1.1 考虑下表中的事务性数据集：

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

1) 每个事务 ID 对应一条事务，计算{e}，{b, c}，{b, c, e}的支持度。
2) 使用（1）的计算结果，计算关联规则{b, c}→{e}和{e}→{b, c}的置信度。
3) 从（2）的结果看，置信度是对称的吗？请根据计算公式分析其对称性。

1) $s(\{e\}) = \frac{8}{10} = \frac{4}{5}$

$s(\{b,c\}) = \frac{3}{10}$

$s(\{b,c,e\}) = \frac{2}{10} = \frac{1}{5}$

2) $s(\{b,c\} \to \{e\}) = \frac{2}{3}$

$s(\{e\} \to \{b,c\}) = \frac{2}{8} = \frac{1}{4}$

3) 置信度不是对称的.

置信度的计算公式为

$c(X \to Y) = \frac{\delta(X \cup Y)}{\delta(X)}$　　$c(Y \to X) = \frac{\delta(X \cup Y)}{\delta(Y)}$

由于 $\delta(X)$ 和 $\delta(Y)$ 可能不同，故不对称.

1.2 考虑如下二元分类的数据集：（设为 D）

User interest	User occupation	Click
Tech	Professional	1
Fashion	Student	0
Fashion	Professional	0
Sports	Student	0
Tech	Student	1
Tech	Retired	0
Sports	Professional	1

1) 计算分别以属性 User interest 和 User occupation 划分时的信息增益。构建决策树将会选择哪个属性？
2) 计算分别以属性 User interest 和 User occupation 划分时的 Gini 指数。构建决策树将会选择哪个属性？

1) 整体熵：（根据是否点击）

$$Ent(D) = -\sum_{k=1}^{2} p_k \log_2 p_k = -\left(\frac{3}{7}\log_2\frac{3}{7} + \frac{4}{7}\log_2\frac{4}{7}\right) \simeq 0.985$$

以 User interest 为特征进行划分

D^1(User interest = Tech)　D^2(User interest = Fashion)

D^3(User interest) = Sports)

$Ent(D^1) = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) \simeq 0.918$

$Ent(D^2) = -(1\times\log_2 1) = 0$

$Ent(D^3) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$

$$\therefore Gain(D, \text{User interest}) = Ent(D) - \sum_{v=1}^{3}\frac{|D^v|}{|D|}Ent(D^v)$$

$$= 0.985 - \left(\frac{3}{7}\times 0.918 + \frac{2}{7}\times 1\right)$$

$$\simeq 0.306$$

同理 以 User occupation（简写为 U_o）为特征进行划分.

D^1(U_o = Professional)　D^2(U_o = Student)　D^3(U_o = Retired)

$Ent(D^1) = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) \simeq 0.918$

$Ent(D^2) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) \simeq 0.918$

$Ent(D^3) = -(1\log_2 1) = 0$

$\therefore Gain(D, U\_o) = 0.985 - \left(\frac{3}{7}\times 0.918 + \frac{3}{7}\times 0.918\right) \simeq 0.198$

前者的信息增益更大，故应选 User interest.

2) $Gini(D) = 1 - \sum_{k=1}^{K} p_k^2$

属性 a 的基尼指数定义为 $Gini\_index(D, a) = \sum_{v=1}^{V}\frac{|D^v|}{|D|}Gini(D^v)$.

以 User interest 为特征进行划分

$Gini(D^1) = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = \frac{4}{9}$

$Gini(D^2) = 1 - 1^2 = 0$

$Gini(D^3) = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = \frac{1}{2}$

故 $Gini\_index(D, \text{User interest}) = \frac{3}{7}\times\frac{4}{9} + \frac{2}{7}\times\frac{1}{2} = \frac{4}{21} + \frac{1}{7} = \frac{7}{21} = \frac{1}{3}$

以 User occupation 为特征进行划分

$Gini(D^1) = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = \frac{4}{9}$

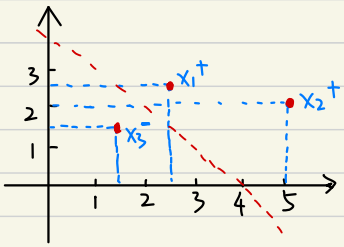
$Gini(D^2) = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = \frac{4}{9}$

$Gini(D^3) = 1 - 1^2 = 0$

故 $Gini\_index(D, \text{User occupation}) = \frac{3}{7}\times\frac{4}{9} + \frac{3}{7}\times\frac{4}{9} = \frac{8}{21}$

前者的 Gini 指数更小，故应选 User interest.

