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:负责搜小数据的取值范围，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：$\mu_A = 42.0214, \sigma_A = 3.95617$

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

（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

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

对协方差矩阵进行特征根分解，可以得到

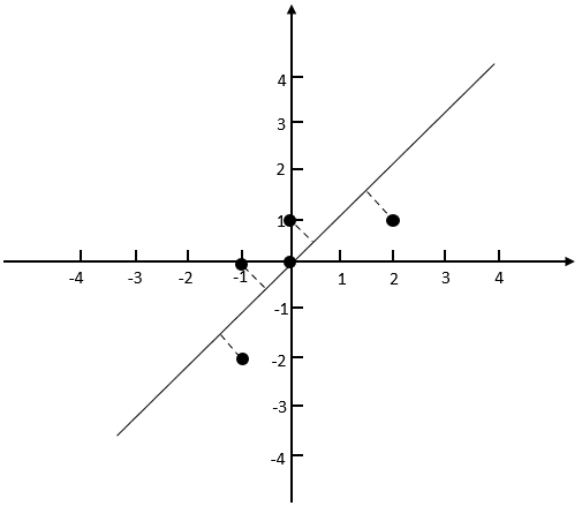
$$\begin{pmatrix} 6 & 4 \\ 4 & 6 \end{pmatrix} \begin{pmatrix} -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix} = \begin{pmatrix} 10 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2}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

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$c_1 = (-\frac{\sqrt{2}}{2} \quad \frac{\sqrt{2}}{2})^T$。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。

$$t_1 = Xp_1 = \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}^T (-\frac{\sqrt{2}}{2} \quad -\frac{\sqrt{2}}{2})^T = (\frac{3\sqrt{2}}{2} \quad \frac{\sqrt{2}}{2} \quad 0 \quad -\frac{3\sqrt{2}}{2} \quad -\frac{\sqrt{2}}{2})^T$$

$$= (\frac{3\sqrt{2}}{2} \quad \frac{\sqrt{2}}{2} \quad 0 \quad -\frac{3\sqrt{2}}{2} \quad -\frac{\sqrt{2}}{2})$$

可视化后的结果图如图所示


4
3
2
1
-4
-3
-2
-1
1
2
3
4
-1
-2
-3
-4