聚　类　方　法

Clustering methond

汇报人：李婧霞 宋梦晗

目　录

CONTENTS

01 介绍

02 相似度或距离

03 类或簇

04 类与类之间的距离

05 层次聚类

01

介绍

聚类分析是将个体或对象分类，使得同一类对象之间的相似性比与其他类的对象的相似性更强。是一种无监督学习，是在缺乏标签的前提下的一种分类模型。

聚类分析

Cluster analysis

目的：聚类分析是把相似的研究对象归成类，通过得到的类或簇来发现数据的特点或对数据进行处理。

分类：1.根据分类对象的不同

Q型聚类分析：对样本进行分类处理

R型聚类分析：对变量进行分类处理

2.根据聚类方法的不同

硬聚类：一个样本只能属于一个类，或类的交集为空集。

软聚类：一个样本可以属于很多个类，属于每个类的概率是不同的。

$$P(Z=1|X_1)=0.7 \qquad P(Z=2|X_1)=0.3$$

聚类分析的应用

- 用户分割 将用户分到不同的组别中，并根据据簇的特性而推送不同的广告。
- 欺诈检测 发现正常与异常的用户数据，识别其中的欺诈行为。

02

相似度或距离

聚类中，可以将样本集合看作是向量空间中点的集合，以该空间的距离来表示样本之间的相似度。

相似度或距离

Similarity or distance

闵科夫斯基距离（闵氏距离）：

对于连续m维空间中的两点 $x_i = (x_{1i}, x_{2i}, \cdots, x_{mi})^{\mathrm{T}}$ $x_j = (x_{1j}, x_{2j}, \cdots, x_{mj})^{\mathrm{T}}$

和

其闵科夫斯基距离为： $d_{ij} = \left(\sum_{k=1}^{m} |x_{ki} - x_{kj}|^p \right)^{\frac{1}{p}}$

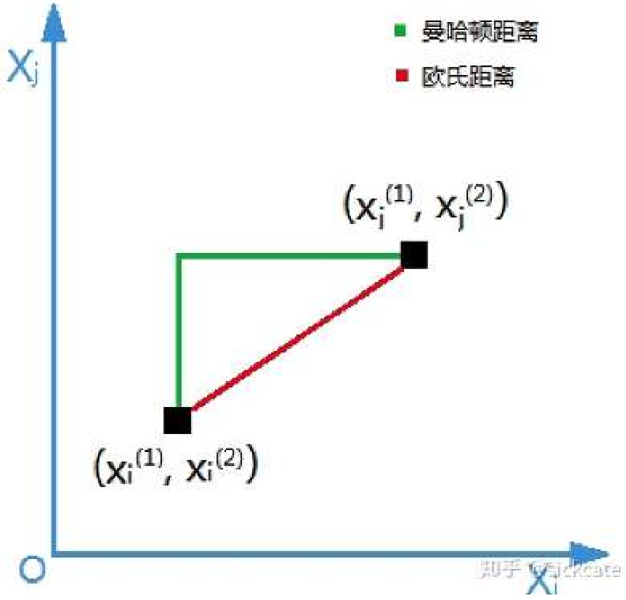
当 $p = 2$ 时称为欧式距离，即 $d_{ij} = \left(\sum_{k=1}^{m} |x_{ki} - x_{kj}|^2 \right)^{\frac{1}{2}}$