1.3 已知正例点 $x_1=(2.5,2.5)^T$，$x_2=(5,2)^T$，和负例点 $x_3=(1.5,1.5)^T$，试用 SVM 对其进行分类，求最大间隔分离超平面，并指出所有的支持向量。



设超平面方程：$\vec{w}^T\vec{x}+b=0$

引入 Lagrange 乘子 α_i，$i=1,2,3$，得到 Lagrange 函数

$L(\vec{w},b,\vec{\alpha})=\frac{1}{2}\|\vec{w}\|^2-\sum_{i=1}^{3}\alpha_i(y_i(\vec{w}^T\vec{x}_i+b)-1)$

令 $L(\vec{w},b,\vec{\alpha})$ 关于 $\vec{w}$ 和 b 的偏导为 0，得：

$\vec{w}=\sum_{i=1}^{3}\alpha_i y_i x_i$　　$\sum_{i=1}^{3}\alpha_i y_i=0$

回代，即求解

$\min \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\alpha_i\alpha_j y_i y_j(\vec{x}_i^T\cdot\vec{x}_j)-\sum_{i=1}^{3}\alpha_i$

约束条件 $\begin{cases}\alpha_1+\alpha_2-\alpha_3=0\\ \alpha_i\geq 0 \quad i=1,2,3\end{cases}$

对上式展开

$\frac{1}{2}(12.5\alpha_1^2+2\alpha_2^2+4.5\alpha_3^2)+17.5\alpha_1\alpha_2-7.5\alpha_1\alpha_3-10.5\alpha_2\alpha_3-\alpha_1-\alpha_2-\alpha_3$

将 $\alpha_3=\alpha_1+\alpha_2$ 代入上式，得

$\frac{1}{2}(12.5\alpha_1^2+2\alpha_2^2+4.5(\alpha_1+\alpha_2)^2)+17.5\alpha_1\alpha_2-7.5\alpha_1(\alpha_1+\alpha_2)-10.5\alpha_2(\alpha_1+\alpha_2)-2\alpha_1-2\alpha_2$

$=(6.25\alpha_1^2+14.5\alpha_2^2+2.25\alpha_1^2+4.5\alpha_1\alpha_2+2.25\alpha_2^2)+17.5\alpha_1\alpha_2-7.5\alpha_1^2-7.5\alpha_1\alpha_2-10.5\alpha_1\alpha_2-10.5\alpha_2^2)-2\alpha_1-2\alpha_2$

$=\alpha_1^2+6.25\alpha_2^2+4\alpha_1\alpha_2-2\alpha_1-2\alpha_2$　（设为 A）

分别对 α_1，α_2 求偏导，并使其为 0

$\frac{\partial A}{\partial\alpha_1}=2\alpha_1+4\alpha_2-2=0$

$\frac{\partial A}{\partial\alpha_2}=12.5\alpha_2+4\alpha_1-2=0$

$\Rightarrow\begin{cases}\alpha_1=\frac{17}{9}\\ \alpha_2=-\frac{4}{9}\\ \alpha_3=\alpha_1+\alpha_2=\frac{13}{9}\end{cases}$

（若不满足以上条件，令其中一个为 0 即为边界情况）

不满足 $\alpha_i\geq 0$

令 $\alpha_1=0$，$A: 6.25\alpha_2^2-2\alpha_2$　$\alpha_2=\frac{2}{12.5}=\frac{4}{25}$　$\min=-\frac{4}{25}$

