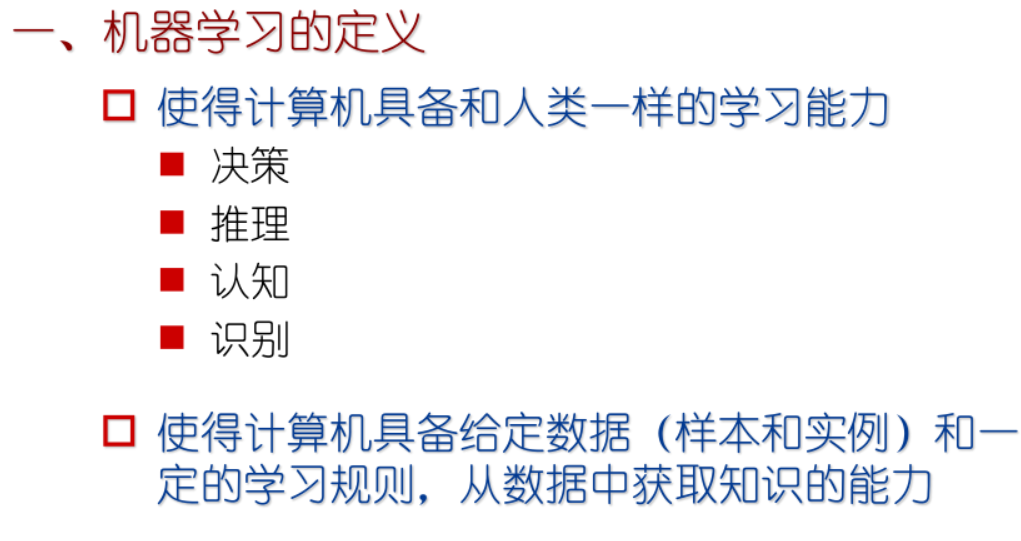
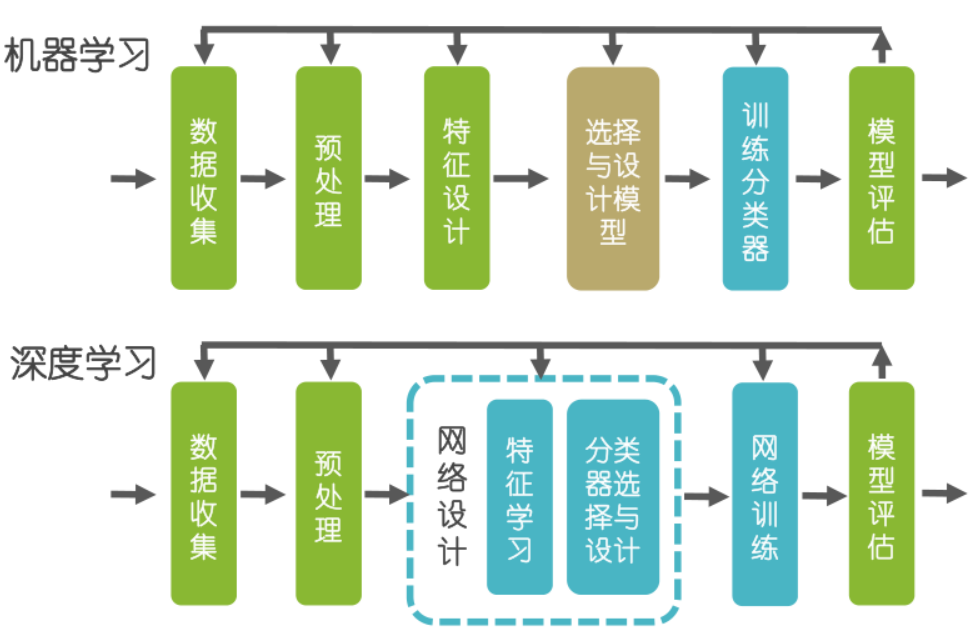
**1、 什么是监督学习？什么是非监督学习？什么是训练集？什么是测试集？各个指标的含义？什么是回归任务？什么是分类任务？其他机器学习概念。**

