

第二章 作业

1. 假设所分析的数据包括属性 Power (W)，它在数据元组中的值（以递增序）为 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70。

- (1) 该数据的均值是什么？中位数是什么？
- (2) 该数据的众数是什么？
- (3) 该数据的中列数是什么？
- (4) 该数据的第一个四分位数 (Q_1) 和第三个四分位数 (Q_3) 是什么？
- (5) 绘制该数据的箱线图。

2. 现有两个随机变量 X 和 Y，有以下样本：

序号	X	Y
1	153	44
2	181	64
3	170	70
4	172	57
5	174	61
6	168	67
7	189	84

试分别求 X 和 Y 的期望、方差以及两者的协方差。

3. 给定两组数据 $X=[1,7,7]$, $Y=[3,9,7]$ ，分别计算 X 与 Y 间的欧氏距离，曼哈顿距离，切比雪夫距离和马氏距离。

4. 简要概述如何计算如下属性描述对象的相异性：

- (1) 标称属性。
- (2) 非对称的二元属性。
- (3) 数值属性。

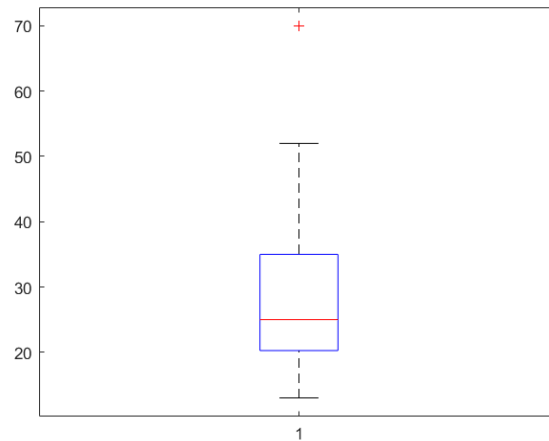
5. 简述协方差和皮尔逊相关系数的区别。

注：第 1、2、3 题可另进行编程实现。

第一次作业答案

Answer1:

- (1) 均值: 29.963 中位数: 25
- (2) 众数: 25 和 35
- (3) 41.5
- (4) 第一个四分位: 20 第三个四分位: 35
- (5) 一个分布的五个数汇总由最小值、第一个四分位数、中间值、第三个四分位数和最大值组成。这个数据是: 13, 20, 25, 35, 70



Answer2:

$$E(X) = \frac{(153 + 181 + \dots + 189)}{7} \approx 172.43$$

$$E(Y) = \frac{(44 + 64 + \dots + 84)}{7} \approx 63.86$$

$$\sigma(X) = \frac{\sum_{i=1}^7 (X_i - E(X))^2}{7} \approx 107.67$$

$$\sigma(Y) = \frac{\sum_{i=1}^7 (Y_i - E(Y))^2}{7} \approx 128.98$$

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^7 (X_i - E(X))(Y_i - E(Y))}{7 - 1} \approx 115.07$$

Answer3:

$$(1) D_p = \sqrt{\sum_{i=1}^3 (x_i - y_i)^2} = 2\sqrt{2}$$

$$(2) D_M = \sum_{i=1}^3 |x_i - y_i| = 4$$

$$(3) D_C = \max_{i=1,2,3} (|x_i - y_i|) = 2$$

(4) 由于没有给出数据的协方差，无法计算马氏距离

Answer4:

$$(1) D_R = \frac{p-m}{p}, \text{ 其中 } m \text{ 为对象相同的属性数, } p \text{ 为对象的属性总数}$$

$$(2) D_J = \frac{q}{q+r+s}, \text{ 其中 } q \text{ 是样本 } x \text{ 和 } y \text{ 都取 } 1 \text{ 的属性数; } r \text{ 是样本 } x \text{ 中取 } 1, \text{ 在}$$

样本 y 中取 0 的属性数; s 是在样本 x 中取 0, 在样本 y 中取 1 的属性数

(3) 可用欧式距离、曼哈顿距离、切比雪夫距离、闵可夫斯基距离等距离来描述

Answer5:

定义不同、取值不同、对单位敏感程度不同

作业 1.1

1. MATLAB 代码:

```
clc
clear all

% 给定数据
data = [13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70];

% 计算均值和中位数
mean_value = mean(data);
median_value = median(data);

% 计算众数
[counts, values] = hist(data, unique(data));
mode_value = values(counts == max(counts));

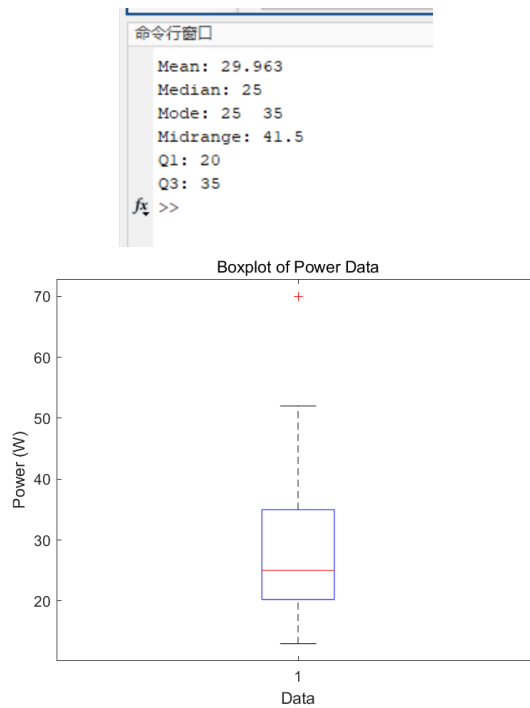
% 计算中列数
midrange_value = (max(data) + min(data)) / 2;

% 计算四分位数
sorted_data = sort(data);
n = length(sorted_data);
q1 = sorted_data(ceil(0.25 * n));
q3 = sorted_data(ceil(0.75 * n));

% 绘制箱线图
figure;
boxplot(data);
title('Boxplot of Power Data');
xlabel('Data');
ylabel('Power (W)');

% 显示计算结果
disp(['Mean: ', num2str(mean_value)]);
disp(['Median: ', num2str(median_value)]);
disp(['Mode: ', num2str(mode_value)]);
disp(['Midrange: ', num2str(midrange_value)]);
disp(['Q1: ', num2str(q1)]);
disp(['Q3: ', num2str(q3)]);
```