当 $p = 1$ 时称为曼哈顿距离，即 $d_{ij} = \sum_{k=1}^{m} |x_{ki} - x_{kj}|$

当 $p = \infty$ 时称为切比雪夫距离，即 $d_{ij} = \max_{k} |x_{ki} - x_{kj}|$

闵科夫斯基距离

Minkowski　distance

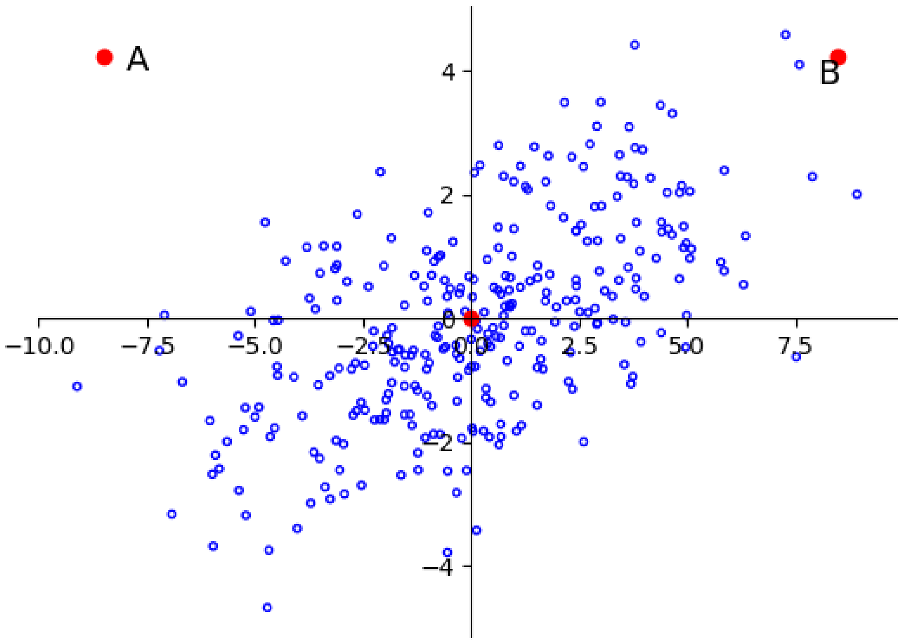


关系：闵氏距离越大相似度越小，距离越小相似度越大。

缺点：1、“距离”的大小与指标的单位有关

2、闵氏距离没有考虑变量间的相关关系

3、没有考虑各个变量的分布（期望等）可能是不同的



A
B
4
2
0
−2
−4
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5

国家和地区	100 米 秒	200 米 秒	400 米 秒	800 米 分	1500 米 分	5000 米 分	10000 米 分	马拉松 分
阿根廷	10.39	20.81	46.84	1.81	3.7	14.04	29.36	137.72
澳大利亚	10.31	20.06	44.84	1.74	3.57	13.28	27.66	128.3
奥地利	10.44	20.81	46.82	1.79	3.6	13.26	27.72	135.9
比利时	10.34	20.68	45.04	1.73	3.6	13.22	27.45	129.95
百慕大	10.28	20.58	45.91	1.8	3.75	14.68	30.55	146.62
巴西	10.22	20.43	45.21	1.73	3.66	13.62	28.62	133.13
缅甸	10.64	21.52	48.3	1.8	3.85	14.45	30.28	139.95
加拿大	10.17	20.22	45.68	1.76	3.63	13.55	28.09	130.15
智利	10.34	20.8	46.2	1.79	3.71	13.61	29.3	134.03
中国	10.51	21.04	47.3	1.81	3.73	13.9	29.13	133.53
哥伦比亚	10.43	21.05	46.1	1.82	3.74	13.49	27.83	131.35

马哈拉诺比斯距离

Mahalanobis distance

马氏距离:（考虑各个分量之间的相关性并与各个分量的尺度无关）

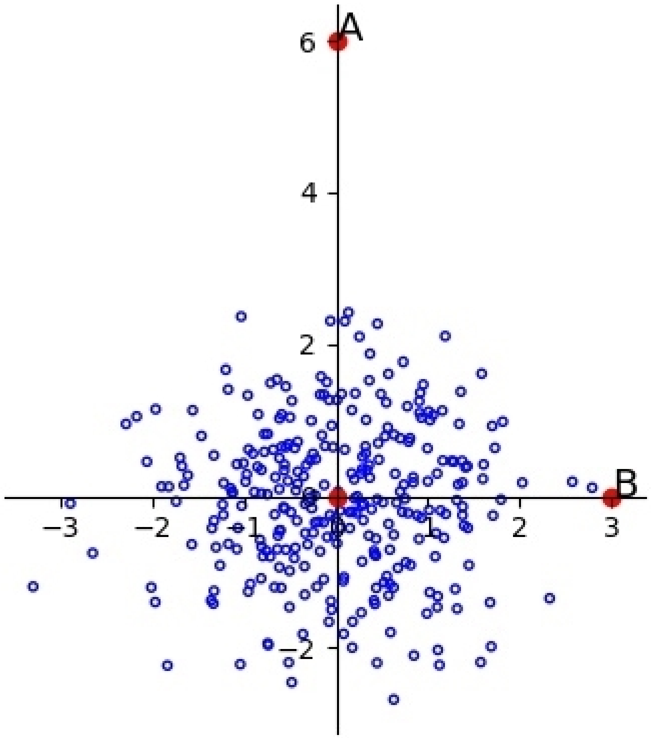
设 X 和 Y 是从均值向量为 μ，协方差阵为 Σ 的总体G中抽取的两个样本，定义 X 和 Y 之间的马氏距离为: $d_m(X,Y)=[(X-Y)'\Sigma^{-1}(X-Y)]^{\frac{1}{2}}$

