数据挖掘第一次作业

BraveY

2020 年 1 月 17 日

1 第一次作业written part

1.1 题目1

1.Suppose that a data warehouse consists of four dimensions, date, spectator, location, and game, and two measures, count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.

(a) Draw a star schema diagram for the data warehouse. (b) Starting with the base cuboid [date, spectator, location, game]，what specific OLAP operations should one perform in order to list the total charge paid by spectators in Chicago in 1999? (c) Bitmap indexing is a very useful optimization technique. Please present the pros and cons of using bitmap indexing in this given data warehouse.

1.1.1 解

(a):数据仓库的星座模式如图 1所示。

(b):首先在date维度上面进行上卷roll-up操作，上卷到year这个层次，game维度也上卷到所有比赛项目total events这个层次，然后进行切块操作其中location选择芝加哥，date选择1999年。

(c):使用位图的**优点**是：对于基数较小的值域，位图可以将其比较、连接和聚集操作都简化成位算术运算，从而大大减少了处理时间。同时使用bit来表示一个字符串，位图索引可以显著降低内存空间以及I/O开销。

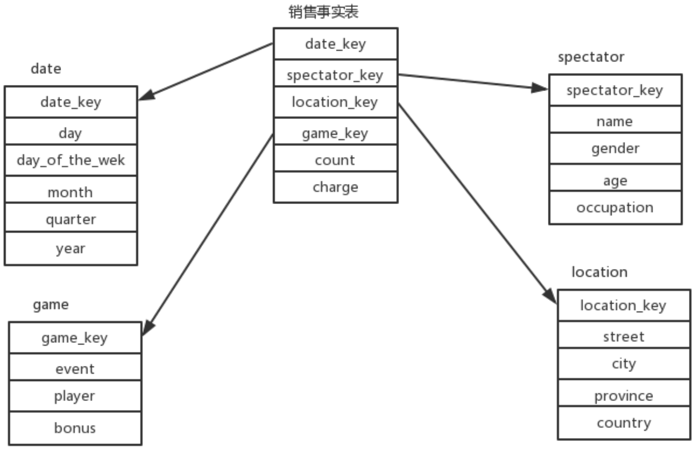
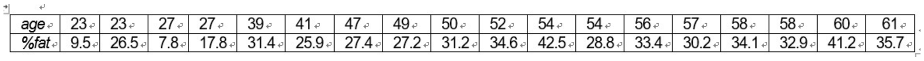


图 1: star schema

缺点：因为值域中的每一个可能的值都需要一个位向量来记录，所以当基数较大时，会需要开辟很大的存储空间，因为每一个记录其实只使用到了一位，因此造成了很大的浪费。同时当可取的值连续时无法使用位图来记录。

1.2 题目2

2. Suppose a hospital tested the age and body fat data for 18 random selected adults with the following result:



age	23	23	27	27	39	41	47	49	50	52	54	54	56	57	58	58	60	61
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

图 2: data table

(a) Calculate the mean, median, and standard deviation of age and %fat.

(b) Draw the boxplots for age and %fat.

(c) Draw a scatter plot based on these two variables.