2. 运行结果:



作业 1.2

1. MATLAB 代码:

```
clc;
```

```
clear all;
```

```
% 给定样本数据
```

```
X = [153, 181, 170, 172, 174, 168, 189];
```

```
Y = [44, 64, 70, 57, 61, 67, 84];
```

```
% 计算期望
```

```
mean_X = mean(X);
```

```
mean_Y = mean(Y);
```

```
% 计算方差
```

```
var_X = sum((X - mean_X).^2) / length(X);
```

```
var_Y = sum((Y - mean_Y).^2) / length(Y);
```

```
% 计算协方差
```

```
covariance = sum((X - mean_X) .* (Y - mean_Y)) / (length(X) - 1);
```

```
% 显示计算结果
```

```
disp(['期望 E(X): ', num2str(mean_X)]);
```

```
disp(['期望 E(Y): ', num2str(mean_Y)]);
```

2. 运行结果:

```
命令窗口
期望 E(X): 172.4286
期望 E(Y): 63.8571
方差 Var(X): 107.6735
方差 Var(Y): 128.9796
协方差 Cov(X, Y): 115.0714
fx >>
```

作业 1.3

1.MATLAB 代码:

```
clc
clear all;

X = [1, 7, 7];
Y = [3, 9, 7];

% 欧氏距离
euclidean_dist = norm(X - Y);

% 曼哈顿距离
manhattan_dist = sum(abs(X - Y));

% 切比雪夫距离
chebyshev_dist = max(abs(X - Y));

% 显示计算结果
disp(['欧氏距离: ', num2str(euclidean_dist)]);
disp(['曼哈顿距离: ', num2str(manhattan_dist)]);
disp(['切比雪夫距离: ', num2str(chebyshev_dist)]);

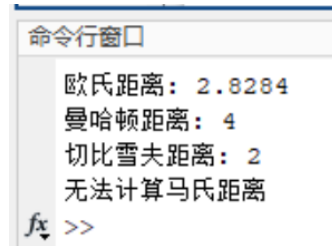
try
    % 将 X 和 Y 组成一个集合
    data = [X; Y];

    % 计算协方差矩阵
    cov_matrix = cov(data);

    % 检查协方差矩阵是否对称正定
    [~, p] = cholcov(cov_matrix);
    if p == 0
        % 计算马氏距离
        mahalanobis_dist = mahal(X, Y, cov_matrix);
        disp(['马氏距离: ', num2str(mahalanobis_dist)]);
    else
        disp('无法计算马氏距离，协方差矩阵不满足要求');
```

```
end  
catch ME  
    disp('无法计算马氏距离');  
end
```

2.运行结果:



A screenshot of the MATLAB Command Window titled '命令行窗口'. It displays the following output:

- 欧氏距离: 2.8284
- 曼哈顿距离: 4
- 切比雪夫距离: 2
- 无法计算马氏距离

The prompt 'fx >>' is visible at the bottom left of the window.

第三章 作业

时间	风速 (m/s)	功率(MW)
01/01/2016 03:00:00	8.94967	35.971
01/01/2016 03:15:00	8.58567	34.51933
01/01/2016 03:30:00	9.20167m/s	38.435
01/01/2016 03:45:00	9.67667	41.13733
01/01/2016 04:00:00	10.10067	42.351
01/01/2016 04:15:00	9.77767	40.66067
01/01/2016 04:30:00	10.34333	42.91667
01/01/2016 04:45:00	NAN	43.41533
01/01/2016 05:00:00	11.24233	43.847
01/01/2016 05:15:00	11.75767	46.13233
01/01/2016 05:30:00	12.16167	48.19467
01/01/2016 05:45:00	11.20233	46.06533
01/01/2016 06:00:00	10.52533	44.60667
01/01/2016 06:15:00	10.99967	45.14967
01/01/2016 06:30:00	10.94933	45.66333
01/01/2016 06:45:00	10.677	45.00834
01/01/2016 07:00:00	0	-1000
01/01/2016 07:15:00	9.99967	42.41767
01/01/2016 07:30:00	9.63633	42.24667
01/01/2016 07:45:00	8.97967	38.286
01/01/2016 08:00:00	8.48467	33.403

表 1 山西某风电场实测风速和实测发电功率数据

1. 表 1 中的数据，如果没有经过数据预处理就进行数据挖掘的话，会有哪些问题？简述数据预处理的意义和步骤。
2. 请对表 1 中的数据进行数据清洗，说明数据清洗步骤。
3. 请使用表 1 中的数据，回答以下问题：
 - (a) 使用最小-最大规范化将风速和功率值变换到[0.0, 1.0]区间。
 - (b) 使用 z 分数规范化变换风速 10.677 m/s 和功率值 42.351 MW。
 - (c) 使用小数定标规范化变换功率值 35.971 MW。
 - (d) 指出对于给定数据，你愿意使用哪种方法，陈述你的理由。

4. 不考虑时序特征，试采用分箱法分别对表 1 中的风速数据和功率数据进行平滑去噪。
5. 对一个 5×2 的二维数据矩阵 X 进行主成分分析，累计方差百分比阈值为 0.8，并用散点图对结果进行可视化。

$$X = \begin{bmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{bmatrix}^T$$

注意：作业要详细说明计算过程，可编程辅助完成。

第二次作业答案

Answer1:

(1) 存在的问题：无法保证数据挖掘的结果的有效性。

(2) 数据预处理主要包括数据清洗、数据集成、数据变换、数据归约等内容。数据清洗:负责解决填充空缺值、识别孤立点、去掉噪声和无关数据等问题;

数据集成:负责解决不同数据源的数据匹配问题、数值冲突问题和冗余问题;

数据变换:将原始数据转换为适合数据挖掘的形式。包括数据的汇总、聚集、概化、规范化,同时可能需要对属性进行重构;

数据归约:负责缩小数据的取值范围,使其更适合数据挖掘算法的需要。

Answer2:

缺失值处理: 原始数据中可能会出现数据值缺失,即数据集中存在无数据的数据单元
格

01/01/2016 04:30:00	10.34333	42.91667
01/01/2016 04:45:00	NAN	43.41533
01/01/2016 05:00:00	11.24233	43.847

其中 NAN 表示数据缺失,此处可采用前后数据的平均值填补,即 10.7928 或者其他变量的平均值填充,即 10.1711

一致化处理: 数据集中会存在某一个数据列的数据至标准不一致或命名规则不一致的情况

01/01/2016 03:30:00	9.20167m/s	38.435
---------------------	------------	--------