令 $\alpha_2=0$，$A: \alpha_1^2-2\alpha_1$　$\alpha_1=1$　$\min=-1$

故取 $\alpha_1=1$，$\alpha_2=0$，$\alpha_3=1$

$\vec{w} = 1 \times 1 \times (2.5, 2.5)^T + 1 \times (-1) \times (1.5, 1.5)^T$

$= (1, 1)^T$

$b = 1 - \vec{w}^T x_1 = 1 - (1, 1)^T (2.5, 2.5) = -4$

∴ 超平面方程： $(1, 1)^T \vec{x} - 4 = 0$

x_1、x_3 对应的 α 均不为0，故 x_1、x_3 为支持向量

问答题见下一页

问答题

2.1

主成分分析的基本流程是什么？与特征值有何关系？

- 基本流程

1. 中心化（均值化），目的是为了方便后面的求解。以二维的情况为例，从协方差矩阵的定义看：`Σ=E{(x-E(x)) * (x-E(x))T}`，PCA的第一步就是要去均值化。求分别求x和y的平均值，然后对于所有的样例，都减去对应的均值。
2. 求特征协方差矩阵。协方差是衡量两个变量同时变化的变化程度。协方差大于0表示x和y若一个增，另一个也增；小于0表示一个增，一个减。如果 x 和 y 是统计独立的，那么二者之间的协方差就是 0；但是协方差是 0，并不能说明 x 和 y 是独立的。协方差绝对值越大，两者对彼此的影响越大，反之越小。
3. 根据协方差矩阵计算特征值和对应的特征向量
4. 将特征值按照从大到小的顺序排序，选择其中最大的k个，然后将其对应的k个特征向量分别作为列向量组成特征向量矩阵
5. 将样本点投影到选取的特征向量上。就将原始样例的n维特征变成了k维，这k维就是原始特征在k维上的投影（样例数为m，特征数为n，减去均值后的样本矩阵为`DataAdjust(m*n)`，协方差矩阵是`n*n`，选取的k个特征向量组成的矩阵为`EigenVectors(n*k)`，则投影后的数据矩阵`FinalData`为`FinalData(m*k) = DataAdjust(m*n矩阵) x 特征向量`）

- 参考周志华老师的《机器学习》，一般化的基本流程可以精炼为如下
    1. 对所有样本进行中心化：$x_i \leftarrow x_i - \frac{1}{m}\sum_{i=1}^{m} x_i$
    2. 计算样本的协方差矩阵XX^T
    3. 对协方差矩阵XX^T作特征值分解
    4. 取最大的K个特征值所对应的特征向量w1、w2 。。。 w_K
- 与特征值的关系
    - 基本流程中需要根据特征值求特征向量，并根据特征值大小进行排序，选取对应的特征向量，通过它们对应的线性组合形成K个新的指标
    - 最大特征值对应的特征向量可以最大化投影方差

2.2

如果从信息检索的视角， 可以将寻找最近邻的过程视作检索最相关的 K 个文档的过程。那么， 这一过程是否可以利用倒排索引的思路加以实现？如何实现？

- 实现：
    1. 将待分类的文本doc表示为其特征词项组成的向量V
    2. 找到将V中每一维对应的词项在倒排索引表中的文本链表
    3. 将这些文本链表合并，去掉重复的文档ID，得到文档ID的集合
    4. 分别计算doc和该集合中文本的相似度（如余弦相似度），取相似度最大的前K个作为K近邻（即最相关的K个文档）
- 这样在查找样本的K个邻居时，只查找与待分类文本的词项有重叠的文档，减少了搜索空间和计算开支，搜索速度更快

- 如果采用聚类的思想，可以将训练文档分为K个簇，用中心向量代表这个簇。通过倒排索引表，找到与查询文本有交集的簇。可以人为规定一个阈值，对于超过该阈值的簇，计算查询向量与簇内文档的相似度。将所有的结果放在一起，取相似度最大的前K个作为K近邻（即最相关的K个文档）