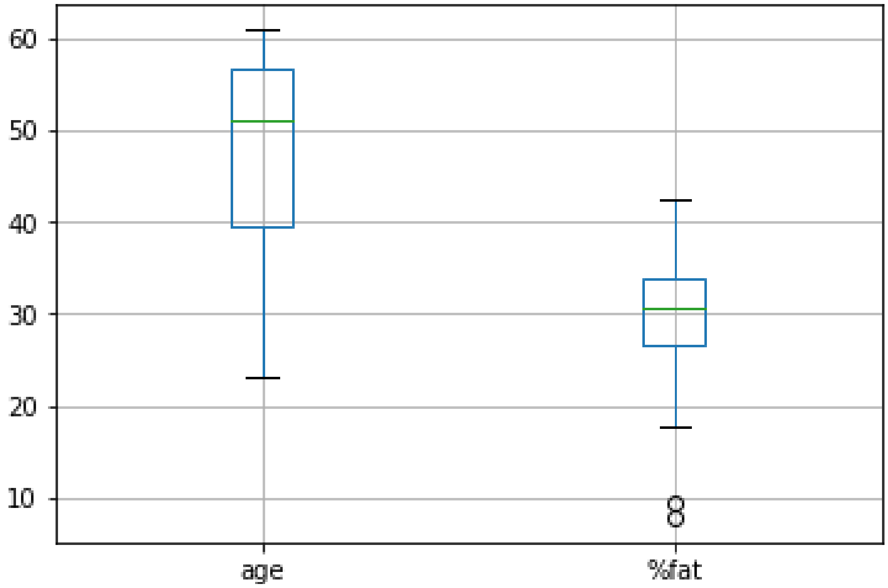


图 3: boxplot

(d) Normalize age based on min-max normalization.

(e) Calculate the correlation coefficient (Pearson’s product moment coefficient). Are these two variables positively or negatively correlated?

(f) Smooth the fat data by bin means, using a bin depth of 6.

(g) Smooth the fat data by bin boundaries, using a bin depth of 6.

1.2.1 解

(a)

- 年龄age的均值mean为：46.44 中位数median为：51 标准差standard deviation为12.85
- 体脂率%fat的均值mean为：28.78 中位数median为：30.7 标准差standard deviation为8.99

(b)盒图如图 3所示。

(c)散点图如图 4所示。

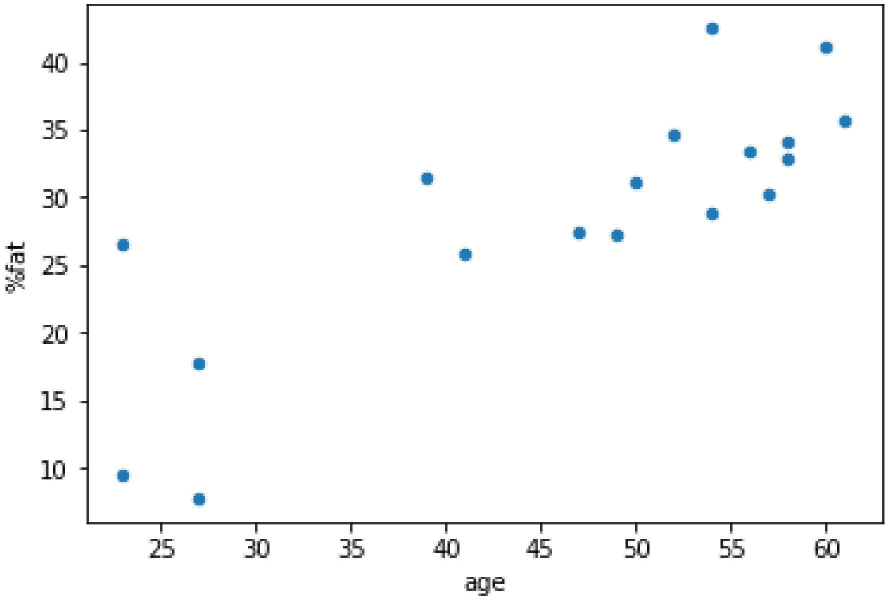


图 4: scatter

(d)年龄数据使用最大最小归一化后为：[0.37704918, 0.37704918, 0.44262295, 0.44262295, 0.63934426, 0.67213115, 0.7704918 , 0.80327869, 0.81967213, 0.85245902, 0.8852459 , 0.8852459 , 0.91803279, 0.93442623, 0.95081967, 0.95081967, 0.98360656, 1.]

(e)根据相关系数的计算公式,计算出来的age和%fat的相关系数为：0.82,是正相关即体脂率有随着年龄的增长而增长。

(f)体脂率使用箱均值光滑后的数据为： [17.5, 17.5, 17.5, 17.5, 17.5, 17.5, 29.8, 29.8, 29.8, 29.8, 29.8, 29.8, 35.8, 35.8, 35.8, 35.8, 35.8, 35.8]

(g)体脂率使用箱边界值光滑后的数据为： [7.8, 7.8, 26.5, 26.5, 26.5, 27.4, 27.4, 31.4, 31.4, 31.4,33.4, 33.4, 33.4, 33.4, 41.2]

1.3 题目3

3. Given a data set below for attributes Height, Hair, Eye and two classes C1, C2. Construct a decision tree with Information Gain.

