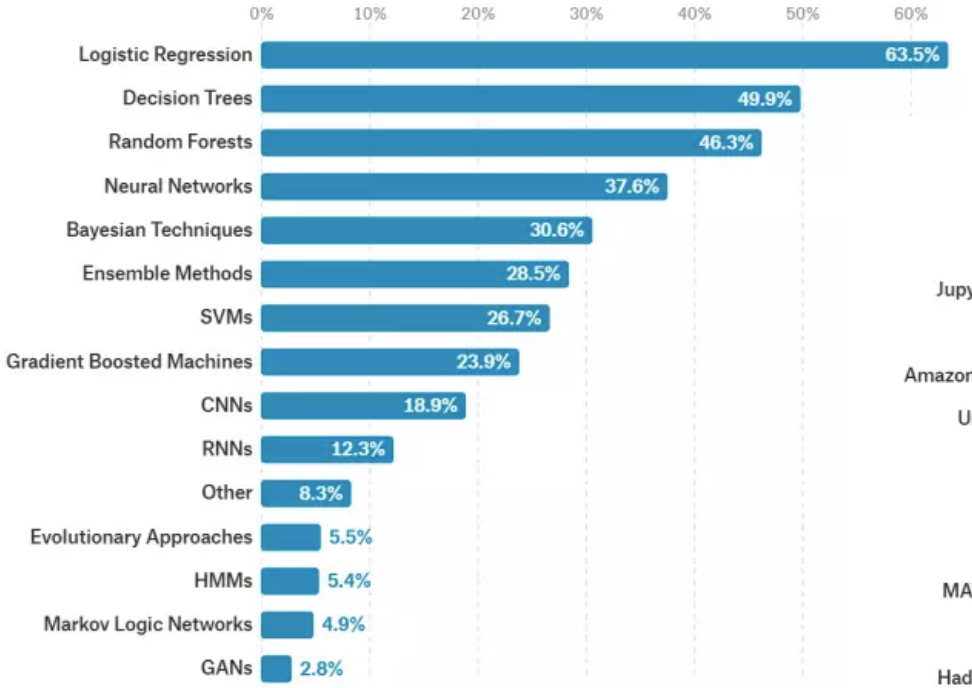
- Kaggle: 为开发商和数据科学家提供举办机器学习竞赛、托管数据库、编写和分享代码的平台
- 数据科学家：使用代码分析数据的人
  - 职业头衔和收入

