第七章机器学习

**1.概念**

机器学习≈寻找一个函数

泛化能力是指机器学习算法对新鲜样本的 **适应能力**。通常期望学习模型具有**较强的泛化能力**

特征向量（Feature Vector）：每个样本的**特征**对应的**特征空间**中的一个坐标向量。

标记（Label）：训练数据中可能会指出训练结果的信息

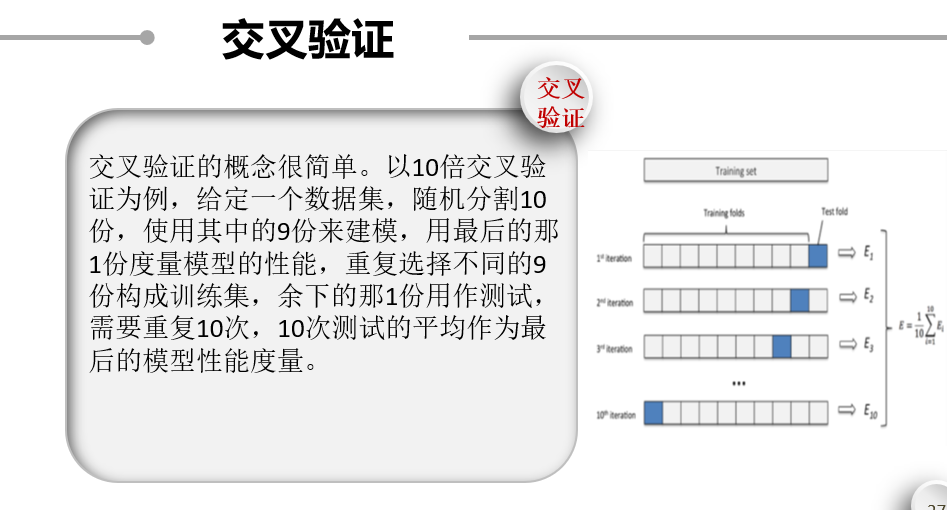
训练集（平时练习）——验证集（月考）——测试集（高考）

输出误差：误差

测试误差 训练误差或经验误差

—模型在测试样本集上的整体误差 ——指模型在训练样本集上的整体误差

交叉验证：



**2.机器学习算法简介**

**监督学习：数据被特别标记**——分类 离散值（二分类，多分类）——**决策树模型、**

回归：进行预测的是 连续值

分类与回归区别：（1）输出不同

分类问题输出的是物体所属的**类别**，回归问题输出的是物体的**值**。

分类问题输出的值是**定性**的，回归问题输出的值是**定量**的。

（2）目的不同

分类的目的是为了**寻找决策边界**。

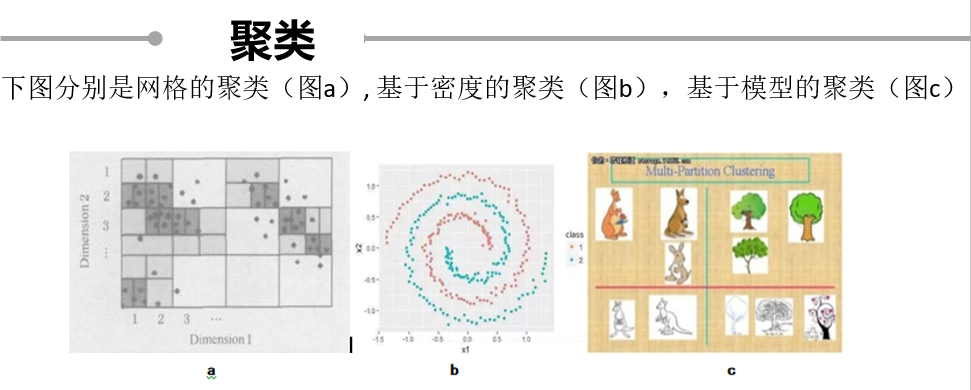
回归的目的是为了**找到最优拟合**。

（3）结果不同

分类的结果**没有逼近**，对就是对，错就是错。

回归是对真实值的一种**逼近预测**。

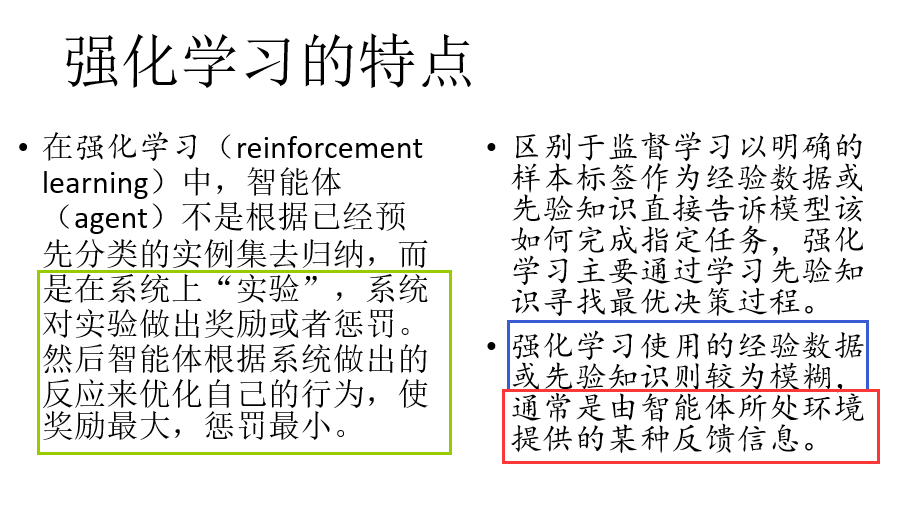
**无监督学习：数据并不被特别标记 ，**计算机**自行学习**分析数据内部的规律、特征等，进而**得出**一定的**结果——聚类算法（基于划分的聚类算法，密度聚类算法）**