	Height	Hair	Eye	Class
1	Tall	Blond	Brown	C1
2	Tall	Dark	Blue	C1
3	Tall	Dark	Brown	C1
4	Short	Dark	Blue	C1
5	Short	Blond	Brown	C1
6	Tall	Red	Blue	C2
7	Tall	Blond	Blue	C2
8	Short	Blond	Blue	C2
9	Medium	Dark	Blue	C2

表 1: 数据

1.3.1 解

首先计算完整的数据元组的期望信息也就是熵，因为完整数据有5个C_1类4个C_2类所以有熵为

$$Info(D)=-\frac{5}{9}\log_2\frac{5}{9}-\frac{4}{9}\log_2\frac{4}{9}=0.991bit$$

Height属性有5个tall其中3个C_1类2个C_2类；3个short其中2个C_1类1个C_2类；1个medium且为C_2类,所以Height属性的熵为：

$$Info_{Height}(D)=\frac{5}{9}\times(-\frac{3}{5}\log_2\frac{3}{5}-\frac{2}{5}\log_2\frac{2}{5})+\frac{3}{9}\times(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3})+\frac{1}{9}\times(-\frac{1}{1}\log_2\frac{1}{1})=0.846bit$$

Hair属性有4个Blond属性，其中2个C_1类2个C_2类；4个Dark其中3个C_1类1个C_2类，一个Red而且属于C_2类。所以Hair属性的熵为：

$$Info_{Hair}(D)=\frac{4}{9}\times(-\frac{2}{4}\log_2\frac{2}{4}-\frac{2}{4}\log_2\frac{2}{4})+\frac{4}{9}\times(-\frac{3}{4}\log_2\frac{3}{4}-\frac{1}{4}\log_2\frac{1}{4})+\frac{1}{9}\times(-\frac{1}{1}\log_2\frac{1}{1})=0.805bit$$

Eye属性有3个Brown，都属于C_1类；6个blue其中2个C_1类4个C_2类，所以Eye属性的熵为：

$$Info_{Eye}(D)=\frac{3}{9}\times(-\frac{3}{3}\log_2\frac{3}{3})+\frac{6}{9}\times(-\frac{2}{6}\log_2\frac{2}{6}-\frac{4}{6}\log_2\frac{4}{6})=0.612bit$$

接下来就可以根据先前计算的熵算出每个属性的信息增益为：

$$\begin{aligned}
Gain(Height)=Info(D)-Info_{Height}(D)=0.991-0.846=0.145bit\\
Gain(Hair)=Info(D)-Info_{Hair}(D)=0.991-0.805=0.186bit\\
Gain(Eye)=Info(D)-Info_{Eye}(D)=0.991-0.612=0.379bit
\end{aligned}\tag{1}$$

因为Eye属性的信息增益最大，所以它被选为最好的分裂属性。当使用了Eye属性作为第一个次的决策属性后，剩下的Blue数据集D_1为：

	Height	Hair	Class
1	Tall	Dark	C1
2	Short	Dark	C1
3	Tall	Red	C2
4	Tall	Blond	C2
5	Short	Blond	C2
6	Medium	Dark	C2

此时再次进行信息增益的计算：D_1有2个C_1类4个C_2类。所以有熵为

$$Info(D_1) = -\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6} = 0.918bit$$

Height属性有3个tall其中1个是C_1类2个是C_2类，2个short其中1个C_1类1个C_2类，一个Medium且为C_2类。所以有熵为：

$$Info_{Height}(D_1) = \frac{3}{6}\times(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3})+\frac{2}{6}\times(-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2})+\frac{1}{6}\times(-\frac{1}{1}\log_2\frac{1}{1}) = 0.792bit$$