参考整体的命名规则,此处需要删除单位符号

异常值处理:

01/01/2016 07:00:00	0	-1000
---------------------	---	-------

此时的异常值可以采用直接删除操作

Answer3:

(1) 归一化

0.12646	0.17361
0.02747	0.07547
0.19500	0.34019
0.32418	0.52288

0.43949	0.60494
0.35165	0.49066
0.50548	0.64318
0.62772	0.67689
0.74998	0.70607
0.89013	0.86057
1.00000	1.00000
0.73910	0.85604
0.55498	0.75743
0.68398	0.79414
0.67029	0.82887
0.59623	0.78459
0.41202	0.60944
0.31321	0.59788
0.13462	0.33012
0.00000	0.00000

(2)

风速: $\mu_A = 10.2021, \sigma_A = 1.0345$

$$v' = \frac{v - \mu_A}{\sigma_A} = 0.45897$$

功率: $\mu_B = 42.0214, \sigma_B = 3.95617$

$$w' = \frac{w - \mu_B}{\sigma_B} = 0.083312$$

(3)

$$v' = \frac{v}{10^j} = 0.359, j = 2$$

(4) z 变换。理由: 求解简化了, 同时相比于最小-最大标准法, 它能通过将 z 变换后的数与 0 比较直观判断与其均值的关系。

Answer4:

排序结果

风速	功率
8.48467	33.40300

8.58567	34.51933
8.94967	35.97100
8.97967	38.28600
9.20167	38.43500
9.63633	40.66067
9.67667	41.13733
9.77767	42.24667
9.99967	42.35100
10.10067	42.41767
10.34333	42.91667
10.52533	43.41533
10.67700	43.84700
10.79280	44.60667
10.94933	45.00834
10.99967	45.14967
11.20233	45.66333
11.24233	46.06533
11.75767	46.13233
12.16167	48.19467

均值平滑后

风速	功率
8.84027	36.12287
8.84027	36.12287
8.84027	36.12287
8.84027	36.12287
8.84027	36.12287
9.83820	41.76267
9.83820	41.76267
9.83820	41.76267
9.83820	41.76267
9.83820	41.76267
10.65756	43.95880
10.65756	43.95880
10.65756	43.95880

10.65756	43.95880
10.65756	43.95880
11.47273	46.24107
11.47273	46.24107
11.47273	46.24107
11.47273	46.24107
11.47273	46.24107

Answer5:

首先求取协方差矩阵

$$\Sigma = X^T X = \begin{bmatrix} 6 & 4 \\ 4 & 6 \end{bmatrix}$$

对协方差矩阵进行特征根分解，可以得到

$$\begin{pmatrix} 6 & 4 \\ 4 & 6 \end{pmatrix} \begin{pmatrix} -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} = \begin{pmatrix} 10 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}$$

可知，协方差阵的特征值为 $\lambda_1 = 10, \lambda_2 = 2$ ，对应的特征向量为 $c_1 = (-\frac{\sqrt{2}}{2} \quad -\frac{\sqrt{2}}{2})^T$

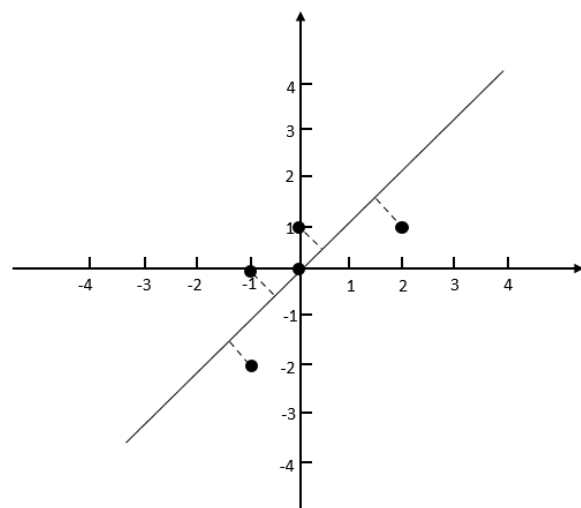
和 $c_2 = (-\frac{\sqrt{2}}{2} \quad \frac{\sqrt{2}}{2})^T$ 。选取累计方差百分比 CPV 的阈值 threshold=0.8，则可知

$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{5}{6} > 0.8$ ，因此在此条件下只用保留第一个主元即可，其对应的负值向量

$p_1 = c_1 = (-\frac{\sqrt{2}}{2} \quad -\frac{\sqrt{2}}{2})^T$ ，因此可以得到降维后的主成分。

$$\begin{aligned} t_1 &= X p_1 = \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}^T \begin{pmatrix} -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix}^T = \begin{pmatrix} \frac{3\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 & -\frac{3\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix}^T \\ &= \begin{pmatrix} \frac{3\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 & -\frac{3\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix} \end{aligned}$$

可视化后的结果图如图所示



第四章 作业

- 1. 给定一个规则 $I_1 \rightarrow I_2$ ，给出该规则的置信度定义，并解释它的含义。
- 2. 简述 Apriori 算法和 FP-tree 算法流程，比较两个算法的优缺点。
- 3. 设最小支持度为 33.3%，最小置信度为 50%。按照 Apriori 算法的步骤，给出每次扫描表 3-1 中的数据库后得到的所有频繁项集。在频繁项集的基础上，产生所有的强关联规则。

表 3-1

TID	故障类型
1	A, B, C, D, E
2	A, B, D, E
3	B, C, D
4	C, D, E
5	A, C, E
6	A, B, D

- 4. 某商店统计了上个季度 10000 笔交易记录，给出如表 3-2 所示的统计信息：

表 3-2

1. {牙刷}在 6000 个事务中出现；
2. {防晒霜}在 5000 个事务中出现；
3. {凉鞋}在 4000 个事务中出现；
4. {太阳镜}在 2000 个事务中出现；
5. {牙刷，防晒霜}在 1500 个事务中出现；
6. {牙刷，凉鞋}在 1000 个事务中出现；
7. {牙刷，太阳镜}在 250 个事务中出现；
8. {牙刷，防晒霜，凉鞋}在 600 个事务中出现。

回答如下问题：

（1）规则“牙刷→防晒霜”与“{牙刷，防晒霜}→{凉鞋}”的置信度分别是多少？

(2) {牙刷}和{防晒霜}是独立的吗?