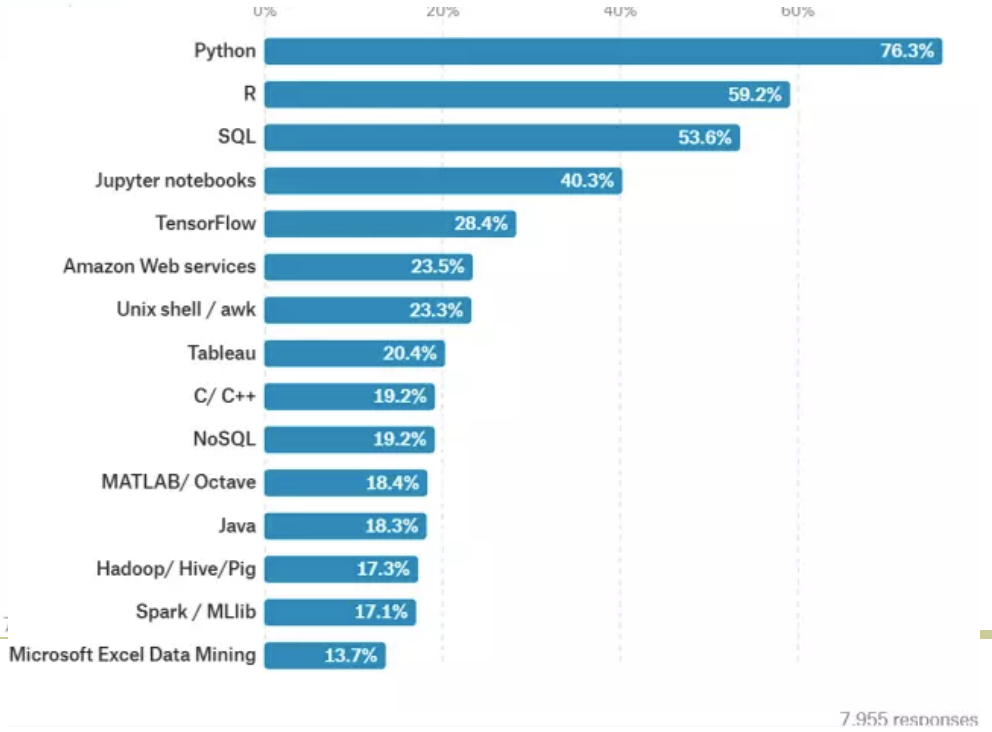




Kaggle首份机器学习大调查（2）

- 常用的数据分析方法
- 常用的语言和工具





机器学习应用的场景示例

- 自然语言理解
  - 歧义判断：我们/三个人一组；我们三个人/一组；这份报告，我写不好
  - 指代消解：小张讨厌小王，（因为/所以）他打了他
- 视频图像处理
  - 文字识别（印刷、手写等）
  - 人脸识别
- 生物计算
  - 蛋白质编码
- 医药学
  - 疾病预测
- 社会学、天文学、气象学
  - 犯罪预测：【洛杉矶警擦局】“利用之前犯罪行为表现出来的规律，全神贯注地分析下一个可能发生犯罪行为的地点”
  - 星体发现、气象预测
- ……

推荐读物

- *机器学习*
  周志华，清华大学出版社，2016

- *课程《机器学习基石》*
  台大林轩田教授

- *Artificial Intelligence: A Modern Approach* (2nd/3rd edition)
  Russell and Norvig, Prentice-Hall, Inc., 2003/2010.

- *Machine Learning*
  Tom M. Mitchell, McGraw Hill, 1997

学习方法的类别——按输入数据分

$\mathcal{A}[(X,Y)] \rightarrow g \in \mathcal{H}\ (g \approx f)$

- 监督学习
  - 输入：数据样例=特征值(s)+标签
  - 分类问题
    - 输出是非连续的有限集合
  - 回归问题
    - 输出是连续的实数集合
  - Decision trees, neural networks, nearest-neighbor algorithms, Bayesian learning, hidden Markov models
- 无监督学习
  - 输入：数据样例=特征值
  - 聚类问题
    - 将以某种规则相似的数据样例集中到一起
- 半监督学习
- 强化学习
  - Markov Decision Processes, temporal difference learning, Q-learning

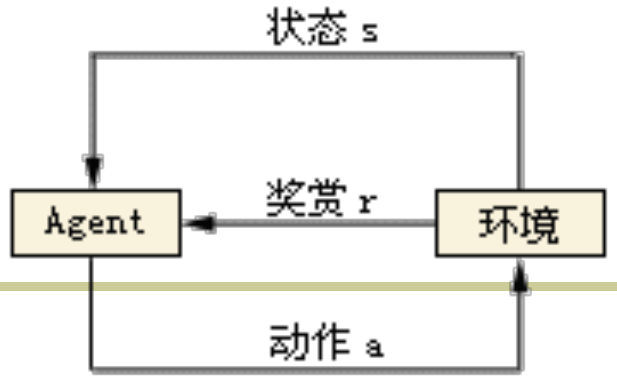
强化学习又是什么？

- Alpha Go - Alpha Zero
- Reinforcement Learning

•环境（标准的为静态stationary，对应的non-stationary）
•agent（与环境交互的对象）
•动作（action space，环境下可行的动作集合，离散or连续）
•反馈（回报，reward，正是有了反馈，RL才能迭代，才会学习到策略链）

•数据是序列的、交互的、并且还是有反馈的

- 方案 (Policy) = 在每个状态下，你会选择哪个行动？



不同学习方法的区别

- https://www.zhihu.com/question/41775291
- 介绍了监督学习和强化学习的不同

Supervised Learning: given **data**, predict **labels**

Unsupervised Learning: given **data**, learn about that **data**

Reinforcement Learning: given **data**, choose **action** to maximize expected **long-term reward**