机器学习和深度学习：

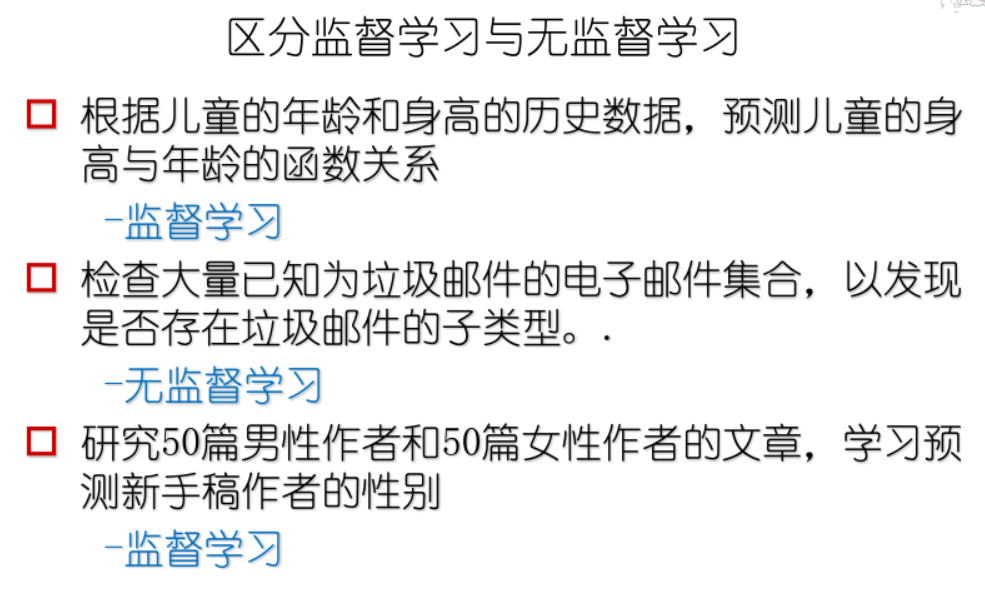
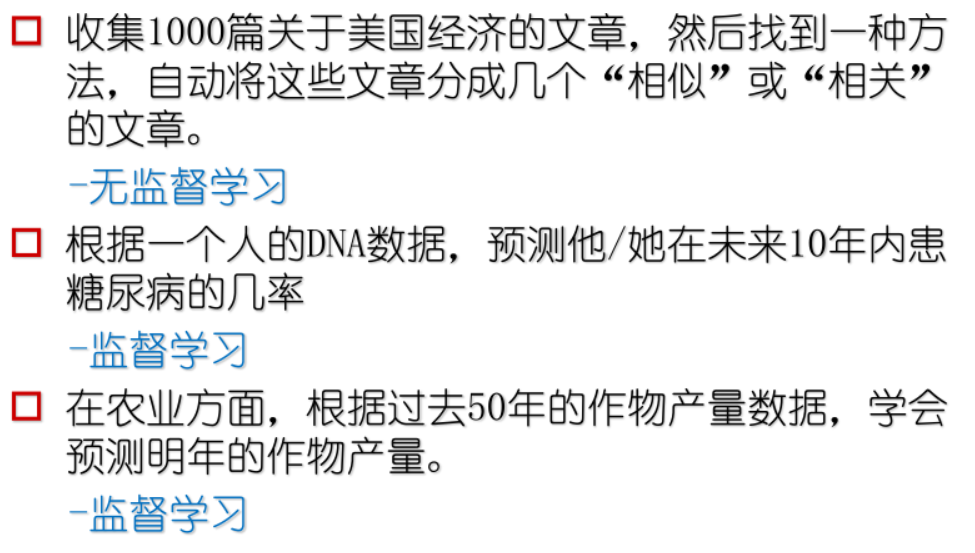
 