(3) 计算 $Lift(\text{牙刷}, \text{太阳镜})$ 。

5. 给定如表 3-3 所示的一个事务数据库,画出 FP-tree 树的生成过程。

表 3-3

TID	Itemset
1	a, b, c
2	b, c, d, e
3	a, c, e
4	b, c, d
5	b, c, d, e

第三次作业答案

Answer1

置信度(confidence): 判断关联规则是否为强关联规则的指标, 关联规则的置信度越高则表明该关联规则越值得信任, 一般通过条件概率计算关联规则 $X \rightarrow Y$ 的置信度:

$$\text{Conf}(I_1 \rightarrow I_2) = P(I_1|I_2) = \frac{\text{Sup}(I_1 \cup I_2)}{\text{Sup}(I_1)}$$

其中 I_1 和 I_2 是频繁项集, $\text{Sup}(I_1 \cup I_2)$ 是包含 I_1 和 I_2 的事务数, 而 $\text{Sup}(I_1)$ 是包含项集 I_1 的事务数, $\text{Conf}(I_1 \rightarrow I_2)$ 是同时包含 I_1 和 I_2 的事务数量占包含 I_1 的事务数量的百分比, 该式在已知频繁项集时可直接用于发现强关联规则。

Answer2

Apriori 算法: 首先从数据库删除支持度少于 minsup 的项, 得到频繁1-项集的集合 L_1 ; 接着将 L_1 与自身进行连接产生长度为2的频繁项集的候选集 C_2 , 通过对其进行剪枝操作, 删除 C_2 中支持度少于 minsup 的项集得到 L_2 ; 迭代并重复该过程, 直到无法再生产新的频繁项集。在每轮迭代过程中, 新的频繁模式候选项集都由前一次迭代挖掘的频繁项集的集合连接自身产生。

优点:

- 1) 采用逐层搜索的迭代方法, 算法简单明了, 易于实现。
- 2) 适合稀疏数据集的关联规则挖掘, 也就是频繁项目集的长度稍小的数据集。

缺点:

- 1) 对数据库的扫描次数过多。
- 2) 可能产生大量的候选项集。
- 3) 在频繁项目集长度变大的情况下, 运算时间显著增加, 不适用于大数据集。

FP-Growth 算法: 首先扫描数据库得到所有频繁项并按支持度降序排列; 之后以空节点作为根节点建立FP树: 对数据库中每个事务创建分支, 其中沿共同前缀的每个结点支持度计数加一, 为前缀之后的项创建结点和链接; 最后以项头表的结构自底依次向上地遍历FP树, 根据频繁项寻找对应的条件模式基和条件FP树, 递归挖掘频繁项集, 直到FP树中没有元素为止。

优点:

- 1) 与Apriori算法相比, FP-Growth算法无需重复扫描数据库, 搜索空间的大小得到了显著压缩, 因此该算法对于长频繁模式的挖掘具有较高的效率。
- 2) 不需要生成候选集。

缺点:

- 1) 内存开销大。
- 2) 只能用于挖掘单维的布尔关联规则。

Answer3

已知最小支持度为 2。扫描数据库得到所有的频繁项集。

候选 1-项集 C1

项集	支持度计数
{A}	4
{B}	4
{C}	4
{D}	5
{E}	4

频繁 1-项集 L1

项集	支持度计数
{A}	4
{B}	4
{C}	4
{D}	5
{E}	4

候选 2-项集 C2

项集	支持度计数	项集	支持度计数
{A,B}	3	{B,D}	4
{A,C}	2	{B,E}	2
{A,D}	3	{C,D}	3
{A,E}	3	{C,E}	3
{B,C}	2	{D,E}	3

频繁 2-项集 L2

项集	支持度计数	项集	支持度计数
{A,B}	3	{B,D}	4
{A,C}	2	{B,E}	2
{A,D}	3	{C,D}	3
{A,E}	3	{C,E}	3
{B,C}	2	{D,E}	3

候选 3-项集 C3

项集	支持度计数	项集	支持度计数
{A,B,C}	1	{A,D,E}	2
{A,B,D}	3	{B,C,D}	2
{A,B,E}	2	{B,C,E}	1
{A,C,D}	1	{B,D,E}	2
{A,C,E}	2	{C,D,E}	2

频繁 3-项集 L3

项集	支持度计数	项集	支持度计数
{A,B,D}	3	{B,D,E}	2
{A,B,E}	2	{C,D,E}	2

{A,C,E}	2		
{A,D,E}	2		
{B,C,D}	2		

候选 4-项集 C4

项集	支持度计数
{A,B,D,E}	2

频繁 4-项集 L4

项集	支持度计数
{A,B,D,E}	2

根据所有频繁项集产生所有的强关联规则（最小置信度为 50%）：

L2 产生的所有强关联规则

关联规则	置信度	关联规则	置信度
{A} → {B}	75%	{B} → {D}	100%
{B} → {A}	75%	{D} → {B}	80%
{A} → {C}	50%	{B} → {E}	50%
{C} → {A}	50%	{E} → {B}	50%
{A} → {D}	75%	{C} → {D}	75%
{D} → {A}	60%	{D} → {C}	60%
{A} → {E}	75%	{C} → {E}	75%
{E} → {A}	75%	{E} → {C}	75%
{B} → {C}	50%	{D} → {E}	60%
{C} → {B}	50%	{E} → {D}	75%