Hair属性有3个dark其中2个为$C_1$1个为C_2,1red且为C_2,2个Blond均为C_2类。所以有熵为：

$$Info_{Hair}(D_1) = \frac{3}{6}\times(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3})+\frac{2}{6}\times(-\frac{2}{2}\log_2\frac{2}{2})+\frac{1}{6}\times(-\frac{1}{1}\log_2\frac{1}{1}) = 0.459bit$$

$$\begin{aligned} Gain(Height) &= Info(D_1) - Info_{Height}(D) = 0.918 - 0.792 = 0.126bit \\ Gain(Hair) &= Info(D_1) - Info_{Hair}(D) = 0.918 - 0.459 = 0.459bit \end{aligned} \tag{2}$$

因为Hair属性的信息增益最大所以在D_1子数据集中选择Hair属性作为分裂属性。剩下的Dark属性数据集D_2为: 选择Height属性来进行决策。最终的

	Height	Class
1	Tall	C1
2	Short	C1
3	Medium	C2

决策树如图 5所示。

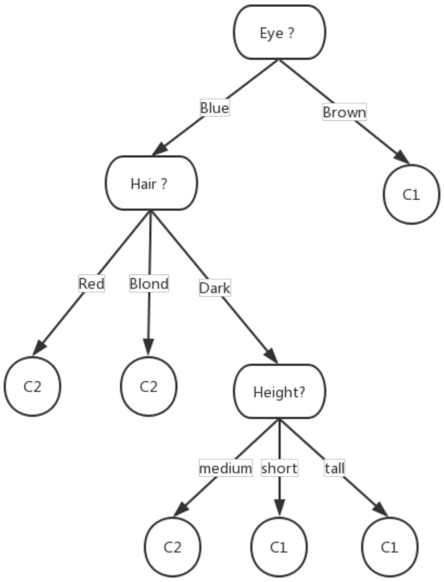


图 5: decision tree

1.4 题目4

4. Design a multilayer feed-forward neural network (one hidden layer) for the data set in Q3. Label the nodes in the input and output layers. Using the neural network obtained above, show the weight values after one iteration of the back propagation algorithm, given the training instance “(Medium, Dark, Brown, C2)”. Indicate your initial weight values and biases and the learning rate used.

1.4.1 解

因为Height，Hair，Eye三个属性总共有8种输入可能，所以可以分配8个输入单元，其中Height属性的Tall、Short、Medium对应三个输入单元I_0, I_1, I_2，Hair属性的Blond、Dark、Red对应三个输入单元I_3, I_4, I_5，Eye属性的Brown、BlueI_6, I_7分配两个输入单元。输出的只有C_1和C_2两类，所以可以只有一个输出单元O_0输出为0表示为C_1，输出为1表示C_2。隐藏层单元数目设置为2个，学习率l设置为0.9。神经网络如图 6所示。

