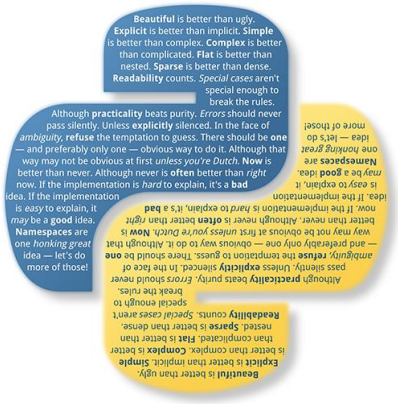
CDA 数据分析师
CERTIFIED DATA ANALYST

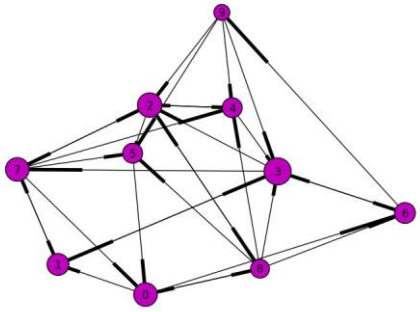
精英班系列课程

机器学习入门



python™

机器学习定义及分类



机器学习

CDA 数据分析师 CERTIFIED DATA ANALYST

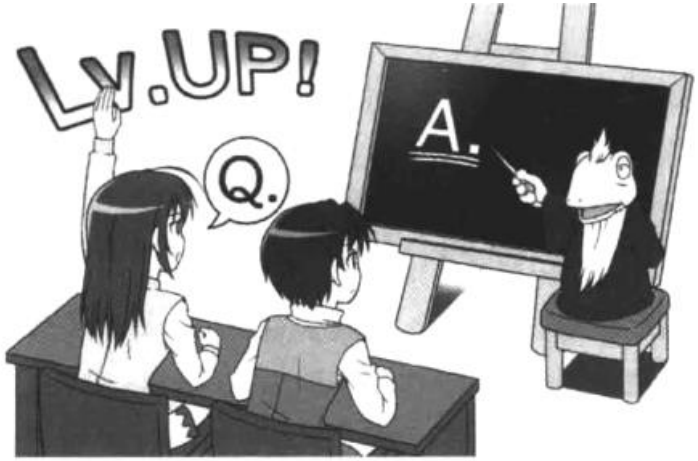
- 机器学习的定义：
  - ✓ 机器学习(Machine Learning，ML)是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构（利用数据或经验等）使之不断改善自身的性能
  - ✓ 它是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域，包括网络搜索、垃圾邮件过滤、推荐系统、广告投放、信用评价、欺诈检测、股票交易和医疗诊断等应用

机器学习

- 机器学习的分类：
  - ✓ 监督学习（supervised learning）
  - ✓ 无监督学习（unsupervised learning ）
  - ✓ 半监督学习（semi-supervised learning ）
  - ✓ 强化学习（reinforcement learning，增强学习）

机器学习

- 监督学习：



监督学习

- ✓ 定义：主要特点是要在训练模型时提供给学习系统训练样本以及样本对应的类别标签，因此又称为有导师学习。例：学生从老师那里获取知识、信息，老师提供对错指示、告知最终答案的学习过程
- ✓ 最终目标：根据在学习过程中获得的经验技能，对没学习过的问题也可以做出正确解答，使计算机获得这种泛化能力

机器学习

- 监督学习：
  - ✓ 典型的监督学习方法：决策树、支持向量机、监督式神经网络等分类算法和线性回归等回归算法
  - ✓ 应用：手写文字识别、声音处理、图像处理、垃圾邮件分类与拦截、网页检索、基因诊断、股票预测等
  - ✓ 典型任务：预测分类标签的分类、预测顺序的排列、预测数值型数据的回归

机器学习

- 无监督学习：

无监督学习

- ✓ 定义：主要特点是训练时只提供给学习系统训练样本，而没有样本对应的类别标签信息。例：没有老师的情况下，学生从书本或网络自学的过程
- ✓ 最终目标：无监督学习不局限于解决有正确答案的问题，所以目标可以不必十分明确

机器学习

- 无监督学习：
  - ✓ 典型的无监督学习方法：聚类学习、自组织神经网络学习
  - ✓ 应用：人造卫星故障诊断、视频分析、社交网站解析、声音信号解析、数据可视化、监督学习的前处理工具等
  - ✓ 典型任务：聚类、异常检测

机器学习

- 半监督学习：
  - ✓ 定义：在半监督学习方式下，训练数据有部分被标识，部分没有被标识，这种模型首先需要学习数据的内在结构，以便合理的组织数据来进行预测。算法上，包括一些对常用监督式学习算法的延伸，这些算法首先试图对标识数据进行建模，在此基础上再对未标识的数据进行预测
  - ✓ 例：给学生很多未分类的书本与少量的清单，清单上说明哪些书属于同一类别，要求对其他所有书本进行分类