关联规则分析：想知道哪些商品顾客可能会在一次购物时同时购买？——营销策略。

**半监督式学习：**输入数据部分被标识，部分没有被标识**，没**标识数据的数量常常远远**大于**有标识数据数量。

强化学习：根据 反馈信息 来调整机器行为以**实现自动决策**的一种机器学习方式。

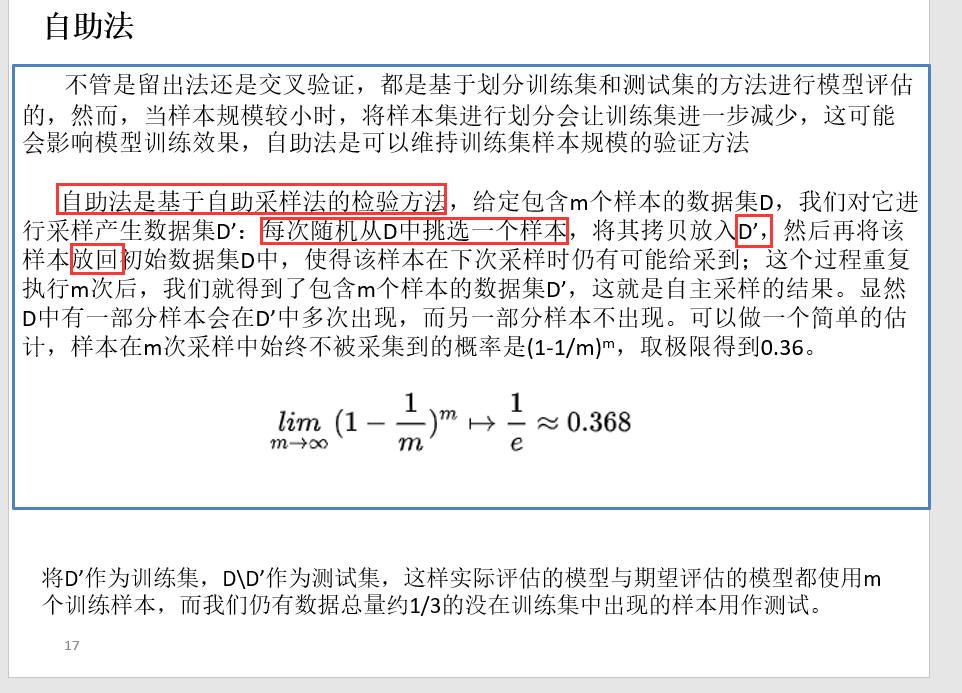
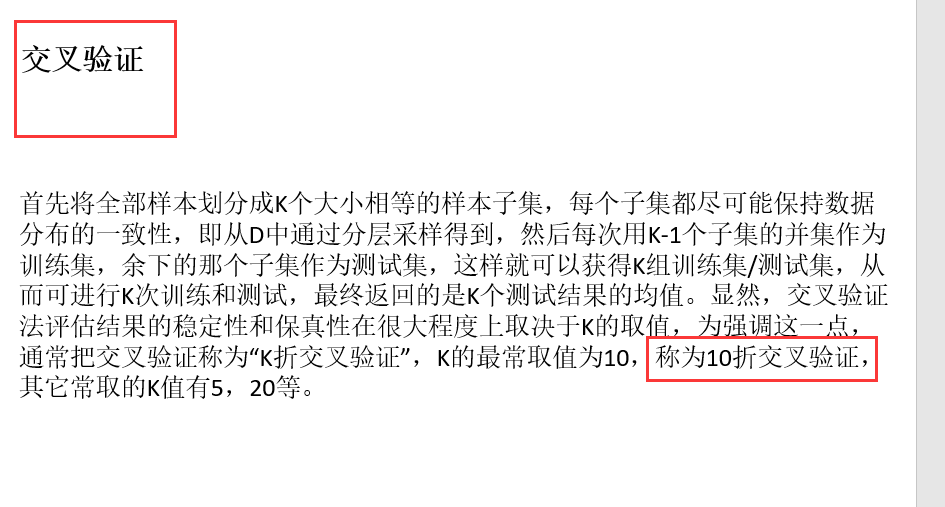