L3 产生的所有强关联规则

关联规则	置信度	关联规则	置信度	关联规则	置信度	关联规则	置信度
{A} → {B,D}	75%	{A} → {C,E}	50%	{B} → {C,D}	50%	{C} → {D,E}	50%
{B,D} → {A}	75%	{C,E} → {A}	66.7%	{C,D} → {B}	66.7%	{D,E} → {C}	66.7%
{B} → {A,D}	75%	{C} → {A,E}	50%	{C} → {B,D}	50%	{D} → {C,E}	40%
{A,D} → {B}	100%	{A,E} → {C}	66.7%	{B,D} → {C}	50%	{C,E} → {D}	66.7%
{D} → {A,B}	60%	{E} → {A,C}	50%	{D} → {B,C}	40%	{E} → {C,D}	50%
{A,B} → {D}	100%	{A,C} → {E}	100%	{B,C} → {D}	100	{C,D} → {E}	66.7%
关联规则	置信度	关联规则	置信度	关联规则	置信度		
{A} → {B,E}	50%	{A} → {D,E}	50%	{B} → {D,E}	50%		
{B,E} → {A}	100%	{D,E} → {A}	66.7%	{D,E} → {B}	66.7%		
{B} → {A,E}	50%	{D} → {A,E}	40%	{D} → {B,E}	40%		
{A,E} → {B}	66.7%	{A,E} → {D}	66.7%	{B,E} → {D}	100%		
{E} → {A,B}	50%	{E} → {A,D}	50%	{E} → {B,D}	50%		
{A,B} → {E}	66.7%	{A,D} → {E}	66.7%	{B,D} → {E}	50%		

L4 产生的所有强关联规则

关联规则	置信度	关联规则	置信度
{A} → {B,D,E}	50%	{A,B} → {D,E}	66.7%
{B,D,E} → {A}	100%	{D,E} → {A,B}	66.7%
{B} → {A,D,E}	50%	{A,D} → {B,E}	66.7%
{A,D,E} → {B}	100%	{B,E} → {A,D}	100%

$\{D\} \rightarrow \{A,B,E\}$	40%	$\{A,E\} \rightarrow \{B,D\}$	66.7%
$\{A,B,E\} \rightarrow \{D\}$	100%	$\{B,D\} \rightarrow \{A,E\}$	50%
$\{E\} \rightarrow \{A,B,D\}$	50%		
$\{A,B,D\} \rightarrow \{E\}$	66.7%		

表中除 $\{D\} \rightarrow \{A,E\}$, $\{D\} \rightarrow \{B,C\}$, $\{D\} \rightarrow \{B,E\}$, $\{D\} \rightarrow \{C,E\}$, $\{D\} \rightarrow \{A,B,E\}$ 外, 均为强关联规则。

Answer4

(1) $\text{con}(\{\text{牙刷}\} \rightarrow \{\text{防晒霜}\}) = \sigma(\{\text{牙刷}, \text{防晒霜}\}) / \sigma(\{\text{牙刷}\}) = 1500/6000 = 25\%$

$\text{con}(\{\text{牙刷}, \text{防晒霜}\} \rightarrow \{\text{凉鞋}\}) = \sigma(\{\text{牙刷}, \text{防晒霜}, \text{凉鞋}\}) / \sigma(\{\text{牙刷}, \text{防晒霜}\}) = 600/1500 = 40\%$

(2) $I(\{\text{牙刷}\} \rightarrow \{\text{防晒霜}\}) = \text{sup}(\{\text{牙刷}\} \rightarrow \{\text{防晒霜}\}) / (\text{sup}(\{\text{牙刷}\}) \times \text{sup}(\{\text{防晒霜}\})) = (N \times \sigma(\{\text{牙刷}\} \rightarrow \{\text{防晒霜}\})) / (\sigma(\{\text{牙刷}\}) \times \sigma(\{\text{防晒霜}\})) = (10000 \times 1500) / (6000 \times 5000) = 0.5 < 1$

所以 $\{\text{牙刷}\}$ 和 $\{\text{防晒霜}\}$ 是负相关。

(3) $\text{Lift}(\{\text{牙刷}\}, \{\text{太阳镜}\}) = I(\{\text{牙刷}\} \rightarrow \{\text{太阳镜}\}) = \text{sup}(\{\text{牙刷}\} \rightarrow \{\text{太阳镜}\}) / (\text{sup}(\{\text{牙刷}\}) \times \text{sup}(\{\text{太阳镜}\})) = (N \times \sigma(\{\text{牙刷}\} \rightarrow \{\text{太阳镜}\})) / (\sigma(\{\text{牙刷}\}) \times \sigma(\{\text{太阳镜}\})) = (10000 \times 250) / (6000 \times 2000) = 0.208$

Answer5

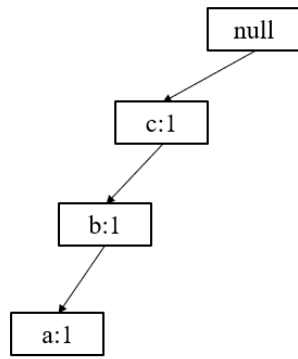
扫描数据库中对各项计数, 按支持度递减降序对频繁项排序

$L = \{\{c:5\}, \{b:4\}, \{d:3\}, \{e:3\}, \{a:2\}\}$

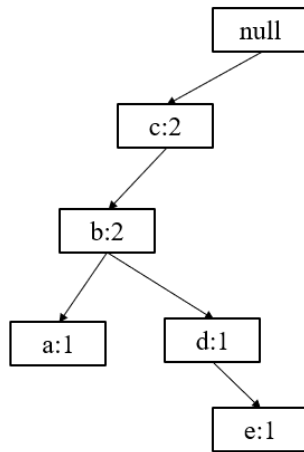
去除频繁项并重新排序

TID	Itemset
1	c, b, a
2	c, b, d, e
3	c, e, a
4	c, b, d
5	c, b, d, e

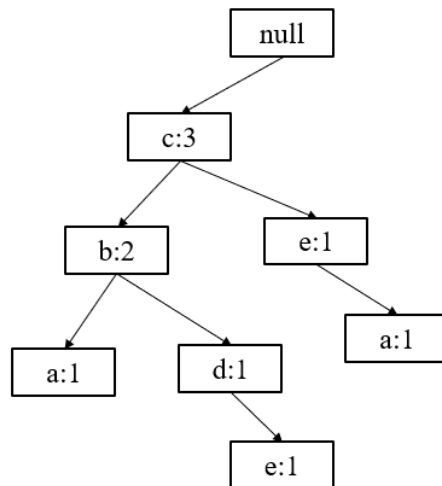
扫描第一个事务:



