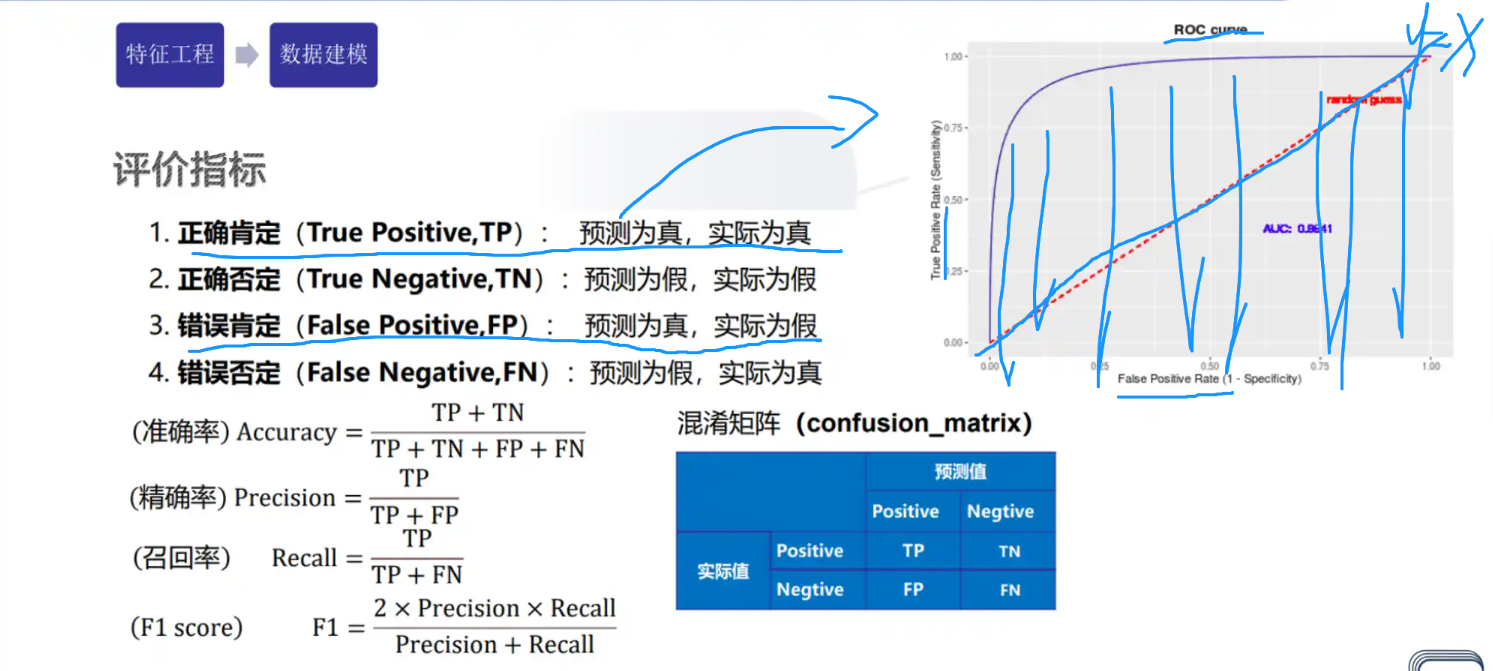
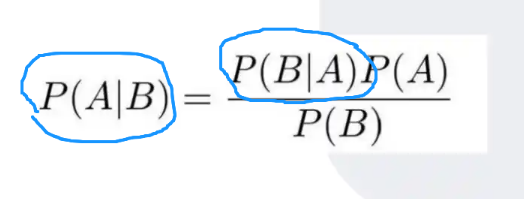
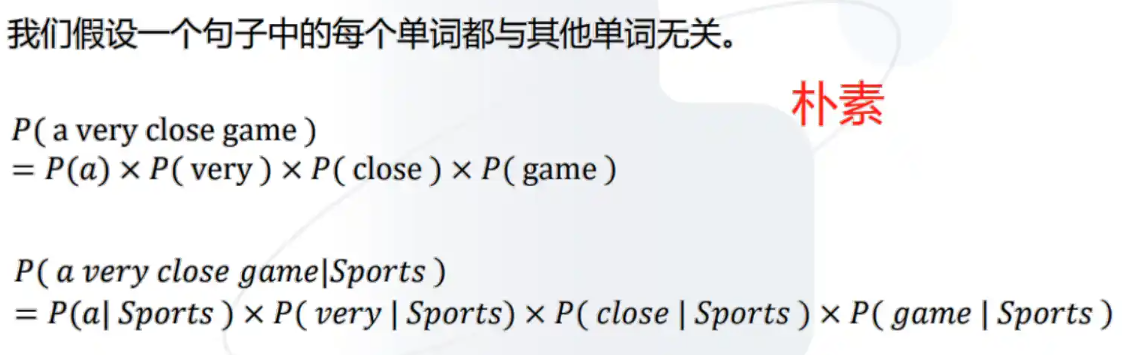
# 机器学习概述+朴素贝叶斯算法

1. 机器学习的一般步骤
   1. 数据搜集
      1. 数据划分
         1. 训练集
         2. 测试集（使用多次的测试【相当于周测小测月考】）
         3. 验证集（只使用一次，用于验证模型的效果【相当于高考】）
      2. K折交叉验证
         1. 解决数据量不够大的问题
         2. 解决参数调优的问题
      3. 数据集不平衡
         1. 采用SMOTE算法（在偏少的数据中间生成一个该类型数据，不改变数据集的离散程度和分布状态）
   2. 数据清洗
      1. 脏数据
         1. 拼写错误
         2. 不合法值
         3. 空值
      2. 数据清洗
         1. 异常检测
         2. 重复处理
         3. 数据挖掘
   3. 特征工程
      1. 减少特征数量（数据之间有联系和推导），降维，增强模型泛化能力
      2. 增强特征和特征值之间的理解
      3. 方法
         1. 降维方面的：PCA,ICA,LDA
         2. 图像方面的：SHIFT，Gabor,HOG
         3. 文本方面的：词袋模型，词嵌入模型
      4. 深度学习的优势：不用想破脑袋找合适的特征工程
   4. 数据建模
      1. 概率模型
      2. 非概率模型
      3. 损失函数
         1. 损失函数值越小，模型性能越好
         2. 优化算法
            1. 用于找到最优全局损失函数
            2. 梯度下降法
            3. 牛顿法
            4. 拟牛顿法
            5. ADAM等
         3. 结果评测指标
            1. 
2. 机器学习的类型
   1. 监督学习
      1. 工业应用多