输入和偏倚初始化如表2所示。

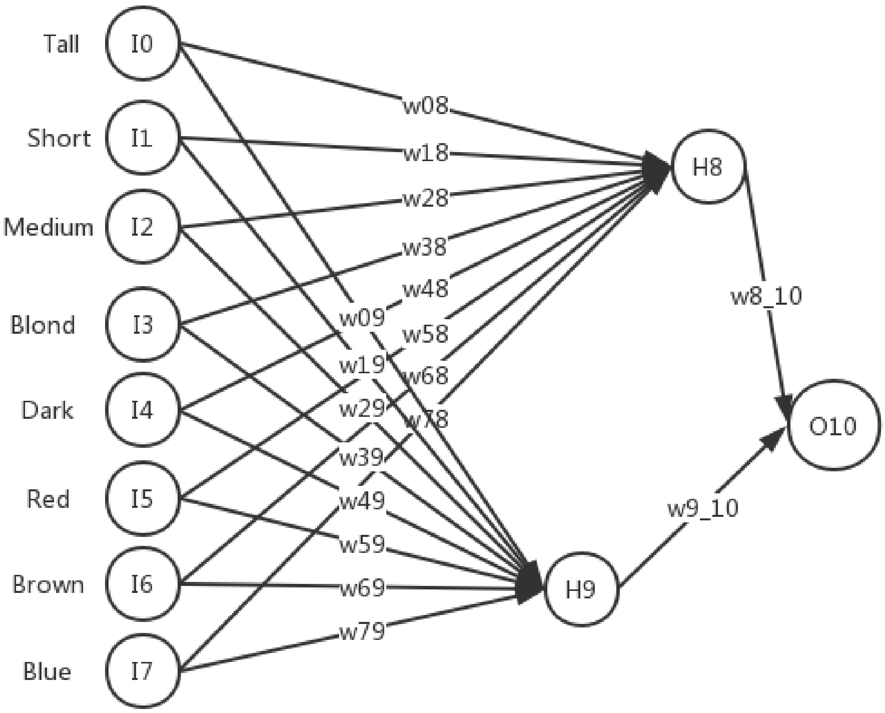


图 6: neural network

I_0	I_1	I_2	I_3	I_4	I_5	I_6	I_7	θ_8	θ_9	θ_{10}
0	0	1	0	1	0	1	0	-0.4	0.2	0.1

表 2: 输入和偏倚初始化

权重初始化如表3所示。

净输入和输出的计算如表4所示。

每个输出单元的误差计算如表5所示。

权重和偏倚的更新计算如表6所示，根据公式只有输出不为0的单元其权重才会进行更新。

1.5 题目5

5. Classify the unknown sample Z based on the training data set in Q3:Z = (Height = Short, Hair = blond, Eye = brown). What would a naïve

w_{08}	w_{18}	w_{28}	w_{38}	w_{48}	w_{58}	w_{68}	w_{78}	w_{09}	w_{19}	w_{29}	w_{39}	w_{49}	w_{59}	w_{69}	w_{79}	w_{810}	w_{910}
0.2	-0.3	0.4	0.1	-0.5	0.2	-0.2	-0.2	0.5	-0.4	-0.1	0.3	-0.1	0.5	0.3	-0.5	0.2	-0.5

表 3: 权重初始化

单元j	净输入I_j	输出O_j
8	$0.4-0.5-0.2-0.4=-0.7$	$\frac{1}{1+e^{0.7}}=0.332$
9	$-0.1-0.1+0.3+0.2=0.3$	$\frac{1}{1+e^{-0.3}}=0.574$
10	$0.332\times0.2-0.5\times0.574+0.1=-0.121$	$\frac{1}{1+e^{-0.121}}=0.530$

表 4: 计算净输入和输出

Bayesian classifier classify Z?

1.5.1 解

根据表1所示，有

$$\begin{aligned}
P(C_1) &= \frac{5}{9}\\
P(C_2) &= \frac{4}{9}\\
P(Height=short|C_1) &= \frac{3}{5}\\
P(Hair=blond|C_1) &= \frac{2}{5}\\
P(Eye=brown|C_1) &= \frac{3}{5}\\
P(Height=short|C_2) &= \frac{2}{4}\\
P(Hair=blond|C_2) &= \frac{3}{4}\\
P(Eye=brown|C_2) &= \frac{0}{4}=0
\end{aligned} \tag{3}$$

因为$P(Eye=brown|C_2)=\frac{0}{4}=0$所以需要使用拉普拉斯校准，校准后有：

$$P(Eye=brown|C_2)=\frac{1}{5} \tag{4}$$

单元j	误差Err_j
10	$0.530 \times (1-0.530) \times (1-0.530) = 0.117$
9	$0.574 \times (1-0.574) \times 0.117 \times (-0.5) = -0.014$
8	$0.332 \times (1-0.332) \times 0.117 \times 0.2 = 0.005$

表 5: 误差计算

所以有：

$$\begin{aligned} P(Z|C_1) &= P(Height=short|C_1) \times P(Hair=blond|C_1) \times P(Eye=brown|C_1) = \frac{3}{5} \times \frac{2}{5} \times \frac{3}{5} = 0.144 \\ P(Z|C_2) &= P(Height=short|C_2) \times P(Hair=blond|C_2) \times P(Eye=brown|C_2) = \frac{2}{4} \times \frac{3}{4} \times \frac{1}{5} = 0.075 \end{aligned} \tag{5}$$

所以可以计算出：

$$\begin{aligned} P(Z|C_1)P(C_1) &= 0.144 \times \frac{5}{9} = 0.08 \\ P(Z|C_2)P(C_2) &= 0.075 \times \frac{4}{9} = 0.03 \end{aligned} \tag{6}$$