相关机器学习算法的简要介绍

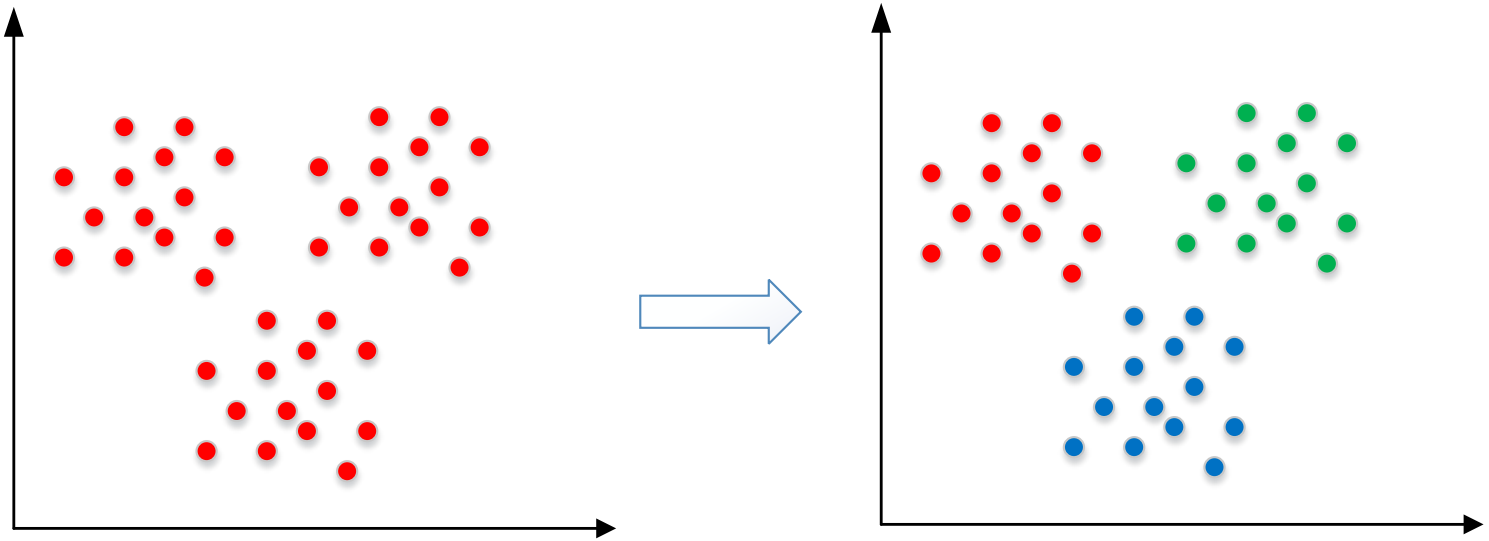
- 聚类算法

- 决策树
- 神经网络
- SVM

- 朴素贝叶斯
- 隐马尔科夫模型

无监督：聚类（1）

- 将没有标签的数据样例分布到由不相交的类簇构成的子集中，达到：
  - 在同一个子集中的两个数据样例是相似的
  - 在不同子集中的两个数据样例是不相似的



  - 相似程度用两个样例之间的距离表示

无监督：聚类（2）

- 距离函数
  - 闵可夫斯基距离(Minkowski distance)

$$L(\vec{x},\vec{y}) = (\sum_{i=1}^{m} |x_i - y_i|^P)^{\frac{1}{p}}$$

  - 欧氏距离(Euclidean distance)

$$L_2(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$$

  - 曼哈顿距离(Manhattan distance)