2.3

无论是 K 最近邻分类还是 K 均值聚类，都涉及到 K 的取值问题。请简述两个问题各自选取合适 K 值的思路，并比较两者在思路上有何不同？

- K最近邻分类：
  - K太小，容易受到噪声的干扰，容易发生过拟合（模型过于复杂）
    - 相当于用较小的邻域中的训练样例进行预测分类，“学习”的近似误差会减小，只有与输入样例较近的（相似的）训练实例才会对预测结果起作用。但“学习”的估计误差会增大，预测结果会对近邻的实例点非常敏感。如果邻近的实例点恰巧是噪声，预测分类就会出错。
  - K太大，可能导致错误涵盖其他类别的样本（模型过于简单）
    - 相当于在较大邻域中的训练实例进行预测分类。可以减少学习的估计误差，但学习的近似误差会增大。这是与输入实例较远的（不相似的）训练实例也会对预测起作用，使预测发生错误。
  - K的选取：
    - K一般取一个较小的数值，经验表明K一般小于训练样本数的平方根
    - K是奇数，这样使用投票法能得出分类结果
    - 采用交叉验证法来选取最优的K值
- K均值聚类中K的选取
  - 尝试用不同的K值来聚类，检验各自得到聚类结果的质量，从而推测最优的K值。聚类结果的质量可以用类的平均直径来衡量。一般来说，类别数变小时，平均直径会增加；类别数变大超过某个值以后，平均直径会不变，而这个值正是最优的K值。
  - 实际可以采用二分查找，快速找到最优的K值
  - 也可采用手肘法。手肘法的核心指标是SSE(sum of the squared errors，误差平方和)，SSE是所有样本的聚类误差，代表了聚类效果的好坏。
    - 核心思想：随着聚类数k的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么SSE自然会逐渐变小。并且，当k小于真实聚类数时，由于k的增大会大幅增加每个簇的聚合程度，故SSE的下降幅度会很大，而当k到达真实聚类数时，再增加k所得到的聚合程度回报会迅速变小，所以SSE的下降幅度会骤减，然后随着k值的继续增大而趋于平缓，也就是说SSE和k的关系图是一个手肘的形状，而这个肘部对应的k值就是数据的真实聚类数。当然，这也是该方法被称为手肘法的原因。
- 思路上的不同：
  - K最近邻中采取交叉验证法来获取最优的K值。交叉验证法的基本思想就是将原始数据（dataset）进行分组，一部分作为训练集来训练模型，另一部分作为测试集来评价模型。每次操作并没有用到所有原始数据来训练。目的是为了找到K个样本，满足：这K个样本中类别众数即为待分类样本的实际类别
  - K均值聚类中用不同的K值来聚类，检验各自得到聚类结果的质量，从而推测最优的K值。尝试的K值从小到大，直到找到恰好使平均直径基本不变的拐点，对应的K值即为最优值。每次操作都用到所有原始数据来训练。目的是为了将样本划分为K个簇，簇内越相同越好，簇间差别越大越好，即尽可能按照实际类别划分出样本的类别

2.4

K-mediods 算法描述:
a) 首先随机选取一组聚类样本作为中心点集
b) 每个中心点对应一个簇
c) 计算各样本到各个中心点的距离(如欧几里得距离)，将样本点放入距离中心点最短的那个簇中
d) 计算各簇中，距簇内各样本点距离的绝对误差最小的点，作为新的中心点
e) 如果新的中心点集和原中心点集相同，算法中止；如果新的中心点集与原中心点集不完全相同，返回b)
试着:
a) 阐述 K-mediods 算法和 K-means 算法相同的缺陷
b) 阐述 K-mediods 算法相比于 K-means 算法的优势
c) 阐述 K-mediods 算法相比于 K-means 算法的不足

- 两种算法的主要区别：中心点的选取，在K-means中，将中心点取为当前cluster中所有数据点的平均值，在 K-medoids算法中，将从当前cluster 中选取这样一个点——它到其他所有（当前cluster中的）点的距离之和最小——作为中心点（也就是说K medoids的中心点一定是数据集中存在的点）。
- a)
  - 初始聚类中心的选择对聚类结果都有较大的影响
  - 都有可能陷入局部最优解的困境之中
  - K的含义相同，都需要开始人为设定簇数目，且K值的选定不是很容易，需要慎重
  - 都是无监督算法，结果不一定具有可解释性
- b)
  - k-medoids对噪声和孤立点的鲁棒性比较好，对极值噪声不是特别敏感。例：当一个cluster样本点只有少数几个，如（1,1）（1,2）（2,1）（100,100）。其中（100,100）是噪声。如果按照k-means质心大致会处在（1,1）（100,100）中间，这显然不是我们想要的。这时k-medoids就可以避免这种情况，他会在（1,1）（1,2）（2,1）（100,100）中选出一个样本点使cluster的绝对误差最小，计算可知一定会在前三个点中选取。
  - K-means只适用于数值属性聚类(均值有实际意义)；K-mediods适用范围更广，还适用类别类型的特征
  - K-mediods每次选取的都是实际存在的样本点，不会出现空簇；K-means选取的点可能不对应实际的样本，可能会出现空簇
- c)
  - k-medoids的运行速度较慢，计算质心的步骤时间复杂度是`O(n^2*K*t)`，因为它必须计算任意两点之间的距离。而k-means只需平均即可，时间复杂度为`O(n*K*t)`（t为迭代次数）
  - 只能对小样本起作用，样本一大，它的速度就太慢了，而且当样本多的时候，少数几个噪音对k-means的质心影响也没有想象中的那么重，所以k-means的应用明显比k-medoids更广泛

参考

《统计学习方法》 李航

《机器学习》 周志华

https://blog.csdn.net/databatman/article/details/50445561

https://www.cnblogs.com/190260995xixi/p/5954921.html

https://blog.csdn.net/qq_15738501/article/details/79036255（手肘法）