监督学习：训练含有很多特征的数据集，不过数据集中的样本都有一个标签或者目标。

举例：分类问题、回归问题、物体识别、语义分割、图像字幕等

非监督学习：通过训练含有很多特征的未标注的训练样本，学习到数据集上有用的结构性质。

举例：聚类、降维(PCA, Autoencoder)、特征学习等

半监督学习：半监督学习训练中的数据由部分标记数据和未标记数据组成。

举例：半监督分类等

弱监督学习：标注成本过高，或很难获得全部真值标签。

包括：不完全监督、不确切监督、不准确监督

强化学习：

预测目标：分类：离散值

回归：连续值

聚类：无标记信息

泛化能力：模型适用于新样本的能力

机器学习基本流程：数据采集→数据预处理→特征工程→模型的构建与调优→模型的评价

**数据采集**：爬虫、公开数据集、数据库等

**数据预处理**：数据可能存在缺失或异常，对之进行处理

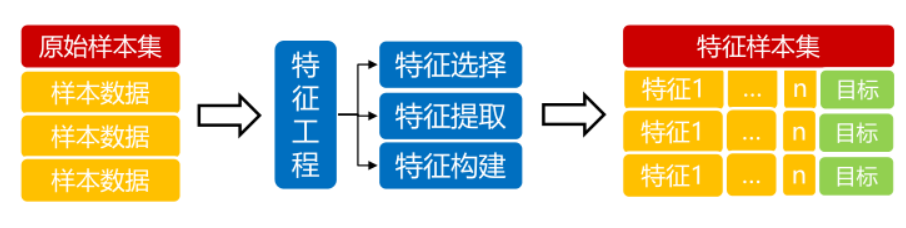
**特征工程**：特征选择、特征提取和特征构建：

特征选择：剔除不相关或者冗余的特征，减少有效特征的个数，减少模型训练的时间，提高

模型的精确度

特征提取：将原始特征转换为一组具有明显物理意的特征，如PCA、ICA、LDA

特征构建：从原始数据中人工的构建新的特征，包括特征组合、特征拆分

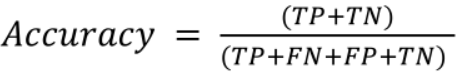


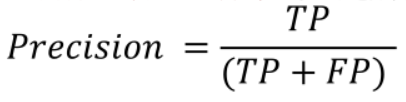
**模型的构建与调优**：1、构建合适的预测模型；2、在数据上训练；3、得到对应的输出