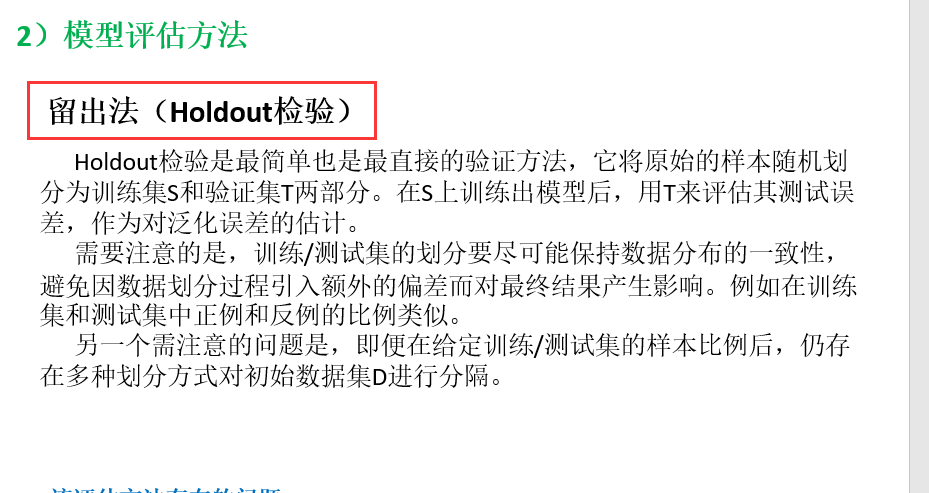
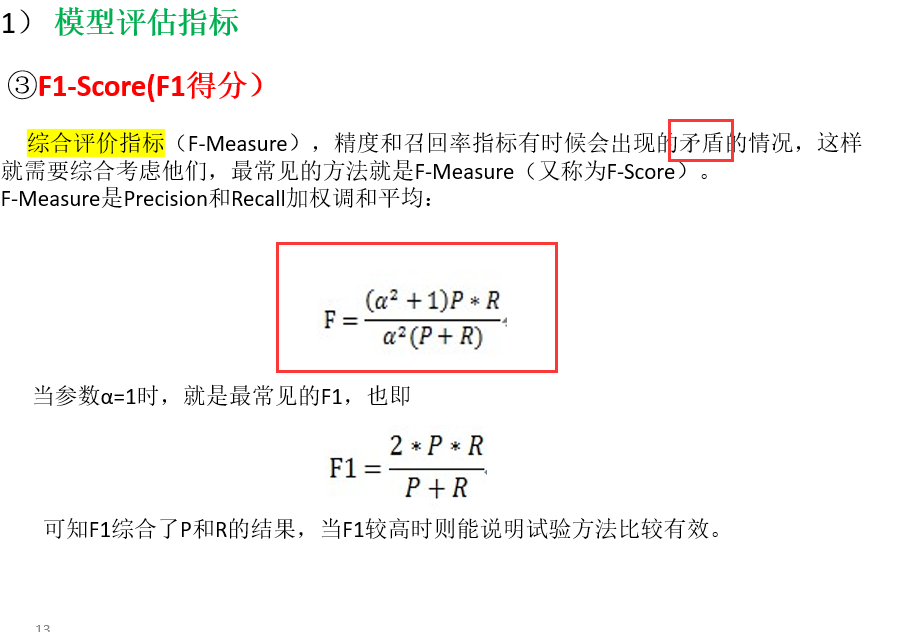
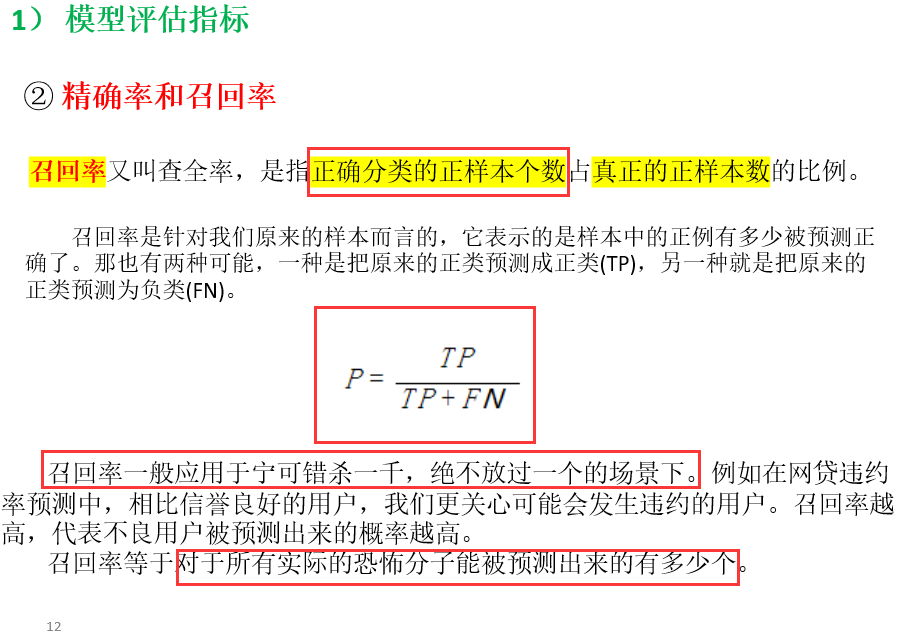
