# 数据挖掘与数据分析
# Homework 1

周子龙 1851201

Mar 21, 2021

## 1 Problem1

### 1.1 维度灾难

维度灾难是由理查德 · 贝尔曼 (Richard E. Bellman) 在考虑动态优化问题时首次提出来的术语。他举例来说，100 个平均分布的点能把一个单位区间以每个点距离不超过 0.01 采样；而当维度增加到 10 后，如果以相邻点距离不超过 0.01 小方格采样一单位超正方体，则需要 1020 个采样点: 所以，这个 10 维的超正方体也可以说是比单位区间大 1018 倍。

对应于数据挖掘中，当样本数量远远小于样本维度的时候，就会产生过度拟合现象。

本文使用了具有 13 个特征，173 个样本的红酒数据作为原始数据，再对各个维度数据进行归一化后，分别计算了不同维度下个样本之间的欧式距离的差异，以及欧式距离的平均值，得到了如下的结果:

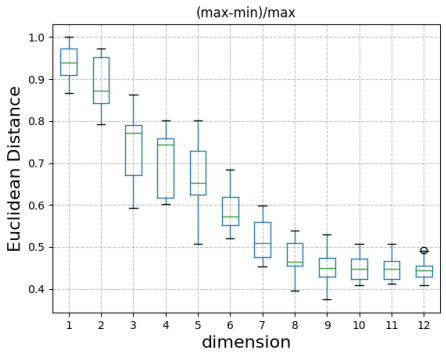


图 1: 不同维度下的欧式距离的差异值，样本容量为 100

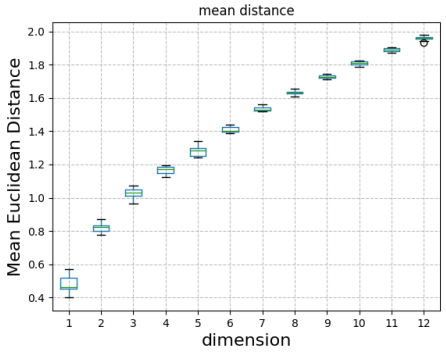


图 2: 不同维度下的平均欧式距离，样本容量为 100

从上图可以很明显的看到，当维度不断增加，各个样本间的欧式距离越小。

## 1.2 数据降维

降维就是一种对高维度特征数据预处理方法。降维是将高维度的数据保留下最重要的一些特征，去除噪声和不重要的特征，从而实现提升数据处理速度的目的。在实际的生产和应用中，降维在一定的信息损失范围内，可以为我们节省大量的时间和成本。降维也成为应用非常广泛的数据预处理方法。

### 1.2.1 主成分分析

PCA(Principal Component Analysis)，即主成分分析方法，是一种使用最广泛的数据降维算法。PCA 的主要思想是将 n 维特征映射到 k 维上，这 k 维是全新的正交特征也被称为主成分，是在原有 n 维特征的基础上重新构造出来的 k 维特征。PCA 的工作就是从原始的空间中顺序地找一组相互正交的坐标轴，新的坐标轴的选择与数据本身是密切相关的。其中，第一个新坐标轴选择是原始数据中方差最大的方向，第二个新坐标轴选取是与第一个坐标轴正交的平面中使得方差最大的，第三个轴是与第 1,2 个轴正交的平面中方差最大的。依次类推，可以得到 n 个这样的坐标轴。通过这种方式获得的新的坐标轴，我们发现，大部分方差都包含在前面 k 个坐标轴中，后面的坐标轴所含的方差几乎为 0。于是，我们可以忽略余下的坐标轴，只保留前面 k 个含有绝大部分方差的坐标轴。事实上，这相当于只保留包含绝大部分方差的维度特征，而忽略包含方差几乎为 0 的特征维度，实现对数据特征的降维处理。