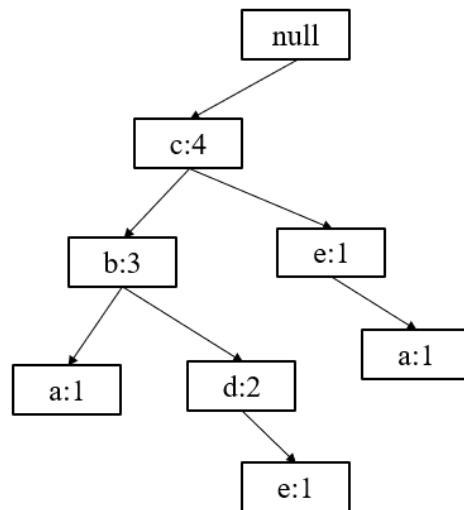
扫描第二个事务:



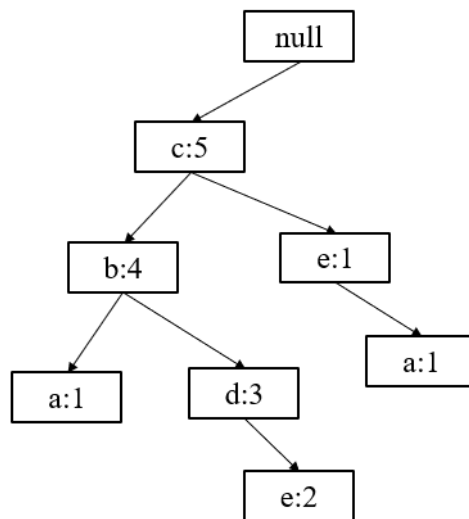
扫描第三个事务:



扫描第四个事务:



扫描第五个事务:



第五章 作业

1. 请简述 K 均值和层次聚类的异同。
2. 将如下的 8 个点聚类为 3 个簇：
 $x_1(2,10), x_2(2,5), x_3(8,4), x_4(5,8), x_5(7,5), x_6(6,4), x_7(1,2), x_8(4,9)$ 。
距离采用欧氏距离，假设初始质心分别是 x_1, x_2, x_7 ，用 K 均值算法给出：
(1) 第一次迭代后的 3 个簇的质心。
(2) 最终的 3 个簇的质心。
3. 指出在何种情况下，基于密度的聚类方法比 K 均值聚类 and 层次聚类方法更合适。通过实例说明。
4. 聚类样本足够多的情况下，K 均值算法会不会返回一个小于 K 个簇的结果？比如 $K=50$ 时，会不会产生 48 个簇的结果？分析原因。
5. 下面给出一个样本事务数据库（表 1），请对它分别实施 AGNES 算法和 DIANA 算法。

表 1 样本事务数据库

序号	属性 1	属性 2
1	1	1
2	1	2
3	2	1
4	2	2
5	3	4
6	3	5
7	4	4
8	4	5

第四次作业答案

1、相同：均需要提前设定聚类数和终止条件，并都以样本之间的距离作为分类的依据；

不同：K-means 需要提前指定初始的聚类中心，而层次聚类不用。

2、

(1) 各样本距离各初始中心的距离如下：

样本	Center1	Center2	Center3	归属簇
1	0	5	8.062258	1
2	5	0	3.162278	2
3	8.485281	6.082763	7.28011	2
4	3.605551	4.242641	7.211103	1
5	7.071068	5	6.708204	2
6	7.211103	4.123106	5.385165	2
7	8.062258	3.162278	0	3
8	2.236068	4.472136	7.615773	1

第一次聚类结果为：

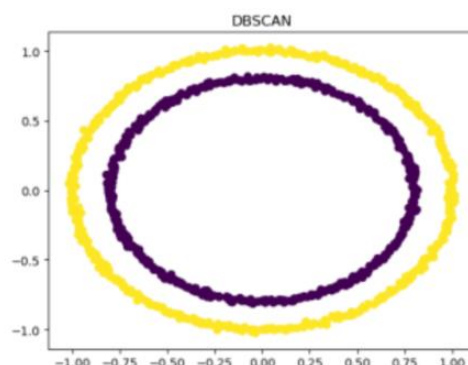
$$C_1 = \{x_1, x_4, x_8\}, C_2 = \{x_2, x_3, x_5, x_6\}, C_3 = \{x_7\},$$

第一次迭代后的 3 个簇的质心：(3.667,9),(5.75,4.5),(1,2)

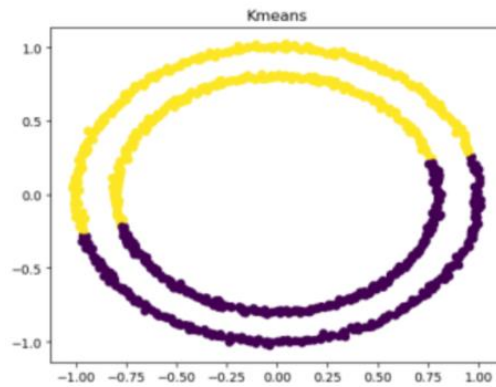
(2)最终的 3 个簇的质心为：(3.667,9),(7,4.333),(1.5,3.5)

