朴素贝叶斯分类

- sklearn工具
  - 三种朴素贝叶斯分类算法

    伯努利朴素贝叶斯是以文件为粒度，如果该单词在某文件中出现了即为 1，否则为 0。而多项式朴素贝叶斯是以单词为粒度，
    会计算在某个文件中的具体次数。而高斯朴素贝叶斯适合处理特征变量是连续变量，且符合正态分布（高斯分布）的情况。
    比如身高、体重这种自然界的现象就比较适合用高斯朴素贝叶斯来处理。而文本分类是使用多项式朴素贝叶斯或者伯努利朴素贝叶斯。
    - 高斯朴素贝叶斯：GaussianNB

      应用场景：特征变量是连续变量，符合高斯分布，比如说人的身高，物体的长度。
    - 多项式朴素贝叶斯：MultinomialNB

      应用场景：特征变量是离散变量，符合多项分布，在文档分类中特征变量体现在一个单词出现的次数，或者是单词的 TF-IDF 值等。
    - 贝努力朴素贝叶斯：BernoulliNB

      应用场景：特征变量是布尔变量，符合 0/1 分布，在文档分类中特征是单词是否出现。
  - TF-IDF值

    TF-IDF 实际上是词频 TF 和逆向文档频率 IDF 的乘积
    - 概念：词频TF，逆向文档频率IDF

      词频TF = 单词出现的次数 / 该文档的总单词数
      逆向文档频率IDF = log(文档总数 / (该单词出现的文档数+1))
    - 使用sklearn求TF-IDF：TfidfVectorizer类
- 如何对文档进行分类
  - 准备阶段：对文档分词，加载停用词，计算单词权重

    英文文档：NTLK
    中文文档：jieba
  - 分类阶段：生成分类器，分类器做预测，计算准确率

幕布 - 思维概要整理工具