这里不对 PCA 方法给出具体的推导，直接给出对上述红酒数据降维后的结果：

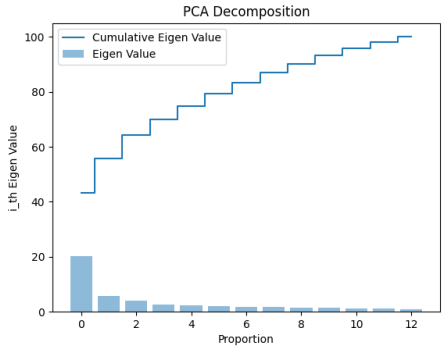


图 3: PCA 方法的到的特征值

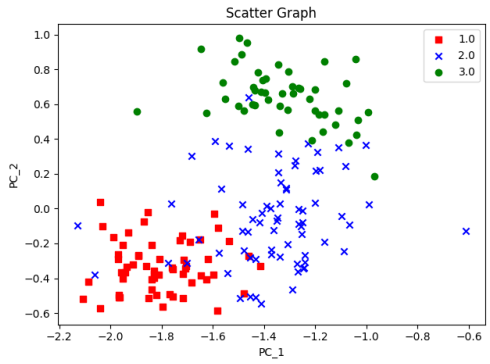


图 4: PCA 降维结果

### 1.2.2 线性判别分析

LDA 是一种监督学习的降维技术，也就是说它的数据集的每个样本是有类别输出的。这点和 PCA 不同。PCA 是不考虑样本类别输出的无监督降维技术。LDA 的思想可以用一句话概括，就是“投影后类内方差最小，类间方差最大”。什么意思呢？我们要将数据在低维度上进行投影，投影后希望每一种类别数据的投影点尽可能的接近，而不同类别的数据的类别中心之间的距离尽可能的大。

这里不对 LDA 方法给出具体的推导，直接给出对上述红酒数据降维后的结果：

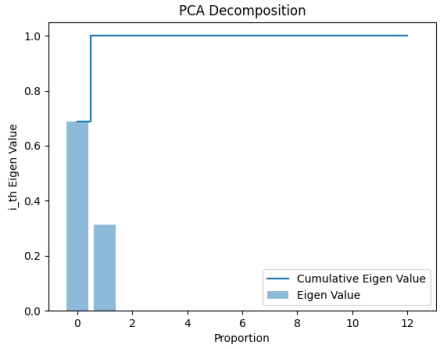


图 5: LDA 方法的到的特征值

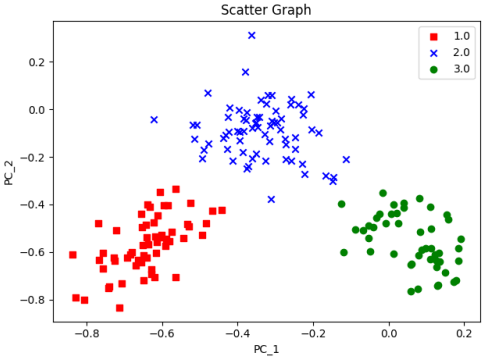


图 6: LDA 降维结果

对比 PCA 和 LDA，不难发现，由于 LDA 线性判别分析是有监督的问题，在降维的时候加入了类别的信息，而 PCA 是无监督的问题，没有标检，基于方差进行降维；对比结果而言，LDA 在区分数据上明显优于 PCA。