3、相对于 K 均值与层次聚类，基于密度聚类方法可以处理不同大小和各种形状的簇，并且不太受噪声和离群点的影响。例如，当簇是圆环形状时如下所示。

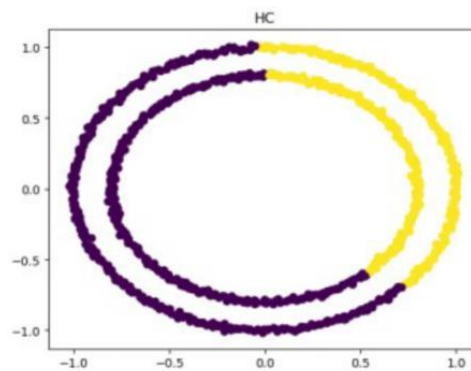
利用 DBSCAN 的聚类结果如下：



利用 K 均值的聚类结果如下：



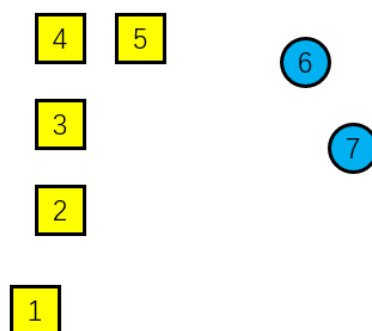
利用层次聚类结果如下：



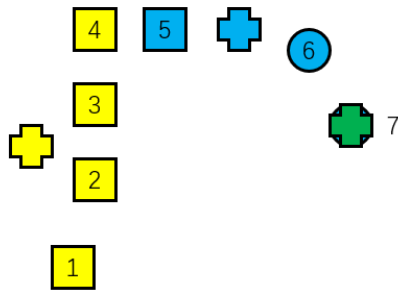
明显的，DBSCAN 更使用于非簇形状的聚类。

4、会出现缺失族的情况，这个问题同聚类过程中产生空簇是一个问题。原因在于初化中心选择不当。举例说明：

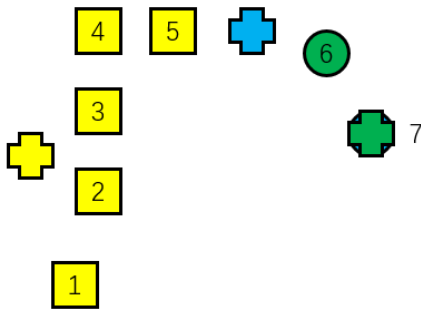
首先，假设有如下数据



聚类数设置为 3，初始中心选择 1，6，7，那么初始分组结果为 $\{1,2,3,4\}, \{5,6\}, \{7\}$



基于更新的类中心再聚类，可以发现，第二类中的样本为空，形成空簇。无法更新，此时只能返回两个簇。



5、

AGNES:

执行过程

在所给的数据集上运行 AGNES 算法，算法的执行过程如表 5.8 所示，设 $n = 8$ ，用户输入的终止条件为两个簇。初始簇为 $\{1\}$ ， $\{2\}$ ， $\{3\}$ ， $\{4\}$ ， $\{5\}$ ， $\{6\}$ ， $\{7\}$ ， $\{8\}$ 。（采用最小距离计算）

步骤	最近的簇距离	最近的两个簇	合并后的新簇
1	1	$\{1\}, \{2\}$	$\{1,2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{7\}, \{8\}$
2	1	$\{3\}, \{4\}$	$\{1,2\}, \{3,4\}, \{5\}, \{6\}, \{7\}, \{8\}$
3	1	$\{5\}, \{6\}$	$\{1,2\}, \{3,4\}, \{5,6\}, \{7\}, \{8\}$
4	1	$\{7\}, \{8\}$	$\{1,2\}, \{3,4\}, \{5,6\}, \{7,8\}$
5	1	$\{1,2\}, \{3,4\}$	$\{1,2,3,4\}, \{5,6\}, \{7,8\}$
6	1	$\{5,6\}, \{7,8\}$	$\{1,2,3,4\}, \{5,6,7,8\}$

(1) 先根据最小距离计算公式，将两两样本点的距离计算出来随机找出距离最小的两个簇，进行合并，最小距离为 1，合并 1、2 点为一个簇。

(2) 对上一次合并后的簇进行簇间计算，找出距离最近的两个簇进行合并，合并后 3、4 合并成为一簇。

(3) 重复第 (2) 步的工作，5、6 成为一簇。

- (4) 重复第 (2) 步的工作，7、8 成为一簇。
- (5) 合并{1,2}，{3,4}成为一簇。
- (6) 合并{5,6}，{7,8}成为一簇，合并后的簇的数目达到终止条件，计算完毕。

DIANA:

执行过程

步骤	具有最大直径的簇	Spliner group	Old party
1	{1,2,3,4,5,6,7,8}	{1}	{2,3,4,5,6,7,8}
2	{1,2,3,4,5,6,7,8}	{1,2}	{3,4,5,6,7,8}
3	{1,2,3,4,5,6,7,8}	{1,2,3}	{4,5,6,7,8}
4	{1,2,3,4,5,6,7,8}	{1,2,3,4}	{5,6,7,8}
5	{1,2,3,4,5,6,7,8}	{1,2,3,4}	{5,6,7,8} 终止

在第 1 步中，根据初始簇计算每个簇之间的距离，对簇中的每个点计算平均相异度（假定使用欧式距离）

平均距离如下：

样本	1	2	3	4	5	6	7	8
平均距离	2.96	2.526	2.68	2.18	2.18	2.68	2.526	2.96

挑出平均相异度最大的点 1 放到 Spliner group 中，剩余点放在 Old party 中。

第 2 步，在 Old party 里找出最近的 Spliner group 中的点的距离不大于到 Old party 中最近的点的距离的点，将该点放入 Spliner group 中，改点是 2。

第 3 步，重复第 2 步的工作，在 Spliner group 中放入点 3。

第 4 步,重复第 2 步的工作，在 Spliner group 中放入点 4。

第 5 步，没有新的 old party 中的点分配给 Spliner group，此时分裂的簇数为 2。达到终止条件。如果没有到终止条件，下一阶段还会从分裂好的簇中选一个直径最大的簇按刚才的分裂方法继续分裂。

1. 请简述聚类和分类的区别。

2. 考虑表1中二元分类问题的训练样本集，回答以下问题。

表1 训练样本集

实例	A	B	C	目标类
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

- (1) 整个训练样本集关于类属性的熵是多少？
- (2) 训练集中属性A、B的信息增益是多少？
- (3) 根据信息增益，A、B哪个是最佳划分？
- (4) 根据信息增益率，A、B哪个是最佳划分？
- (5) 根据gini指标，A、B哪个是最佳划分？

3.主观题 (20分)

考虑表2中的数据，回答以下问题。

表2 数据集

记录	A	B	C	类
1	0	0	1	F
2	1	0	1	T
3	0	1	0	F
4	1	0	0	F
5	1	0	1	T
6	0	0	1	T
7	1	1	0	F
8	0	0	0	F
9	0	1	0	T
10	1	1	1	T

- (1) 估计条件概率 $P(A=1|T)$ 、 $P(B=1|T)$ 、 $P(C=1|T)$ 、 $P(A=1|F)$ 、 $P(B=1|F)$ 、 $P(C=1|F)$ 。
 - (2) 根据 (1) 中的条件概率，使用朴素贝叶斯方法预测测试样本 ($A=1$ 、 $B=1$ 、 $C=1$) 的类标号。
 - (3) 比较 $P(A=1|T)$ 、 $P(B=1|T)$ 、 $P(A=1,B=1|T)$ 。给定类T，变量A和变量B条件独立吗？
4. 作为一种分类算法，支持向量机的基本原理是什么？常用的核函数有哪些？
5. 什么是神经网络？在神经网络中，sigmoid、tanh、ReLU 是十分常见的三种激活函数，他们的特点、优缺点有哪些？
6. 什么是模型的过拟合？如何解决过拟合问题，至少阐述两点解决方案？