机器学习

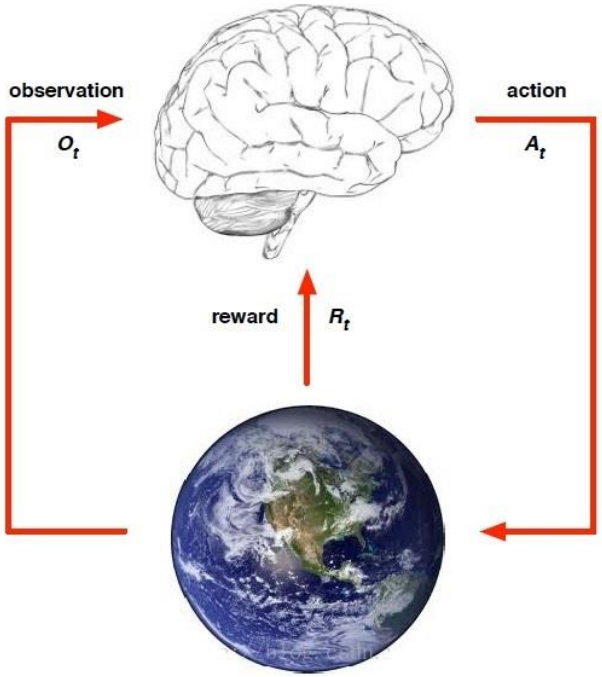
- 强化学习：

强化学习

✓ 定义：主要特点是通过试错来发现最优行为策略而不是带有标签的样本学习。如在没有老师提示的情况下，自己对预测的结果进行评估的方法。通过这样的自我评估，学生为了获得老师的最高评价将而不断的进行学习。强化学习被认为是人类主要的学习模式之一

机器学习

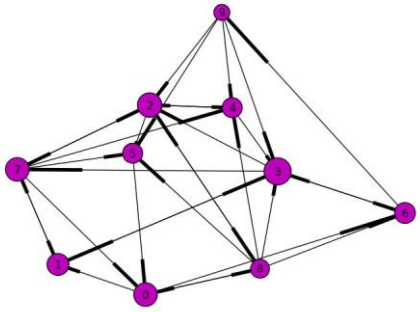
- 强化学习：研究如何基于环境而行动，以获得最大预期收益
- 强化学习的关键要素有：environment，reward，action和state
- 强化学习解决的问题是，针对一个具体问题得到一个最优的策略，使得在该策略下获得的reward最大



机器学习

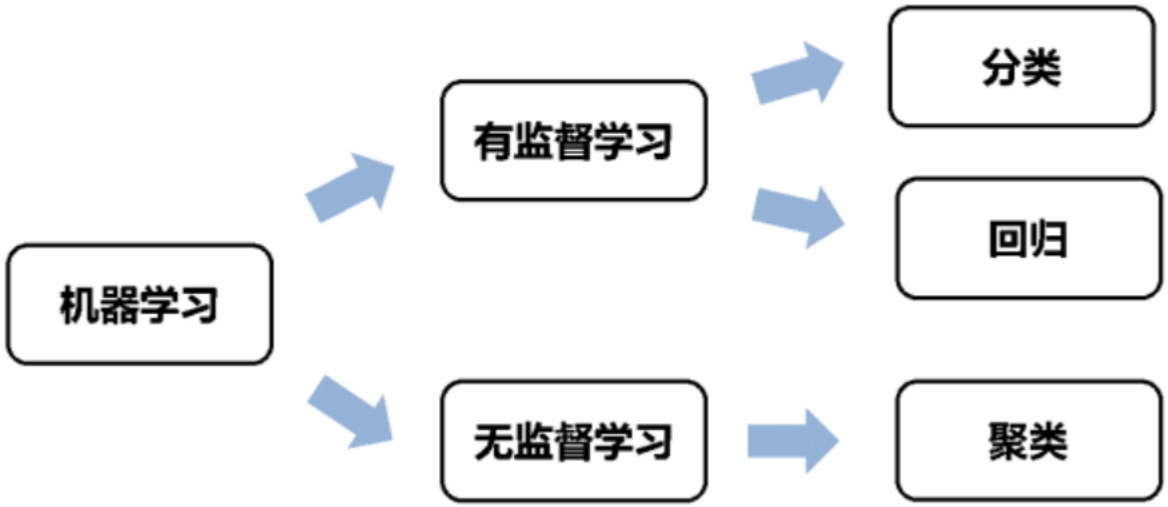
- 强化学习：
  - ✓ 最终目标：使计算机获得对没学习过的问题也可以做出正确解答的泛化能力
  - ✓ 应用：机器人的自动控制、计算机游戏中的人工智能、市场战略的最优化等
  - ✓ 典型任务：回归、分类、聚类、降维。例：下棋、机器人、自动驾驶等