$$L_1(\vec{x},\vec{y}) = \sum_{i=1}^{m}|x_i - y_i|$$

  - Cosine距离

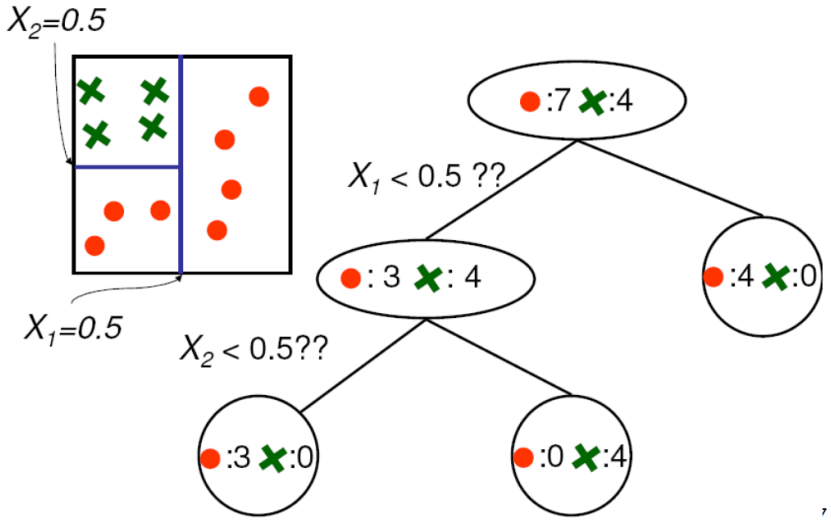
$$\cos(\vec{x},\vec{y}) = \frac{\vec{x}\bullet\vec{y}}{|\vec{x}|\cdot|\vec{y}|}$$

无监督：聚类（3）

- 层次聚类(Hierarchical clustering)
  - 自底向上 or 自顶向下
  - AGNES：自底向上
    - 初始化：把每一个样本当作一个类簇；迭代合并
- 原型聚类(prototype-based clustering)
  - K-means：均值聚类
    - 目标是最小化每一个类簇里平方误差
    - 贪心策略
  - LVQ：学习向量量化
    - 假设数据样本自带类别标记，用这些标记辅助聚类
    - 有点监督学习的意思
  - 高斯混合聚类
    - 假设数据满足高斯分布；用EM算法求这个分布；按照概率划分类簇
- 密度聚类(Density-based clustering)
  - 密度直达、可达、相连概念；找到核心对象；根据核心对象确定对应的类簇；

监督：决策树

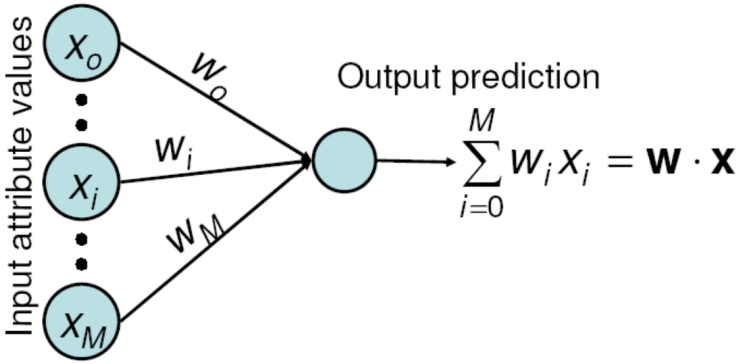
- 监督学习：输入数据带标签
- 构造一个可以用来获取新数据标签的树结构
- 思考：
  - 假设类型？ $\mathcal{A}[(X,Y)] \rightarrow g \in \mathcal{H}\ (g \approx f)$