定义 X 与总体G的马氏距离为: $d_m(X,G)=[(X-\mu)'\Sigma^{-1}(X-\mu)]^{\frac{1}{2}}$

当 Σ 为单位矩阵时，马氏距离就是欧式距离，所以马氏距离是欧式距离的推广。

马氏距离的几何意义

- 将变量按照主成分进行旋转，让维度间相互独立，然后进行标准化，让维度同分布。
- 由主成分分析可知，由于主成分就是特征向量方向，每个方向的方差就是对应的特征值，所以只需要按照特征向量的方向旋转，然后缩放特征值倍。



夹角余弦

Angle cosine

样本 x_i 与样本 x_j 之间的夹角余弦定义为

$$s_{ij} = \frac{\sum_{k=1}^{m} x_{ki} x_{kj}}{\left[\sum_{k=1}^{m} x_{ki}^2 \sum_{k=1}^{m} x_{kj}^2\right]^{\frac{1}{2}}}$$

夹角余弦越接近于1，表示样本越相似；越接近于0，表示样本越不相似。

余弦相似度的特点

- 余弦相似度通常用于正空间，因此给出的值为0到1之间
- 仅仅与向量方向有关，与向量长度无关。
- 对任何维度的向量空间都适用，而且最常用于高维正空间。

余弦相似度的应用

在信息检索中，每个词被赋予不同的维度，而一个文档由一个向量表示，其各个维度上的值对应于该词项在文档中出现的频率，余弦相似度因此可以给出两篇文档在其主题方面的相似度。

另外，它通常用于文本挖掘中的文件比较；在数据挖掘领域中，会用到它来度量集群内部的凝聚力。

相关系数

correlation coefficient

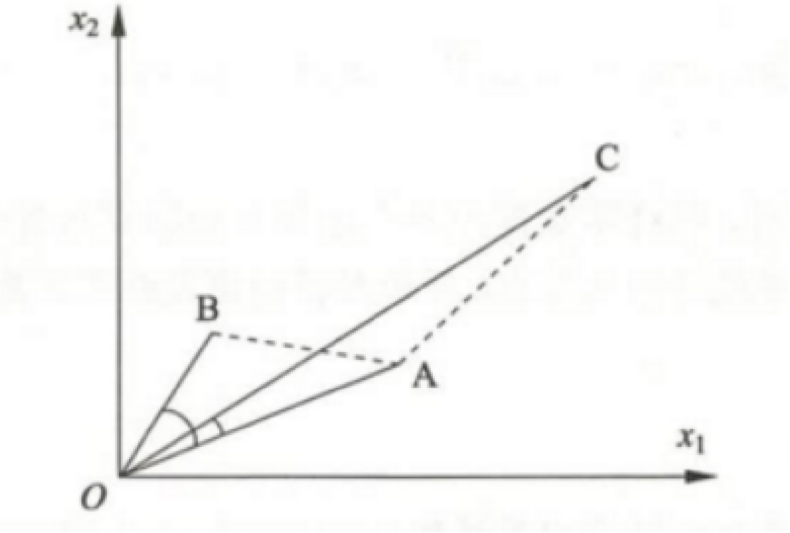
样本 x_i 与样本 x_j 之间的相关系数定义为

$$r_{ij}=\frac{\sum_{k=1}^{m}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\left[\sum_{k=1}^{m}(x_{ki}-\bar{x}_i)^2\sum_{k=1}^{m}(x_{kj}-\bar{x}_j)^2\right]^{\frac{1}{2}}}$$

其中 $x_i=(x_{1i},x_{2i},\cdots,x_{mi})^{\mathrm{T}},\quad x_j=(x_{1j},x_{2j},\cdots,x_{mj})^{\mathrm{T}}$

$$\bar{x}_i=\frac{1}{m}\sum_{k=1}^{m}x_{ki},\quad \bar{x}_j=\frac{1}{m}\sum_{k=1}^{m}x_{kj}$$

相关系数的绝对值越接近于1，表示样本越相似；越接近于0，表示样本越不相似。



从距离的角度看：　A和B比A和C更相似

从夹角余弦的角度看：A和C比A和B更相似