机器学习任务



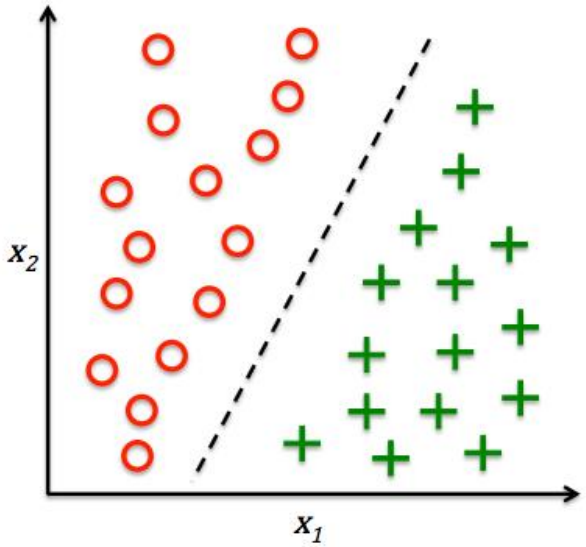
机器学习任务

- 在机器学习中，要解决某一问题，通常把问题转为成分类、回归、聚类、强化学习这四类问题进行解决



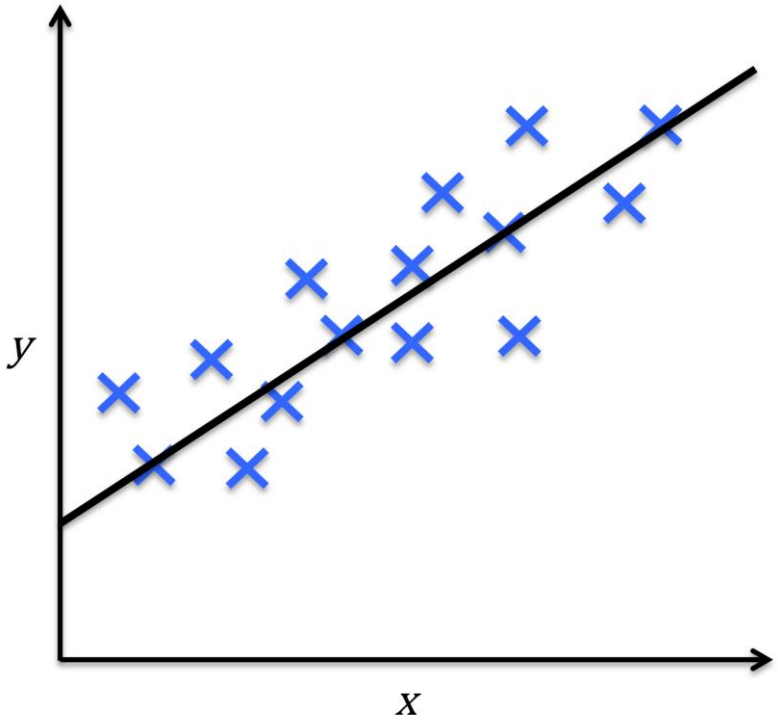
机器学习任务

- 分类问题：根据数据样本抽取出的特征，判定其属于有限个类别中的哪一个。在实际中做分类的时候，大多会产出一个概率值，对概率值做排序得到该样本属于哪个类别的概率最高
- 应用：垃圾邮件识别，结果类别为垃圾邮件和正常邮件；文本情感褒贬分析，结果类别褒、贬，股票的涨跌



