Web作业3

PB18000268 曾勇程 HW3

计算题

1.1 1) 支持度：

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

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

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

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

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

由于 $c(\{b,c\} \to \{e\}) \neq c(\{e\} \to \{b,c\})$

∴ 置信度不是一个对称的度量，即不对称

也可根据公式：$c(X \to Y) = \frac{|X,Y|}{|X|}$

$$c(Y \to X) = \frac{|X,Y|}{|Y|}$$

两者不一定相等（$|X|$ 不一定等于 $|Y|$）

∴ 置信度不对称

1.2 1) 首先，计算整体的熵：

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

$$\approx 0.985$$

特征 User interest：

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

D^3(User interest = Sports)

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

$Ent(D^2) = 0$

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

$\therefore Gain(D, \text{User interest}) = 0.985 - (\frac{3}{7}\times 0.918 + \frac{2}{7}\times 1) = 0.306$

D^6(User occupation = Professional)

D^5(User occupation = Student)

D^4(User occupation = Retired)

$Ent(D^4) = -(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}) = 0.918$

$Ent(D^5) = 0.918$

$Ent(D^6) = 0$

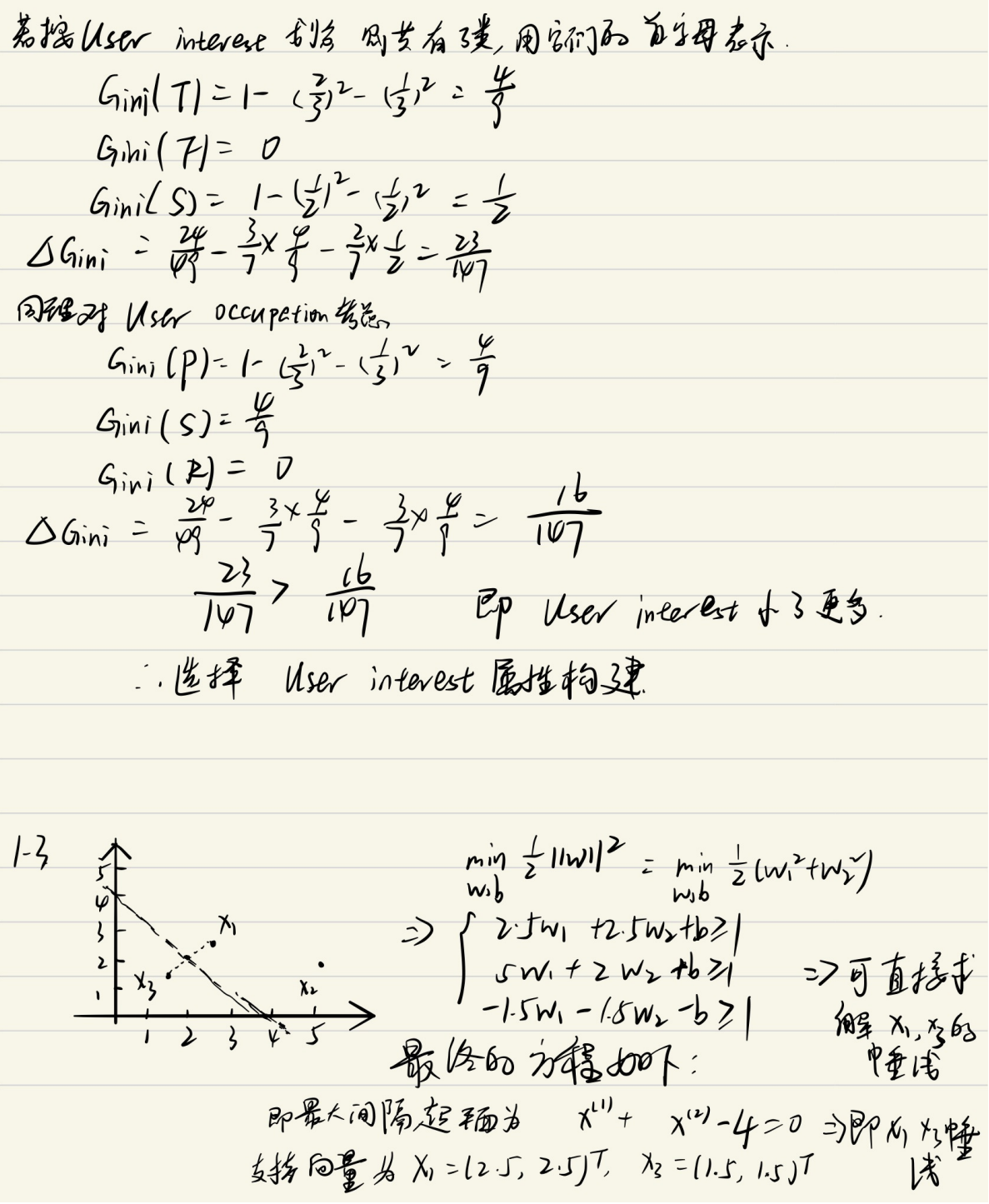
$\therefore Gain(D, \text{User occupation}) = 0.985 - (\frac{3}{7}\times 0.918 + \frac{3}{7}\times 0.918)$