第五次作业答案

1、区别：分类就是按照某种标准给对象贴标签，再根据标签来区分归类，是有监督的学习；聚类是指事先没有“标签”而通过某种成团分析找出事物之间存在聚集性原因的过程，属于无监督的学习。

2、

$$(1) \text{info}(D) = 0.991076$$

(2)

属性 A 的信息增益

$$\text{info}_A(D) = \frac{4}{9} * \text{info}_A(D_T) + \frac{5}{9} * \text{info}_A(D_F)$$

$$\text{info}_A(D_T) = 0.811278$$

$$\text{info}_A(D_F) = 0.721928$$

$$\text{info}_A(D) = 0.761639$$

$$\text{Gain}(A) = \text{info}(D) - \text{info}_A(D) = 0.229437$$

属性 B 的信息增益

$$\text{info}_B(D) = \frac{5}{9} * \text{info}_B(D_T) + \frac{4}{9} * \text{info}_B(D_F)$$

$$\text{info}_B(D_T) = 0.97095$$

$$\text{info}_B(D_F) = 1$$

$$\text{info}_B(D) = 0.983861$$

$$\text{Gain}(B) = \text{info}(D) - \text{info}_B(D) = 0.007215$$

(3)

$$\text{Gain}(A) = \text{info}(D) - \text{info}_A(D) = 0.229437$$

$$\text{Gain}(B) = \text{info}(D) - \text{info}_B(D) = 0.007215$$

$$\text{Gain}(C) = \text{info}(D) - \text{info}_C(D) = 0.143$$

根据信息增益，选择 A 属性作为最佳划分。

(4)

$$\text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}_A(D)} = 0.231$$

$$\text{GainRatio}(B) = \frac{\text{Gain}(B)}{\text{SplitInfo}_B(D)} = 0.007$$

根据信息增益率，选择 A 属性

(5)

$$Gini(D) = 1 - \sum_{i=1}^m p_i^2 = 0.494$$

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) = 0.344$$

$$Gini_B(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) = 0.488$$

根据 gini 系数最小的属性，因此选择 A 属性。

3、

(1)

$$P(A = 1|T) = \frac{3}{5}$$

$$P(B = 1|T) = \frac{2}{5}$$

$$P(C = 1|T) = \frac{4}{5}$$

$$P(A = 1|F) = \frac{2}{5}$$

$$P(B = 1|F) = \frac{2}{5}$$

$$P(C = 1|F) = \frac{1}{5}$$

(2)

$$P(A = 1, B = 1, C = 1|T) = P(A = 1|T) * P(B = 1|T) * P(C = 1|T) = 24/125$$

$$P(A = 1, B = 1, C = 1|F) = P(A = 1|F) * P(B = 1|F) * P(C = 1|F) = 4/125$$

$$P(T) * P(A = 1, B = 1, C = 1|T) = 12/125 > P(F) * P(A = 1, B = 1, C = 1|F) = 2/125$$

因此预测样本的类标号为 T。

(3)

$$P(A = 1|T) = \frac{3}{5}$$

$$P(B = 1|T) = \frac{2}{5}$$

$$P(A = 1|T) * P(B = 1|T) > P(A = 1, B = 1|T)$$

因此说明给定类 T ，变量 A 和 B 条件不独立。

4、

基本原理

支持向量机的基本模型是定义在特征空间上的间隔最大的线性分类器，支持向量机还包括核技巧，这使它成为实质上的非线性分类器。支持向量机的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题。支持向量机的学习算法是求解凸二次规划的最优化算法。

常用的核函数

线性核函数:主要用于线性可分的情况。

多项式核函数:一种非稳态核函数，适合于正交归一化后的数据。

径向基核函数:具有很强的灵活性，应用广泛。大多数情况下有较好的性能。

Sigmoid 核:来源于 MLP 中的激活函数，SVM 使用 Sigmoid 相当于一个两层的感知机网络。

5、

神经网络是由具有适应性的简单单元组成的广泛并行互联的网络，它的组织能够模拟生物神经系统对真实世界物体做出的交互反应。神经网络的训练采用 BP 算法，给定训练集和学习率，随机初始化网络中的连接权重和阈值。不断地根据当前的参数计算当前样本的输出，并和理想的输出计算均方误差，并且根据链式法则计算输出层和隐层神经元的梯度项，来更新网络中的连接权重和阈值，直到达到停止条件。

sigmoid 函数将任意大小的输入都压缩到 $[0,1]$ 之间，输入值越小，压缩后越趋近于 0。它在神经网络中常常用作二分类器最后一层的激活函数，可以将任意实数值转换为概率。sigmoid 函数的导数是其本身的函数，计算方便。其最明显的缺点是容易饱和，出现梯度消失等问题，导致层数较多的深度神经网络难以有效训练;另外，它的输出均大于 0，使得输出不是零均值，所以 sigmoid 函数现在很少在深度神经网络的中间层作为激活函数使用。

tanh 函数将任意大小的输入都压缩到 $[-1,1]$ 之间，其输出均值是 0，使得收敛速度比 sigmoid 要快，但是其仍然具有饱和性，会造成梯度消失，同时还有更复

杂的幂运算。

ReLU将负数部分置零，保留正数部分不变，在计算上非常高效，且避免了梯度消失的问题。ReLU函数的优点包括：计算简单，不会导致梯度消失问题，收敛速度较快。然而，ReLU函数在负数部分输出为0，可能导致神经元死亡；不是零均值激活函数，可能导致训练时的震荡问题。

6、

过拟合是指模型对训练集的学习程度过高，学习到了训练集的数据特性，但没有理解数据背后的规律，泛化能力差，导致模型在训练集上表现很好，但在测试集上却表现很差。发生过拟合一般是由于模型复杂度过高，或者数据集样本不足。

解决过拟合问题，常用的有以下几个方案：

- ①提前停止训练;
- ②设置 Dropout;
- ③获取和使用更多的数据;
- ④控制模型的复杂度;
- ⑤删除冗余特征。