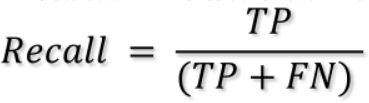
**模型评价**：

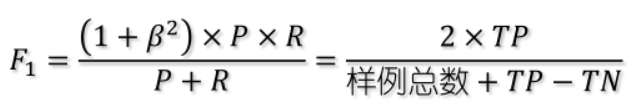
分类问题：



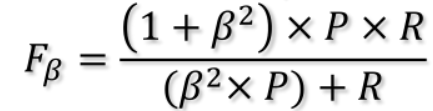
准确率：

查准率：

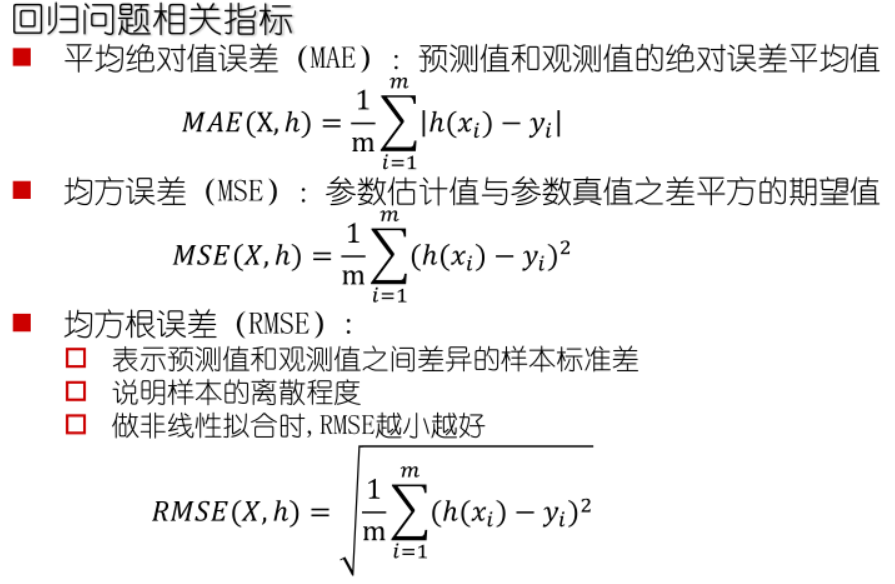
查全率：

F1 score：



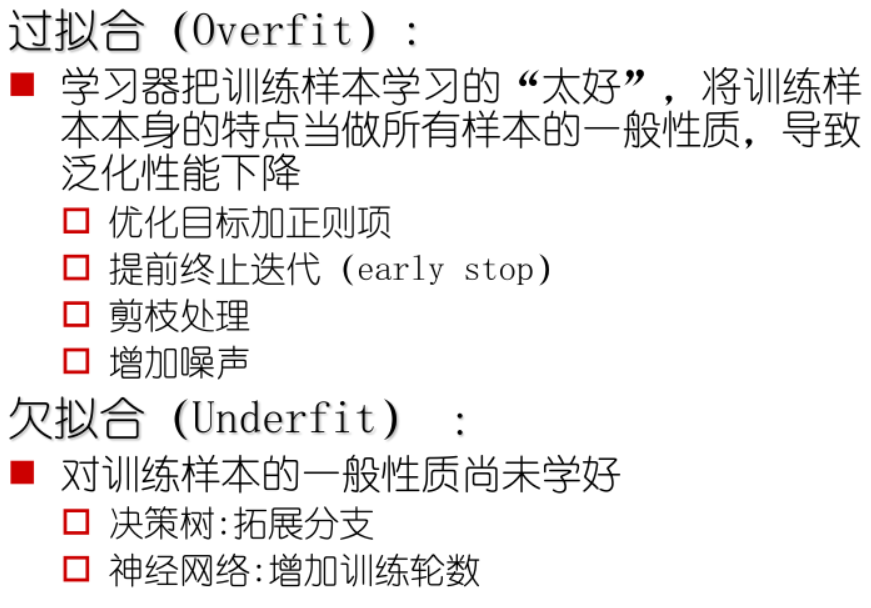
Fβ score: 

回归问题：



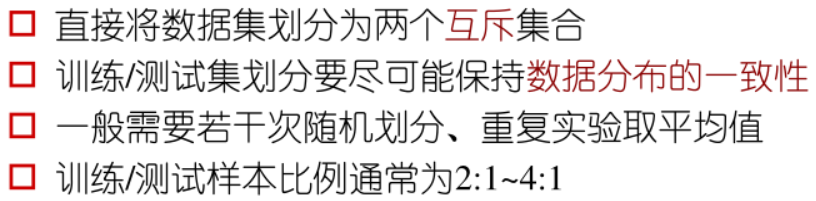
模型评估与选择：过拟合：学习器把训练样本本身特点当做所有潜在样本都会具有的一般性质