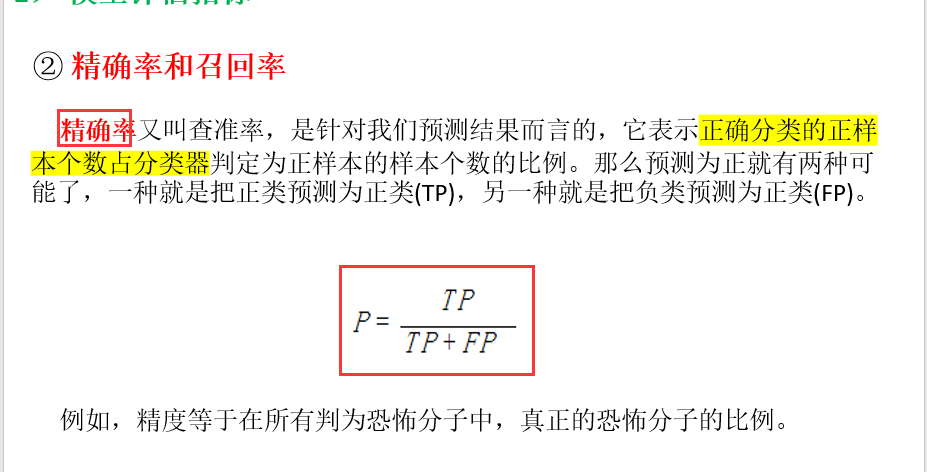
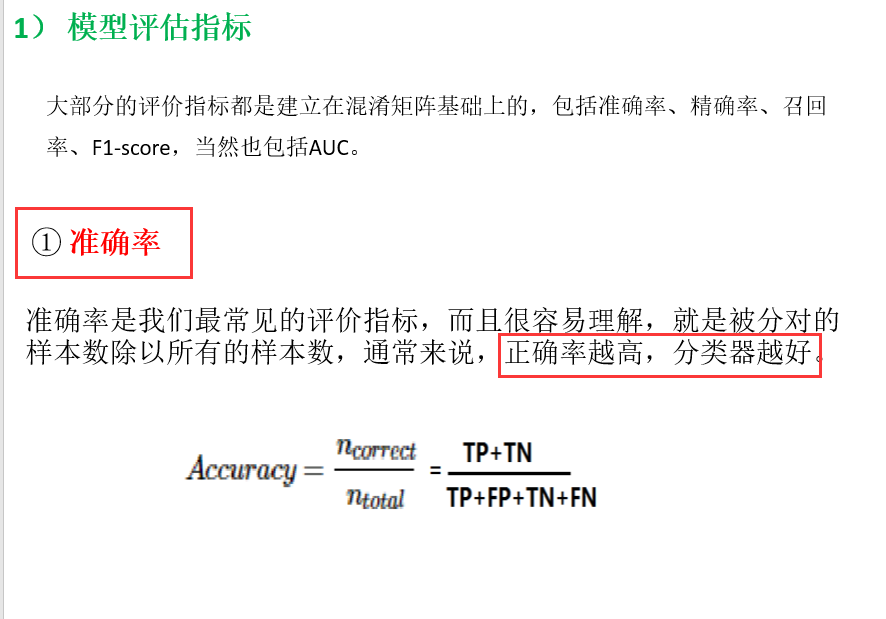
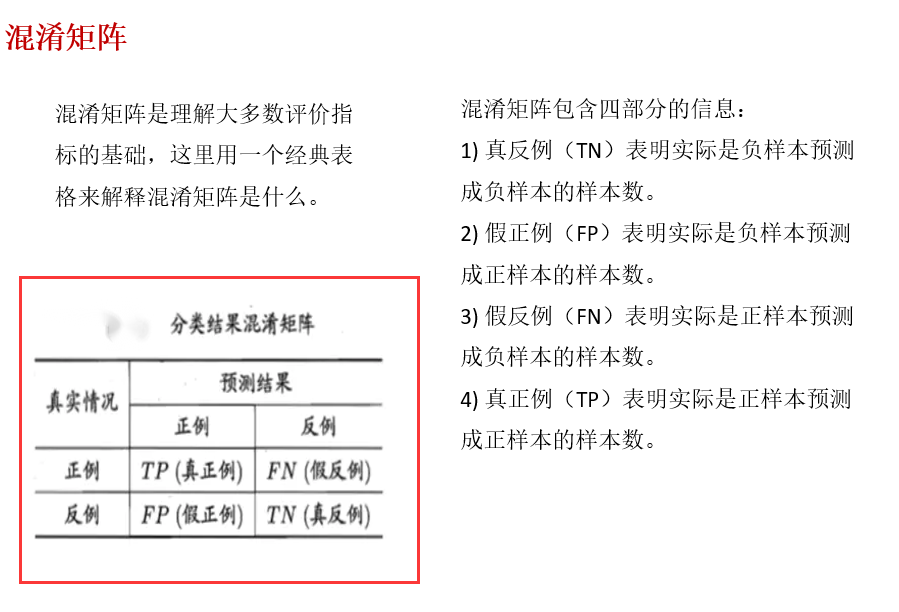