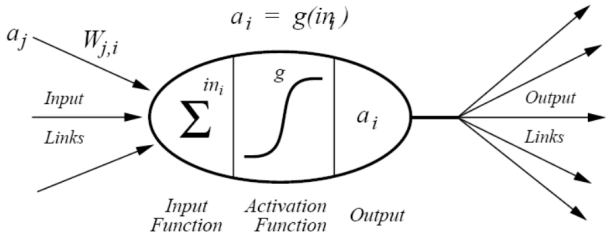


  - 信息熵理论
  - 剪枝
  - 多变量决策树

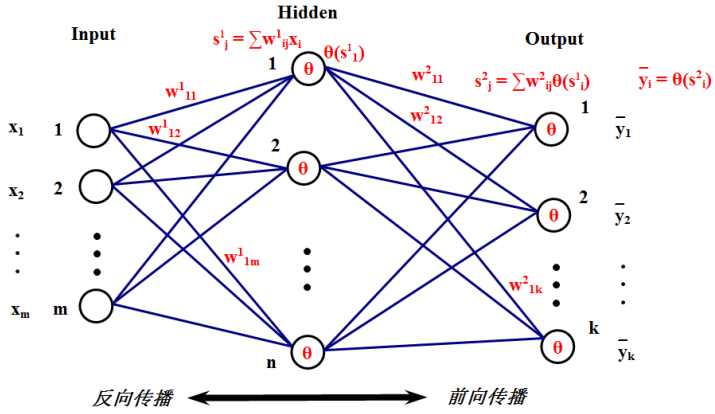
监督：神经网络（1）

- 模拟人脑神经元对信息的传递
- 由神经元链接而成
  - 能够模拟任意函数——假设的类型是什么？ $\mathcal{A}[(X,Y)] \rightarrow g \in \mathcal{H}$ $(g \approx f)$



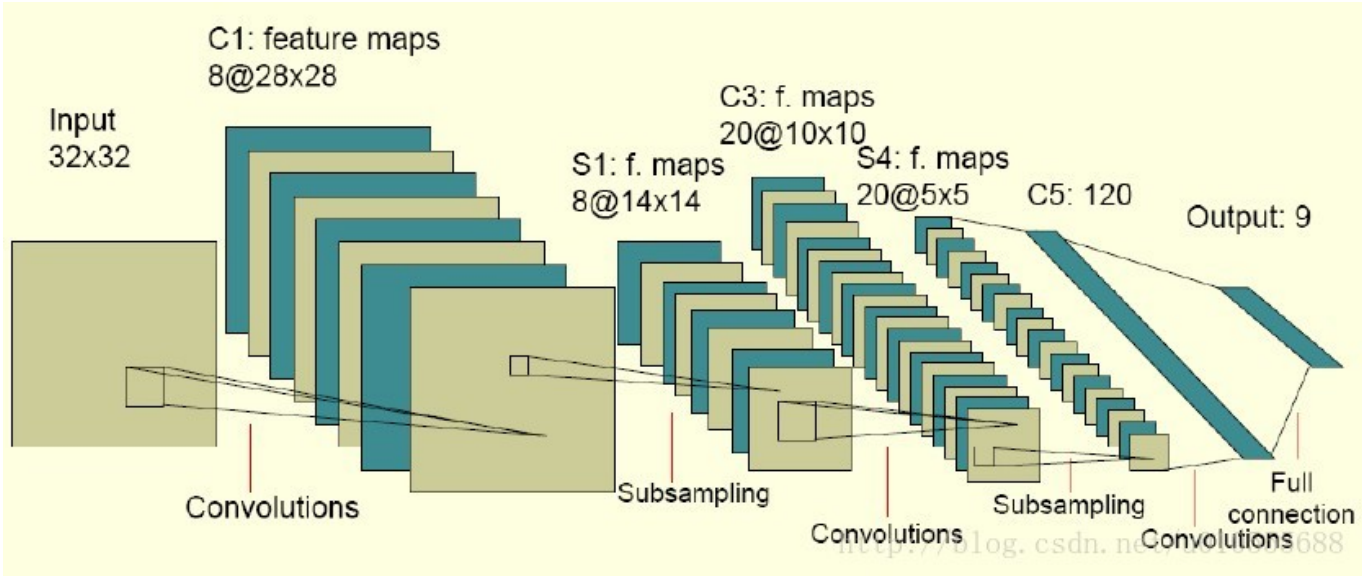


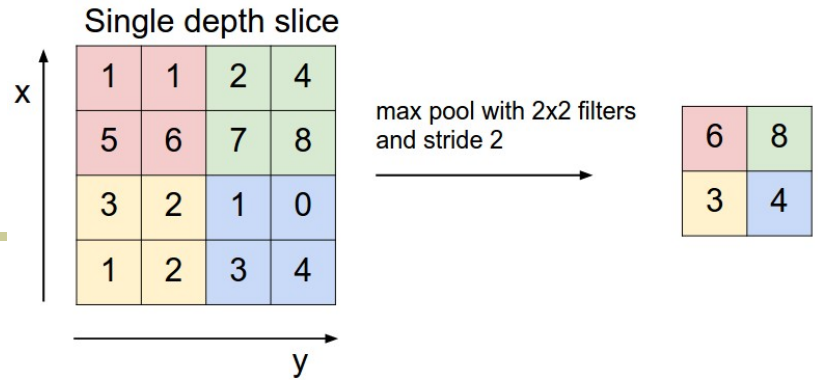
  - 结果对“假设”的一般表达
    - 层数，每层的神经元个数
    - 每两层间的连接变量W
    - 变换函数