欠拟合：训练样本的一般性质尚未被学习器学好。

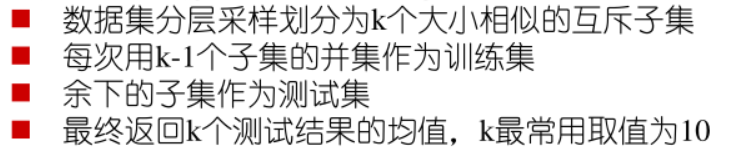


评估方法：

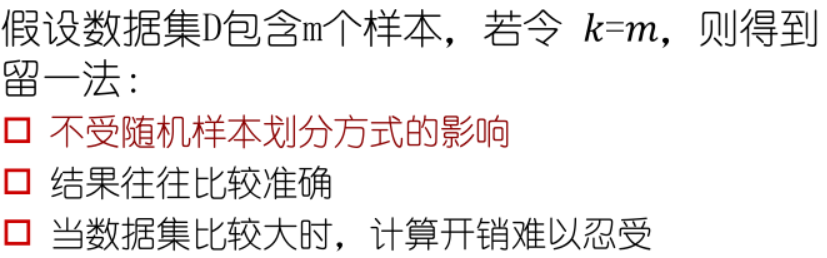
留出法：



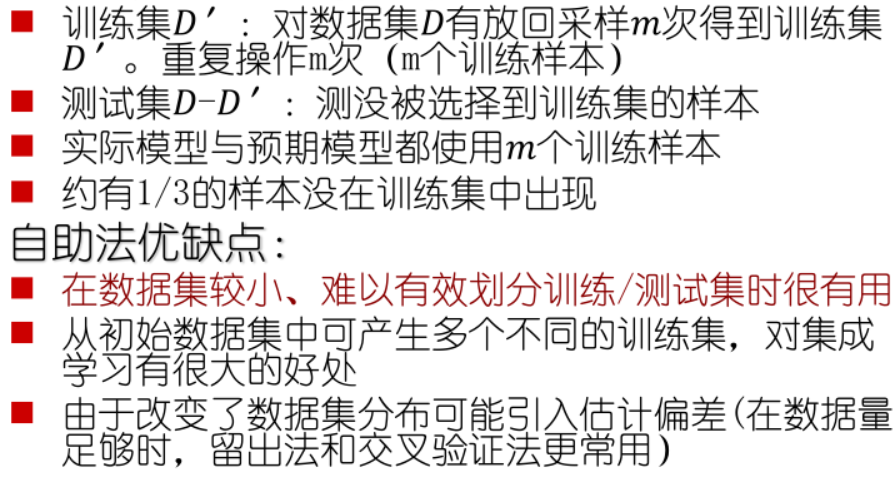
交叉验证法：



留一法：

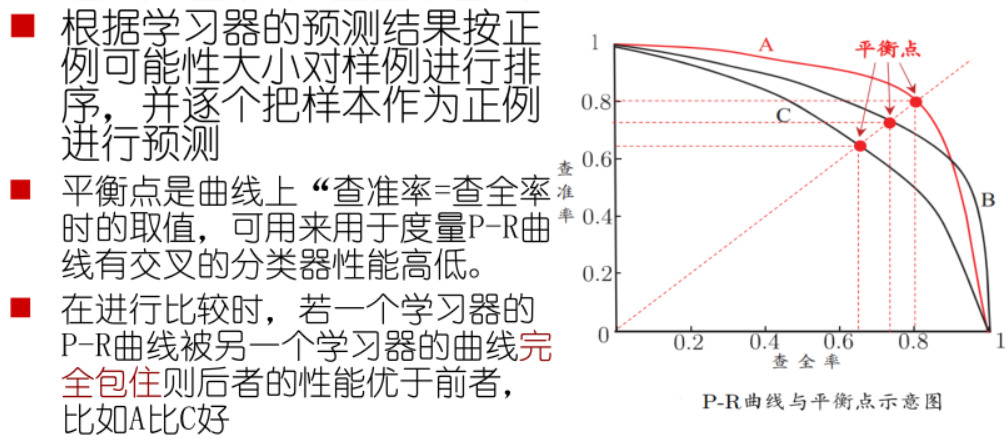


自助法：

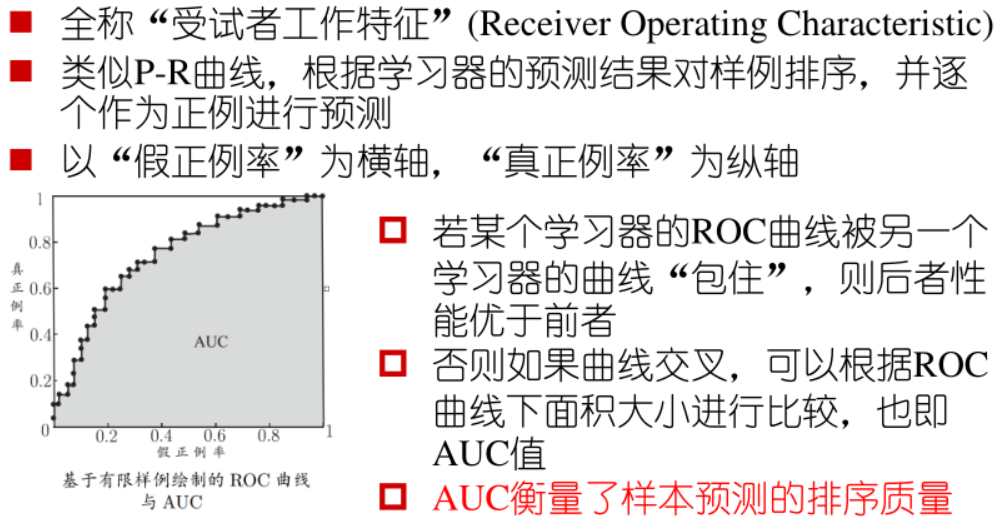


性能度量：回归任务最常用的性能度量是“均方误差”；分类任务最常用的是错误率和精度。

P-R曲线：



ROC曲线：



偏差与方差：

偏差度量了学习算法期望预测与真实结果的偏离程度，刻画了学习算法本身的拟合能力

方差度量了同样大小训练集变动所导致的学习性能变化，刻画了数据扰动所造成的影响