**3．机器学习工作流程**

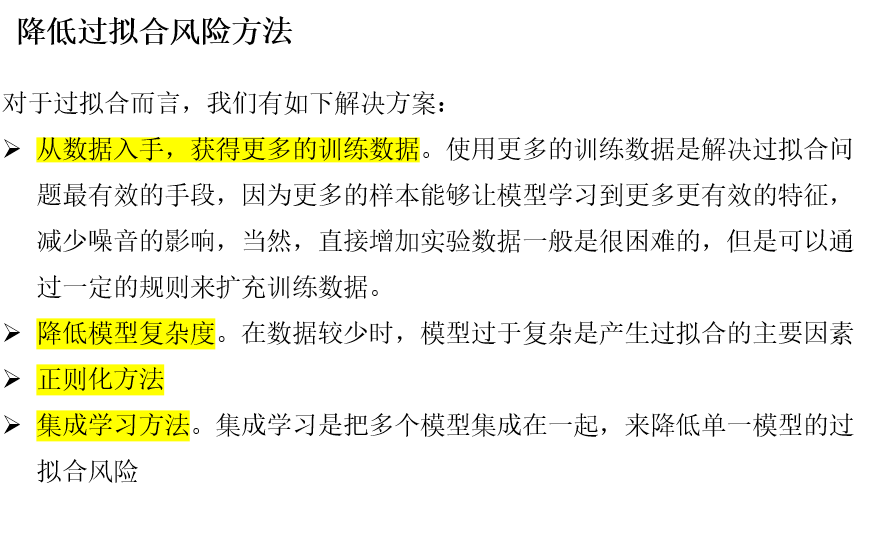
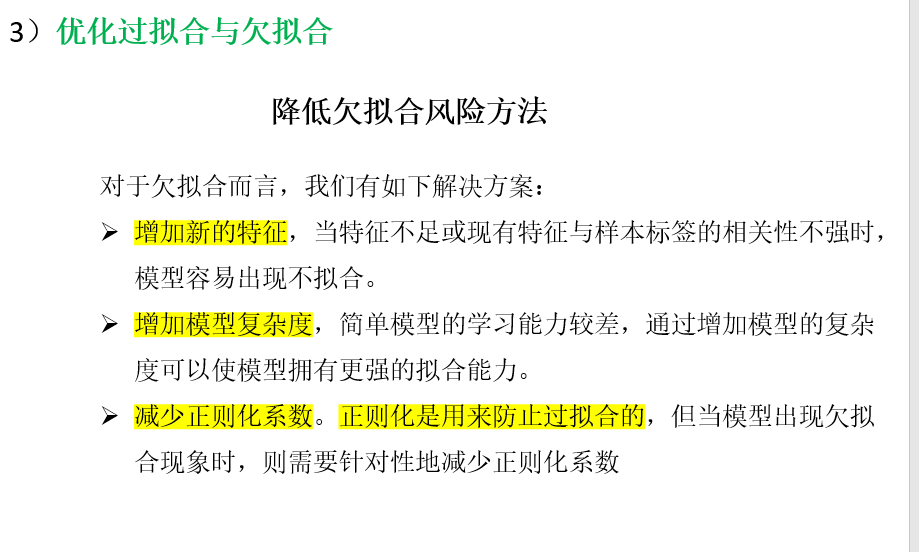
获取数据——数据基本处理——特征工程——机器学习（模型训练）——模型评估

1. **数据采集**：爬虫；API；数据库
2. **数据处理**：数据清洗；数据预处理（缺失值。重复值处理，数据类型的转换、字符串数据的规整）；数据归一化/标准化
3. **特征工程**：把**原始数据**转变为模型的**训练数据**的过程，它的目的就是获取更好的训练数据特征——**特征构建、特征提取、特征选择**三个部分。
4. **构建模型**：① 建立训练数据集和测试数据集，通常80%为训练数据集。② 选择机器学习算法③模型调优④模型融合
5. **模型评估**：训练好的模型进行评估

混淆矩阵：



留一法



（4）调参和最终模型