$= 0.198$

$0.306 > 0.198$

∴ 选择 User interest 属性构建.

2) 首先计算整体的 Gini 值 $G = 1 - (\frac{3}{7})^2 - (\frac{4}{7})^2 = \frac{24}{49}$



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?

答：PCA的思路是通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量，转换后的这组变量叫主成分。通过这种方式，可以采用较少的综合指标综合先前存在于各个属性 （且相关）中的各类信息，而综合指标之间彼此不相关。

获取主成分的过程，同时也是降维的过程。对于二维的实例，如果只有两维特征，我们可以将其表示为一个椭圆，椭圆有长轴短轴，相比之下，短轴方向的数据变化较少，区分度偏低，在这种情况下，我们删去短轴，再将坐标轴变换与长轴平行。对于高维椭球而言，思路是类似的，即找出主轴与几个最长的轴作为新维度。

与特征值有何关系?

特征值的大小代表了矩阵正交化之后所对应特征向量对于整个矩阵的贡献程度。

最大特征值对应的特征向量可以最大化投影方差，K个特征值的比重反应了主成分的信息量，一般应大于0.85。

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?

答：先将所有文档中相似度较大的文档集合合并，称为一个簇，并计算簇的中心向量。把每一簇看作一个整体，簇的中心向量看作这个整体的表征向量，建立从词条到簇的倒排索引。利用词条到簇的倒排索引找到与**检索文档**有交集的簇，并计算待分类文本和簇的**中心向量**的相似度，得到相似度最高的`m`个簇，然后计算**检索文档**与这`m`个簇中每个文档的相似度，得到相似度最高的`K`个文档，实现最近邻的寻找。由于簇的数目小于文档的数目，在一定程度上减少了相似度计算次数，提高了寻找效率。

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?

答：对于 K 最近邻分类(监督学习)：K值一般取一个比较小的数值，然后通过交叉验证法来选择最优的K值。例如：从K=1开始，使用检验集估计分类器的误差率。重复该过程，每次K增值1，允许增加一个近邻。选取产生最小误差率的K。

对于 K 均值聚类(无监督学习)：

1. 对于 K 较小时，可以采少数样本，借助其他聚类（如层次聚类）先确定出初始中心；
2. 选择多于K个的中心，然后从中挑选分隔较为明显的，使用“后处理”来修补生成的簇(如清除较小的、可能代表离群点的簇)；
3. 采用二分K均值方法来确定 K 的取值：为了得到K个簇，先分为 2 个簇，然后不断选择其中一个分裂 ，选择的标准可以是样本数较大的簇，或者SSE较高的簇。
4. 当簇存在不同规模、密度及不规则形状的情况下，可以采用偏大的K，然后进行合并。

思路上的不同：对于 K 最近邻分类的 K 的选取，是在有监督情况下的选取，选取到最适的 K 使得误差率最小。对于每次训练，K值确定后每次结果固定。

对于 K 均值聚类， K 的选取是在无监督的情况下的选取，每次的 K 值的选取可能不同，初始的 K 个中心的选取的随机性较大，不同的初始中心，可能导致截然不同的聚类结果。聚类的依据有样本到簇中心的距离的大小（如根据 SSE）。

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) 计算各簇种，据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

a)缺陷：都是随机选取初始中心，并且要事先确定簇数。不同的初始中心，可能导致截然不同的聚类结果。当簇存在不同规模、密度及不规则形状的情况下，聚类效果较差。一般获得的是局部最优解，而非全局最优解。

b)优势：1. K-means算法使用Means（均值）作为聚点，容易受到离群点（孤立点和噪声数据）的干扰，对离群点敏感；K-mediods算法使用Mediods(中位数)作为聚点，即它选择簇中位置最接近簇中心的对象(称为中心点)作为簇的代表点，有助于消除这种敏感性。2. K-Mediods算法对大规模数据性能更好。

c)不足：K-mediods算法由于按照中心点选择的方式进行计算，算法的时间复杂度也比K-means算法上升了O(n)。