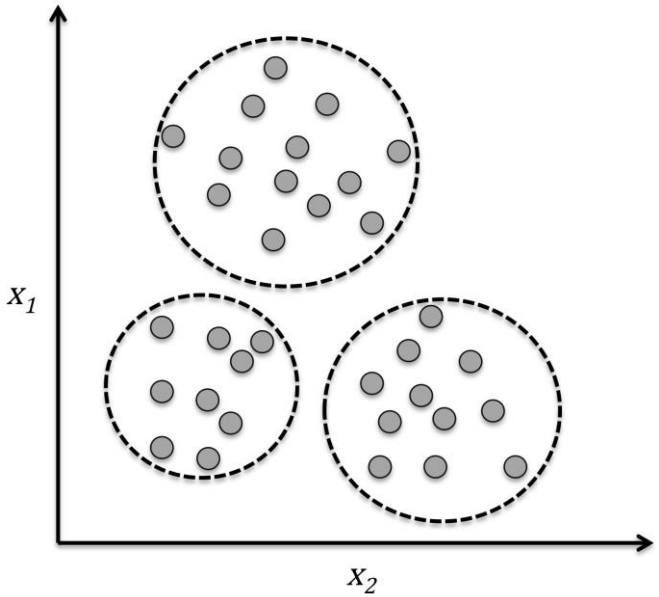
机器学习任务

- 回归问题：根据样本上抽取的特征，预测连续值结果。属于有监督学习
- 应用：预测电影的票房值，预测某城市的房价、股票的价格等



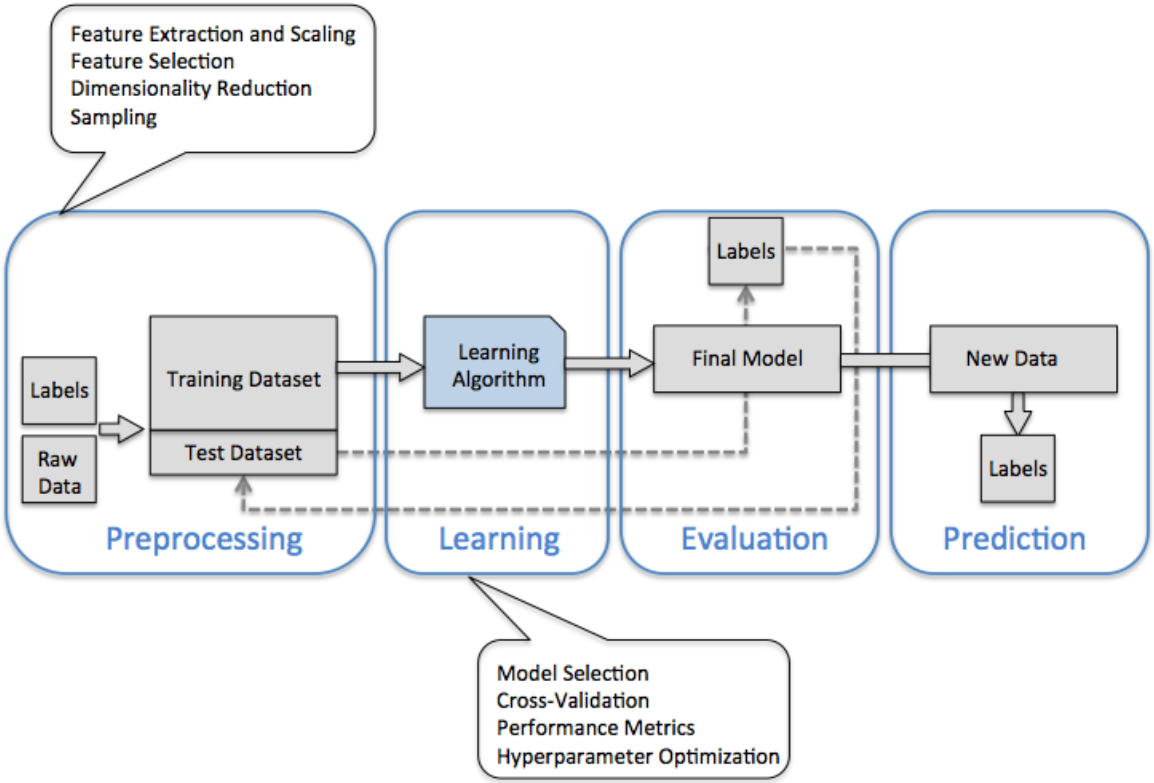
机器学习任务

- 聚类问题：根据数据样本抽取的特征，挖掘出数据的关联模式
- 应用：相似用户挖掘、新闻聚类等



机器学习任务

- 机器学习流程：数据预处理（Preprocessing）、模型学习（Learning）、模型评估（Evaluation）、新样本预测（Prediction）



机器学习任务

CDA 数据分析师 CERTIFIED DATA ANALYST

- 基本术语：
  - ✓ 每一条记录为：一个实例（instance）或样本（sample）
  - ✓ 数据集：所有记录的集合
  - ✓ 训练集：含有参考答案的数据，用来训练模型的已标注数据，用来建立模型，发现规律
  - ✓ 验证集：模型训练过程中单独留出的样本集，用于调整模型的超参数和用于对模型的能力进行初步评估
  - ✓ 测试集：用来评估模最终模型的泛化能力，但不能作为调参、选择特征等算法相关的选择的依据。已标注数据，通常做法是将标注隐藏，输送给训练好的模型，通过结果与真实标注进行对比，评估模型的学习能力