监督：神经网络（2）

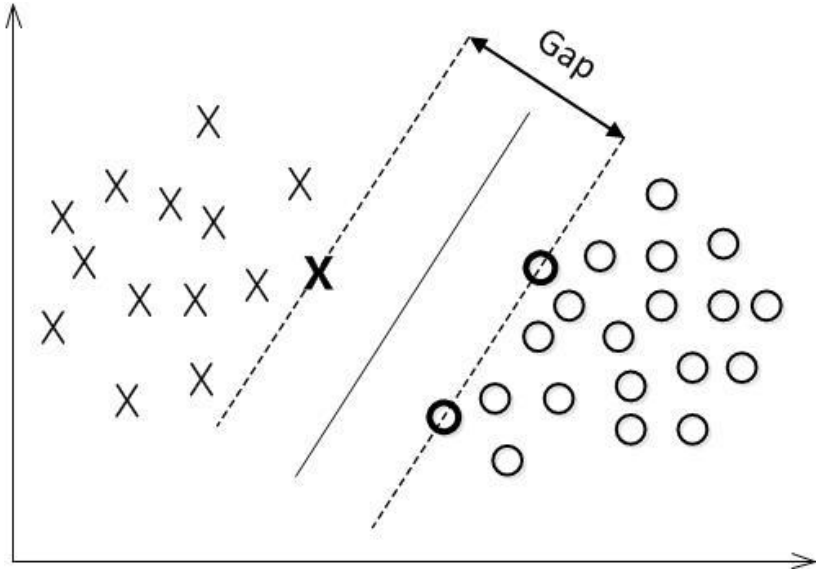
- CNN：卷积神经网络





监督：SVM

- 最大间隔分类器
  - 分类超平面距离数据的间隔越大，分类置信程度越大
- 支持向量
  - 超平面是由距离平面最近的点决定的
  - 也就是图中在虚线上的点
  - 这些点称为“支持向量”点
- 也可以理解为
  - 找到一个平面，使得到平面距离最近的不同类型的点之间最短距离最大



- http://blog.csdn.net/amds123/article/details/53696027

判别模型 vs. 生成模型

- 判别模型
  - 给定(x,y)数据集合，构造计算P(y|x)的模型用于预测y，称为判别模型
  - 前面的“决策树、神经网络、SVM”都是判别模型
- 生成模型
  - 对给定的(x,y)数据集，试图获得联合概率分布P(x,y)，然后计算P(y|x)

$$P(y|x) = \frac{P(x,y)}{P(x)} = \frac{P(y) * P(x|y)}{P(x)}$$

  - 朴素贝叶斯
  - 隐马尔可夫模型

监督：朴素贝叶斯

- 假设样本的各个属性对结果的影响是独立的，所以是Naive的

$$P(y|x) = \frac{P(x,y)}{P(x)} = \frac{P(y) * P(x|y)}{P(x)}$$

- 训练过程
  - 计算$P(y)$和$P(x|y)$
  - $P(x|y)$就是$P(x_1|y)$，$P(x_2|y)$……
  - 根据数据集统计计算出对应的概率即可