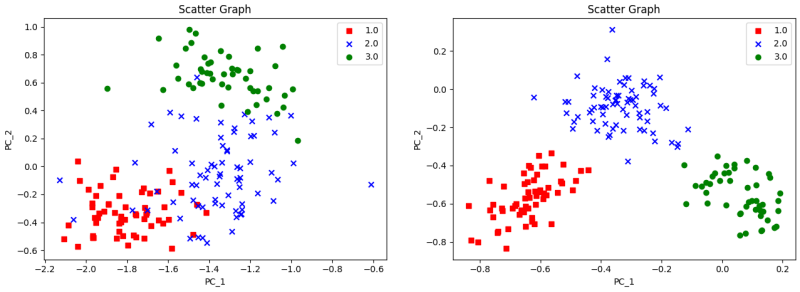


图 7: PCA(图左边) 与 LDA(图右) 降维结果比较

# 2 Problem2

## 2.1 距离与相似度

### 2.1.1 欧氏距离

欧氏距离是最常见的距离度量，衡量的是多维空间中各个点之间的绝对距离。

### 2.1.2 明氏距离

明氏距离是欧氏距离的推广，是对多个距离度量公式的概括性的表述。

### 2.1.3 曼哈顿距离

曼哈顿距离来源于城市区块距离，是将多个维度上的距离进行求和后的结果。

### 2.1.4 切比雪夫距离

切比雪夫距离起源于国际象棋中国王的走法，我们知道国际象棋国王每次只能往周围的 8 格中走一步，那么如果要从棋盘中 A 格 (x1, y1) 走到 B 格 (x2, y2) 最少需要走几步？扩展到多维空间，其实切比雪夫距离就是当 p 趋向于无穷大时的明氏距离。

这里直接给出各种距离计算上述数据集两点之间距离的情况：

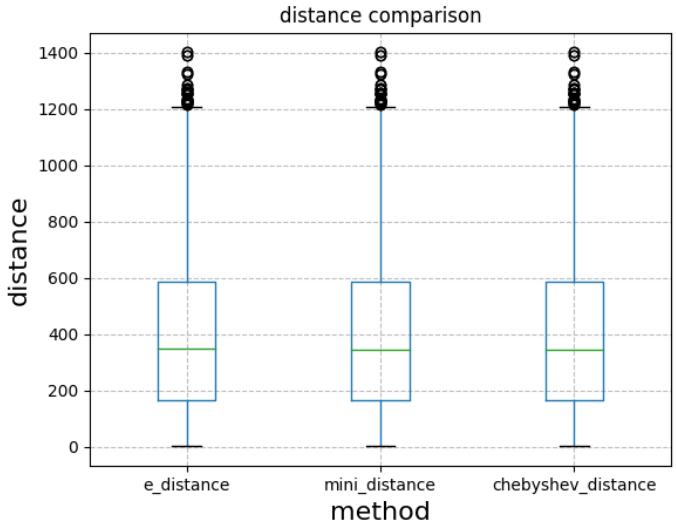


图 8: 各种距离实现方法对比

从结果可以看出，好像在本数据集中似乎没有太大影响。

## 2.2 基于信息论的相似度

### 2.2.1 余弦相似性

余弦相似性通过测量两个向量的夹角的余弦值来度量它们之间的相似性。0 度角的余弦值是 1，而其他任何角度的余弦值都不大于 1；并且其最小值是-1。从而两个向量之间的角度的余弦值确定两个向量是否大致指向相同的方向。两个向量有相同的指向时，余弦相似度的值为 1；两个向量夹角为 90° 时，余弦相似度的值为 0；两个向量指向完全相反的方向时，余弦相似度的值为-1。这结果是与向量的长度无关的，仅仅与向量的指向方向相关。余弦相似度通常用于正空间，因此给出的值为-1 到 1 之间。

### 2.2.2 KL 散度

在概率论或信息论中，KL 散度（ Kullback–Leibler divergence），又称相对熵（relative entropy），是描述两个概率分布 P 和 Q 差异的一种方法。它是非对称的，这意味着 D(P||Q)  D(Q||P)。特别的，在信息论中，D(P||Q) 表示当用概率分布 Q 来拟合真实分布 P 时，产生的信息损耗，其中 P 表示真实分布，Q 表示 P 的拟合分布。