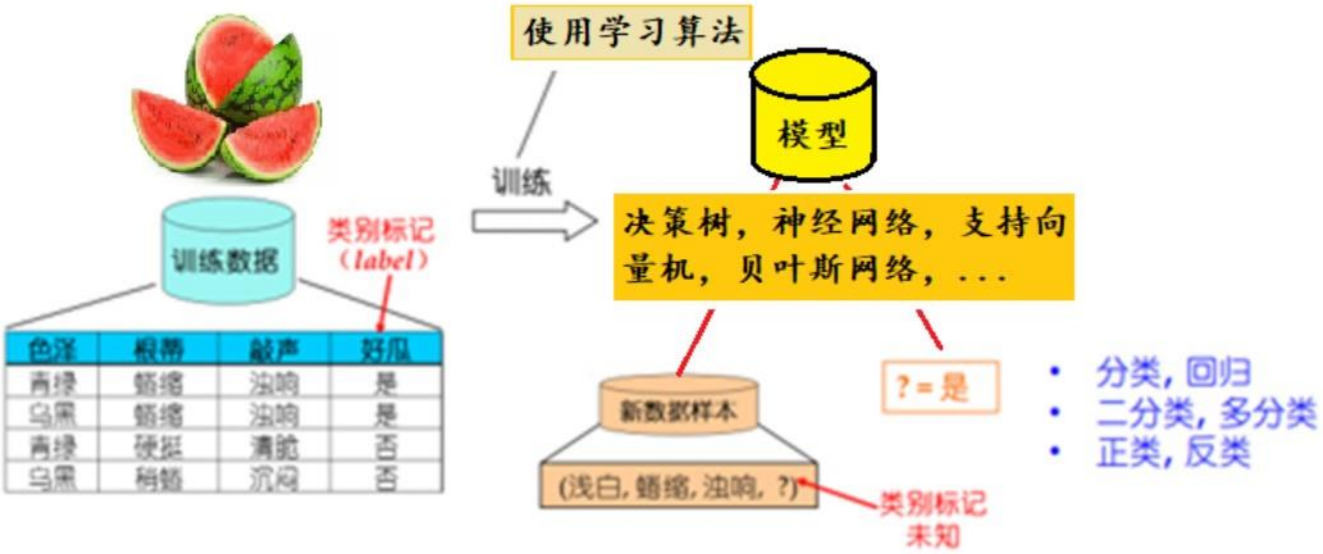
机器学习任务

- 如何理解？
  - ✓ 训练集：学生的课本，学生根据课本里的内容来掌握知识
  - ✓ 验证集：作业，通过作业可以知道不同学生学习情况、进步的速度快慢
  - ✓ 测试集：考试，考的题是平常都没有见过，考察学生举一反三的能力

机器学习任务

- 为什么要测试集？
  - ✓ 训练集直接参与了模型调参的过程，显然不能用来反映模型真实的能力（防止课本死记硬背的学生拥有最好的成绩，即防止过拟合）
  - ✓ 验证集参与了人工调参(超参数)的过程，也不能用来最终评判一个模型（刷题库的学生不能算是学习好的学生）
  - ✓ 所以要通过最终的考试(测试集)来考察一个学(模)生(型)真正的能力（期末考试）

机器学习任务



样本：对一个西瓜的描述

数据集：样本的集合

属性或特征：反映事件或对象在某方面的表现或性质的事项，例如：色泽，根蒂，敲声

属性值：属性上的取值，例如：青绿，蜷缩，浊响

属性空间或样本空间或输入空间：由属性张成的空间，例如，我们把色泽，根蒂，敲声，作为三个坐标轴，则它们张成一个用于描述西瓜的三维空间，每个西瓜都可以在这个坐标中找到自己的坐标位置