监督：隐马尔科夫模型（1）

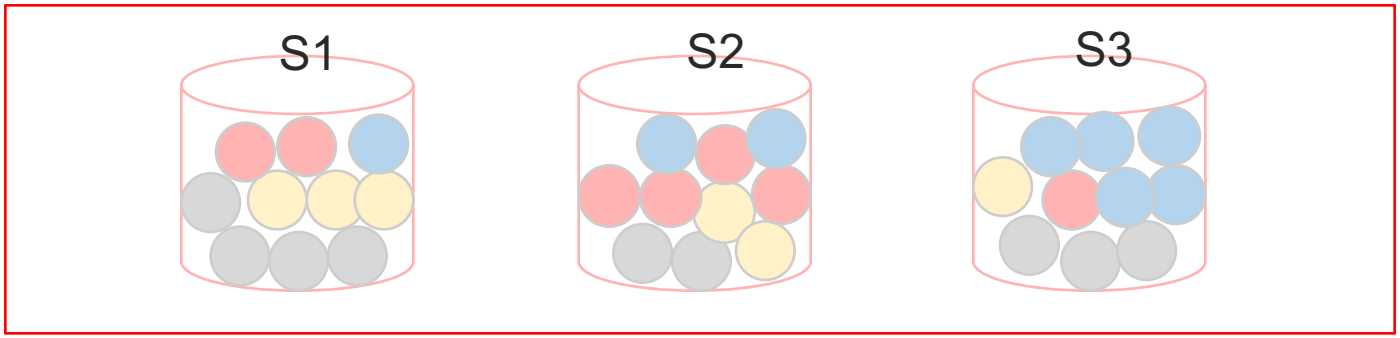
- 马尔可夫系统
  - 给定一个观察到的状态序列，预测下一个出现的状态
- 定义
  - 对有限自动机的扩展：状态集合、状态间转换关系集合（可加权）
  - 有一个有限的状态集合：$S=\{s_1,s_2,\ldots,s_{|S|}\}$
  - 有一个有限的时间序列：$t=\{1, 2, 3, \ldots, T\}$
  - 在每一个时间点，系统处于一个确定的状态$z_t \in \{s_1,s_2,\ldots,s_{|S|}\}$
  - 每个时间点的状态都是随机选择的
  - 当前时刻的状态决定了下一个时间点状态的概率分布
    - 状态转换概率矩阵：$A=\{a_{ij} \mid s_i \rightarrow s_j$转换的概率，$0<i、j<|S|\}$
    - 开始状态概率：$\pi \in R^{|S|}$
- 例子：天气变化模型，晴天、阴天、雨天、多云

监督：隐马尔科夫模型（2）

- 隐马尔科夫模型
  - 给定一个观察到的结果序列，预测对应的状态序列

- 例子：从桶中取球

S:



V:

$\vec{v}$:

监督：隐马尔科夫模型（3）

- 隐藏状态，结果，时间，状态转换概率，状态产生结果概率
  - 有一个有限的不可见的状态集合：$S=\{s_1,s_2,\ldots,s_{|S|}\}$
  - 有一个有限的可见的、由状态产生的结果的集合$V=\{v_1,v_2,\ldots,v_{|V|}\}$
  - 有一个有限的时间序列：$t=\{1, 2, 3, \ldots, T\}$
  - 在每一个时间点，系统处于一个确定的状态$z_t \in \{s_1,s_2,\ldots,s_{|S|}\}$
  - 但是该状态是不可见的，只能看见这个状态产生的结果$v_t \in \{v_1,v_2,\ldots,v_{|V|}\}$
  - 每个时间点的状态都是随机选择的
  - 当前时刻的状态决定了下一个时间点状态的概率分布
    - 状态转换概率矩阵：$A=\{a_{ij} \mid s_i \rightarrow s_j$转换的概率，$0<i、j<|S|\}$
    - 开始状态概率：$\pi \in R^{|S|}$
  - 每个当前结果由当前状态随机产生，产生结果的概率分布为
    - 产生结果概率矩阵：$B=\{b_{jk} \mid s_j \rightarrow v_k$产生的概率，$0<j<|S|、0<k<|V|\}$

学习算法应用的一般步骤

- 定义问题
  - 类型
- 整理数据
  - 定义原始样例（内容+标签）
  - 训练集
  - 验证集
  - 测试集
- 选择机器学习算法
  - 开始选择多个
- 选择特征
  - 对应的特征表达方式
- 算法+特征的组合
- 评估

谢 谢！