### 2.2.3 JS 散度

JS 散度度量了两个概率分布的相似度，基于 KL 散度的变体，解决了 KL 散度非对称的问题。一般地，JS 散度是对称的，其取值是 0 到 1 之间。

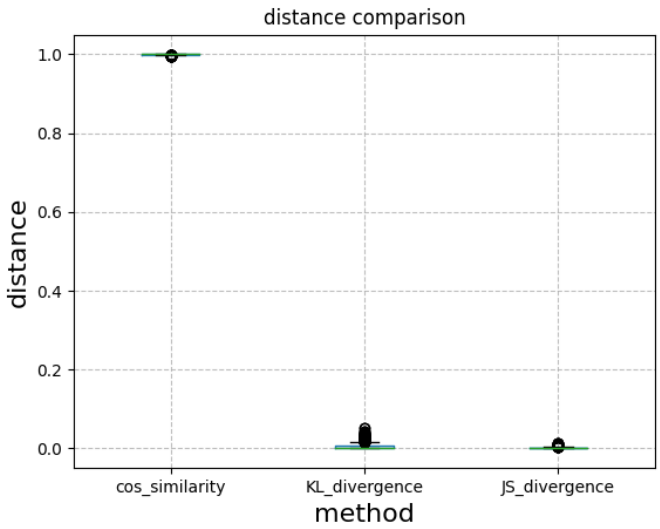


图 9: 各种相似度实现方法对比

# 3 problem3

## 3.1 缺失值处理

缺失值处理有几种常见的方法，首先最为简单也是最为直观的就是直接删除缺失值数据，但这可能造成样本容量的大幅度缩水。

### 3.1.1 平均值填充

将初始数据集中的属性分为数值属性和非数值属性来分别进行处理。 如果空值是数值型的，就根据该属性在其他所有对象的取值的平均值来填充该缺失的属性值； 如果空值是非数值型的，就根据统计学中的众数原理，用该属性在其他所有对象的取值次数最多的值 (即出现频率最高的值) 来补齐该缺失的属性值。与其相似的另一种方法叫条件平均值填充法（Conditional Mean Completer）。在该方法中，用于求平均的值并不是从数据集的所有对象中取，而是从与该对象具有相同决策属性值的对象中取得。

### 3.1.2 K 最近距离邻法

先根据欧式距离或相关分析来确定距离具有缺失数据样本最近的 K 个样本，将这 K 个值加权平均来估计该样本的缺失数据。

## 3.2 数据采样

在实际的分类问题中，数据集的分布经常是不均衡的。虽然不均衡的数据在分类时常常能得到较高的分类准确率，但对于某些情况而言，准确率的意义并不大，并不能提供任何有用的信息。从数据层面上而言，对于不平衡数据主要通过重采样的方法对数据集进行平衡。重采样方法是通过增加小众训练样本数的上采样和减少大众样本数的下采样使不平衡样本分布变平衡，从而提高分类器对小众的识别率。

### 3.2.1 上采样

从小众样本中进行随机采样来增加新的样本。比较常见的算法如 SMOTE 算法；以及对其改进的 Borderline-SMOTE 算法；原始的 SMOTE 算法对所有的小众样本都是一视同仁的，但实际建模过程中发现那些处于边界位置的样本更容易被错分，因此利用边界位置的样本信息产生新样本可以给模型带来更大的提升。Borderline-SMOTE 是一种自适应综合过采样方法，其解决思路是只为那些 K 近邻中有一半以上大众样本的小众样本生成新样本，因为这些样本往往是边界样本。在确定了为哪些小众样本生成新样本后，再利用 SMOTE 生成新样本。

### 3.2.2 下采样

从大众样本中随机选择少量样本（分为有放回和无放回两种），再合并原有小众样本作为新的训练数据集。常见的算法有 EasyEnsemble，可以有效解决了数据不均衡问题，且减少了欠采样造成的大众样本信息损失。但该算法未考虑小众样本极度欠缺的情况，当小众样本数远小于正确训练分类器所需的样本数时，每个基学习器的分类性能都可能会很差，进而导致最终分类器的分类效果差。