特征向量：由于空间中每个点对应一个坐标向量，因此，也把一个样本也称为一个特征向量

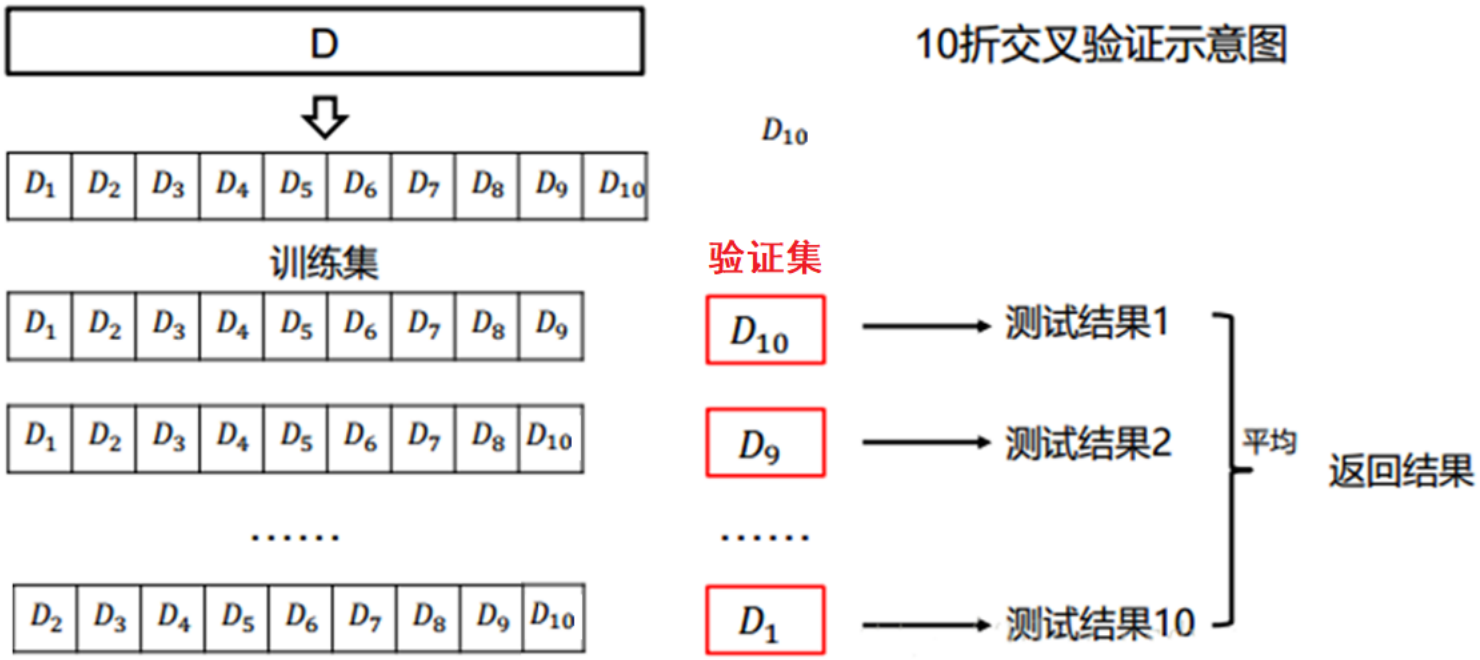
机器学习任务

- 总体数据集划分方法：
  - ✓ 根据已有全部标注数据，随机选出一部分数据（比如70%）作为训练数据，余下的作为测试数据
- 训练集/验证集的划分：两种方法
  - ✓ 方法1：从训练集中，再随机选出一部分数据（比如90%）作为训练数据，余下的作为验证数据
  - ✓ 方法2：交叉验证法先将训练集D划分为k个大小相似的互斥子集，每个子集都尽可能保持数据分布的一致性，即从D中通过分层采样得到。然后，每次用k-1个子集的并集作为训练集，余下的那个子集作为验证集，这样就可获得k组训练/验证集，从而可进行k次训练和验证，最终返回的是这个k个测试结果的均值

机器学习任务

交叉验证法

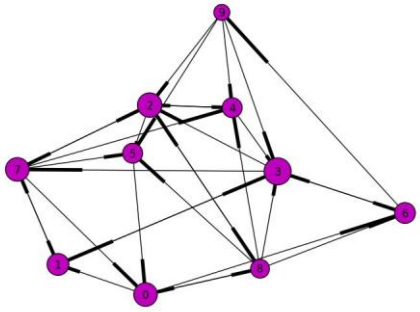
- ✓ 通常把交叉验证法称为“k折交叉验证”，k最常用的取值是10，此时称为10折交叉验证



机器学习任务

- 最终模型的测试：
  - ✓ 首先用训练集训练出模型，然后用验证集验证模型，根据情况不断调整模型，选出其中最好的模型，记录最好的模型的各项设置，然后据此再用（训练集+验证集）数据训练出一个新模型，作为最终的模型，最后用测试集评估最终的模型
  - ✓ 由于验证集数据的信息会被带入到模型中去，因此，验证误差通常比测试误差要小。需要注意的是：测试误差是我们得到的最终结果，即便我们对测试得分不满意，也不应该再返回重新调整模型，因为这样会把测试集的信息带入到模型中去