      2. 输入数据中有监督信息
   2. 无监督学习
      1. 采用聚类方法，学习结果为类别
   3. 强化学习（起步阶段且难度大）
      1. 以环境反馈（奖惩信号）作为输入，以统计和动态规划技术为指导
      2. 最符合ai发展方向
3. 数学基础
   1. 高等数学
      1. 微分
      2. 积分
      3. 梯度
      4. 泰勒公式
      5. 拉格朗日乘子
   2. 线性代数
      1. 向量
      2. 矩阵
      3. 行列式
      4. 秩
      5. 线性方程组
      6. 特征值和特征向量
   3. 概率论
      1. 随机事件和概率
      2. 概率的基本公式
      3. 常见分布
      4. 期望
      5. 协方差
   4. 数理统计
      1. 抽样分布
      2. 参数估计
      3. 假设检验
      4. 相关性分析
   5. 信息论
      1. 熵
      2. 交叉熵
4. 朴素贝叶斯
   1. 贝叶斯定理
      1. 先验概率P(A)
         1. 没有训练数据之前某一事件的初始概率
      2. 后验概率P(A|B)
         1. 在真实观测情况下计算得出的概率
      3. P(A|B)~~P(A)\*P(B|A)
   2. 不同角度观察的概率学派
      1. 概率学派
         1. 认为事件在多次重复之后会趋于一个稳定的概率
      2. 贝叶斯派
         1. 世界是不确定的，对事件的预测结果取决于观察者对世界的预先估计
   3. 举例
      1. 朴素贝叶斯算法
         1. 以输入法中的联想输入为例
         2. 假设每一个单词之间都没关系（称之为朴素贝叶斯）
         3. 
         4. 这样做的好处是可以在相当程度上简化计算
         5. 朴素不是为了提升算法性能，而是不得已而作的假设
      2. 拉普拉斯平滑
         1. 为每一个计数加1，避免出现0
      3. 朴素贝叶斯算法已经非常成熟（1970年至今），不适合作为科研中的创新点。