所以对与数据Z而言，朴素贝叶斯预测的其类为：C_1类。

2 第一次作业lab part

2.1 题目1

Question 1. Assume a supermarket would like to promote pasta. Use the data in “transactions” as training data to build a decision tree (C5.0 algorithm) model to predict whether the customer would buy pasta or not. 1. Build a decision tree using data set “transactions” that predicts pasta as a function of the other fields. Set the “type” of each field to “Flag” , set the “direction” of “pasta” as “out” , set the “type” of COD as “Typeless” , select “Expert” and set the “pruning severity” to 65, and set the “minimum records per child branch” to be 95. **Hand-in:** A figure showing your tree. 2. Use the model (the full tree generated by Clementine in step 1 above) to make a prediction for each of the 20 customers in the “rollout” data to determine whether the customer would buy pasta. **Hand-in:** your prediction for each of the 20 customers. **Hand-in:** rules for positive (yes) prediction of pasta purchase identified from the decision tree (up to the fifth level. The root is considered as level 1).

2.1.1 解

在Clementine中的Stream流程图如图7所示,导入数据后求得决策树模型的部分如图8所示,完整的决策数如图9所示。

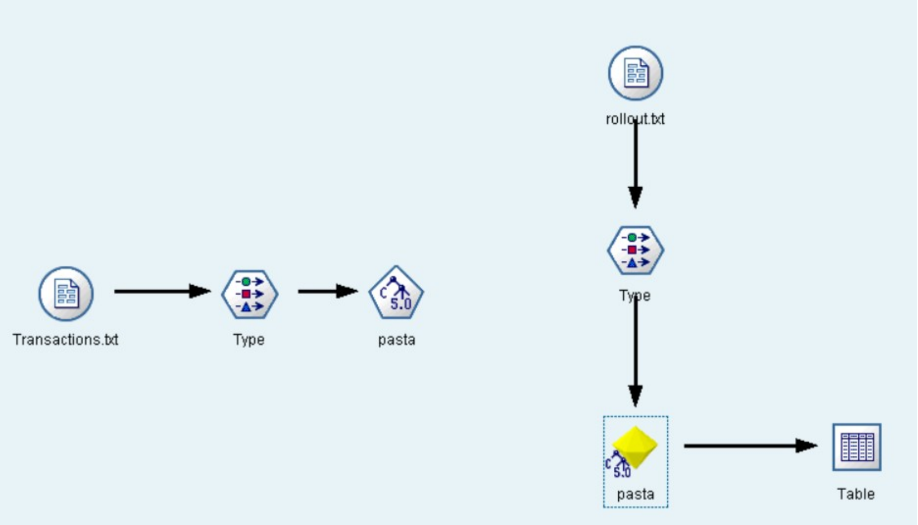


图 7: stream

2.1.2 解

将预测的数据导入到模型后有预测结果如图10 决策树规则如图11所示。

2.2 题目2

Question 2: 考试成绩预测通过对某在线培训系统的标注数据集进行建模，预测其它会员期末考试的结果。数据集来自在线培训系统的日志，数据包括每个会员的在线学习行为。请尝试多种不同的模型、不同的参数，建立高质量的预测模型。训练集有873条记录，测试集有461条记录。训练集和测试集包含如下变量：