模型误差

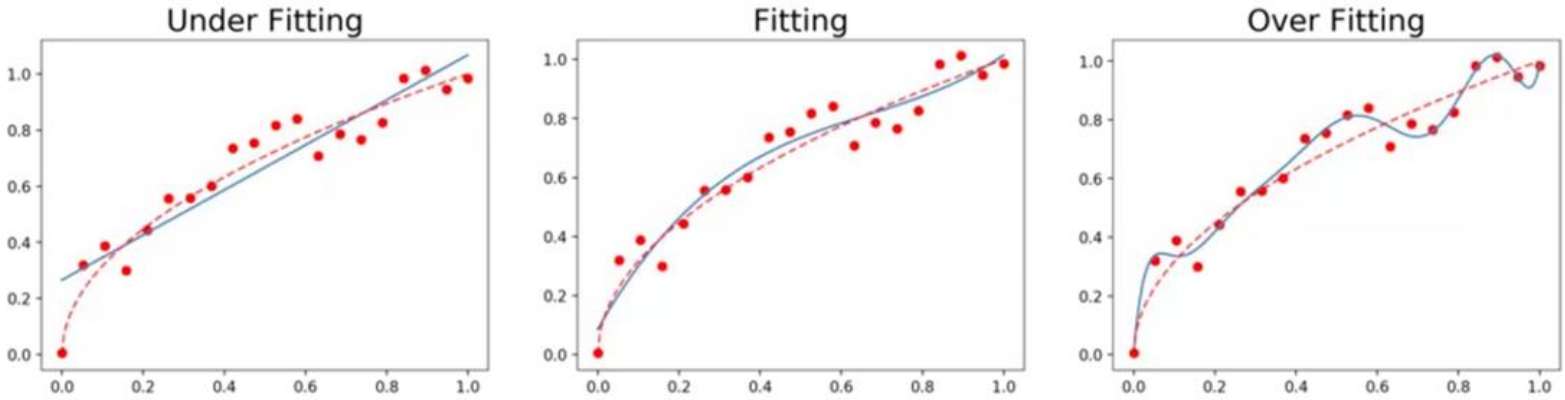


模型误差

- 机器学习算法针对特定数据所训练出来的模型并非是十全十美的，再加上数据本身的复杂性，误差不可避免
- 模型误差 = 偏差（Bias）+ 方差（Variance）+ 数据本身的误差
- 偏差：度量了学习算法的期望预测与真实结果的偏离程度，即刻画了算法本身的拟合能力；
- 方差：度量了同样大小的训练集的变动所导致的学习性能变化，即刻画了数据扰动所造成的影响
- 偏差和方差来源：导致偏差的原因有多种，其中一个就是针对非线性问题使用线性方法求解，当模型欠拟合时，就会出现较大的偏差；产生高方差的原因通常是由于模型过于复杂，即模型过拟合时，会出现较大的方差

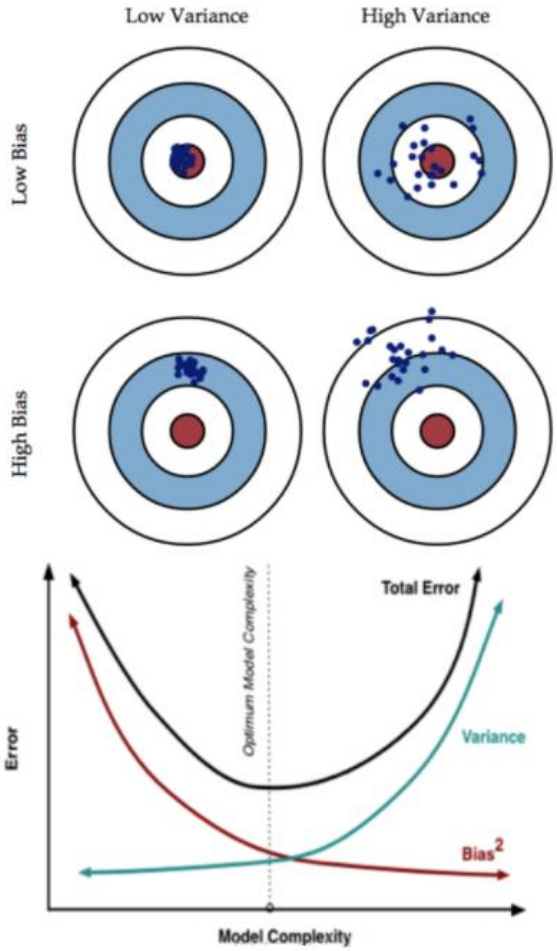
模型误差

- 模型的欠拟合与过拟合：



模型误差

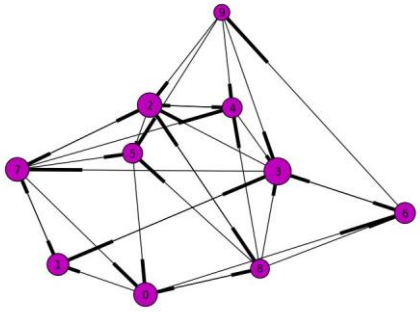
- 模型误差来源：


Low Variance
High Variance
Low Bias
High Bias
Total Error
Optimum Model Complexity
Variance
Bias^2
Error
Model Complexity

模型误差

- 针对偏差和方差的思路
- 偏差：实际上也可以称为避免欠拟合
  - ✓ 寻找更好的特征 -- 具有代表性
  - ✓ 用更多的特征 -- 增大输入向量的维度（增加模型复杂度）
- 方差：避免过拟合
  - ✓ 增大数据集合 -- 使用更多的数据，噪声点比减少（减少数据扰动所造成的影响（紧扣定义）
  - ✓ 减少数据特征 -- 减少数据维度，高维空间密度小（减少模型复杂度）
  - ✓ 正则化方法
  - ✓ 交叉验证法

机器学习评价标准



评价标准

评价标准：分类问题

	预测1	预测0	合计
实际1(P)	TP	FN	$TP+FN(P)$
实际0(N)	FP	TN	$FP+TN(N)$
合计	$TP+FP$	$FN+TN$	$TP+FN+FP+TN$

- ✓ 准确率(Accuracy)：预测正确的样本占所有样本的比例

$$\frac{TP+TN}{TP+FN+FP+TN}$$

评价标准

CDA 数据分析师 CERTIFIED DATA ANALYST

评价标准：分类问题

- ✓ 精确率(Precision)：在所有被分类为正例的样本中，真正是正例的比例

$$\frac{TP}{TP+FP}$$

- ✓ 召回率(Recall)（医学上称作灵敏度）:实际为正例的样本中，被预测为正例的样本比例

$$\frac{TP}{TP+FN}$$

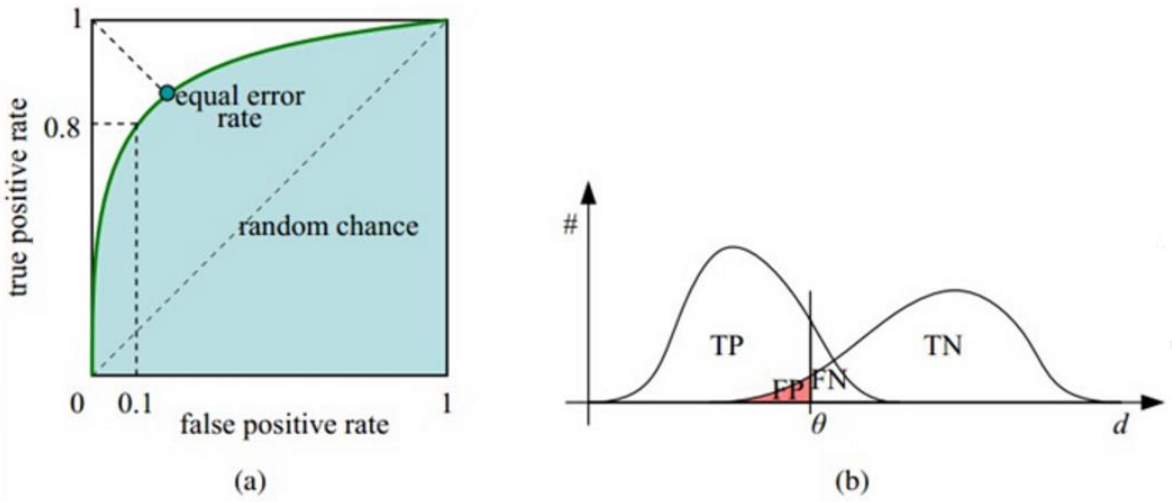
- ✓ 特异度(Specificity):实际为负的样本中，有多大概率被预测出来

$$SP=\frac{TN}{FP+TN}$$

评价标准

评价标准：二分类问题

- ROC曲线：根据一系列不同的二分类方式（阈值或分界值），以真阳性率TPR（灵敏度）为纵坐标，假阳性率FPR（1-特异度）为横坐标绘制的曲线



机器学习任务

- 评价标准：回归问题
  - ✓ MSE（Mean Squared Error）：

$$\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2$$

  - ✓ 均方根误差（RMSE）

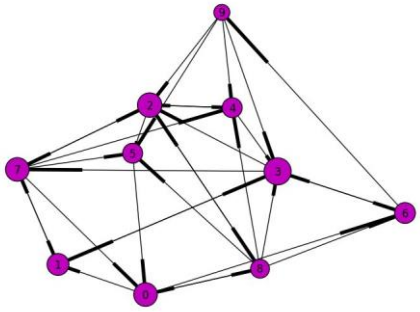
$$\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}$$

机器学习任务

- 评价标准：回归问题
  - ✓ MAE(平均绝对误差)

$$\frac{1}{m}\sum_{i=1}^{m}\left|(y_i-\hat{y}_i)\right|$$

机器学习算法库Scikit-Learn



Scikit-Learn

- Scikit-learn项目最早由数据科学家David Cournapeau在2007年发起，需要NumPy和SciPy等其他包的支持，是Python语言中专门针对机器学习应用而发展起来的一款开源框架
- 和其他众多的开源项目一样，Scikit-learn目前主要由社区成员自发进行维护。可能是由于维护成本的限制，Scikit-learn相比其他项目要显得更为保守。这主要体现在两个方面：一是Scikit-learn从来不做除机器学习领域之外的其他扩展，二是Scikit-learn从来不采用未经广泛验证的算法
- http://scikit-learn.org/stable/index.html

Talk is cheap

Show me the

CODE