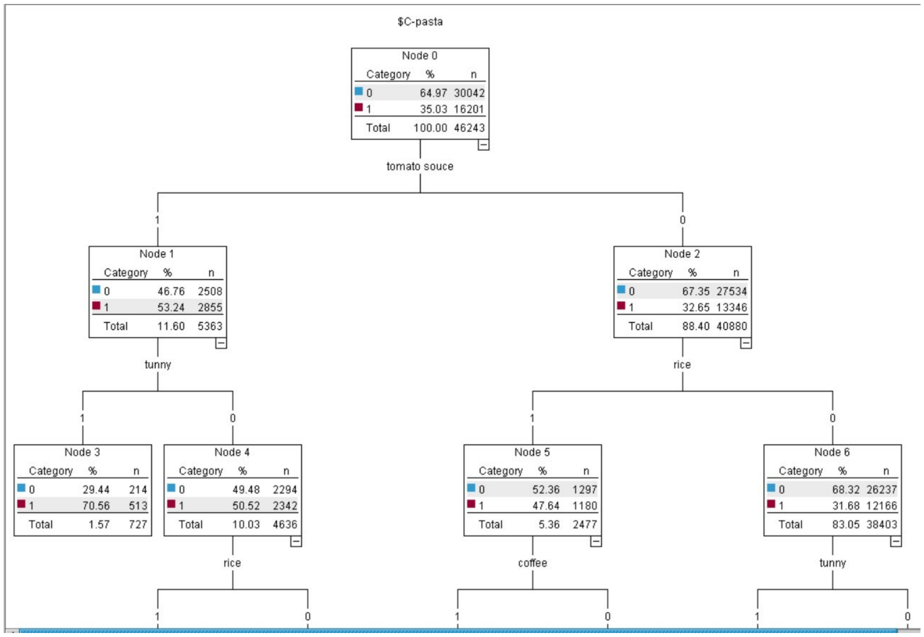


图 8: decision tree part

人员ID 在线总时长（分钟）在线阅读时长（分钟）在线测试时长（分钟）全文阅读次数智能阅读次数知识点阅读次数试题阅读次数回溯原文次数题库测试次数仿真考试次数仿真考试优秀次数仿真考试良好次数仿真考试合格次数仿真考试不合格次数 Class: 期末考试结果及格=1, 期末考试结果不及格=0

1. Perform decision tree classification on training data set. Select all the input variables except ID. Set the “Direction” of class as “out”, “type” as “Flag”. Then, you can specify the “minimum records per child branch” and “pruning severity”, and then click “use global pruning”. Hand-in the confusion matrices for test data. (Provide the best parameters for “minimum records per child branch” and “pruning severity”.)

2. Perform neural network on training data set using default settings. Again, select all the input variables except ID. Hand-in the confusion matrix for test data.

3. Perform logistic regression on training data set using default settings. Again, select all the input variables except ID. Hand-in the confusion matrix

for test data. 4. Hand-in your comments on the model quality for decision tree, neural network and logistic regression using the confusion matrices.

2.2.1 解

经过实验调参，最好的参数为minimum records per child branch=10，pruning severity=50，此时有矩阵如图12所示。

2.2.2 解

使用默认的神经网络设置得到的结果如图13所示。

2.2.3 解

使用默认的Logistic回归设置得到的结果如图14所示。

2.3 解

根据三种模型的输出可以看到，性能最好的是决策树准确率最高，只有10个样例判别错误；其次是logsitic 回归15个样例判别错误；最差的是神经网络共有44个样例判别错误，因为默认的条件只迭代了250次，所以准确率还较低。

权重或偏倚	新值
w_{810}	$0.2 + 0.9 \times 0.117 \times 0.332 = 0.235$
w_{910}	$-0.5 + 0.9 \times 0.117 \times 0.574 = -0.440$
w_{08}	0.2
w_{09}	0.5
w_{18}	-0.3
w_{19}	-0.4
w_{28}	$0.4 + 0.9 \times 0.005 \times 1 = 0.406$
w_{29}	$-0.1 + 0.9 \times -0.014 \times 1 = -0.113$
w_{38}	0.1
w_{39}	0.3
w_{48}	$-0.5 + 0.9 \times 0.005 \times 1 = -0.496$
w_{49}	$-0.1 + 0.9 \times -0.014 \times 1 = -0.113$
w_{58}	0.2
w_{59}	0.5
w_{68}	$-0.2 + 0.9 \times 0.005 \times 1 = -0.196$
w_{69}	$0.3 + 0.9 \times -0.014 \times 1 = 0.287$
w_{78}	-0.2
w_{79}	-0.5
θ_{10}	$0.1 + 0.9 \times 0.117 = 0.205$
θ_9	$0.2 + 0.9 \times -0.014 = 0.187$
θ_8	$-0.4 + 0.9 \times 0.005 = -0.396$

表 6: 权重和偏倚更新

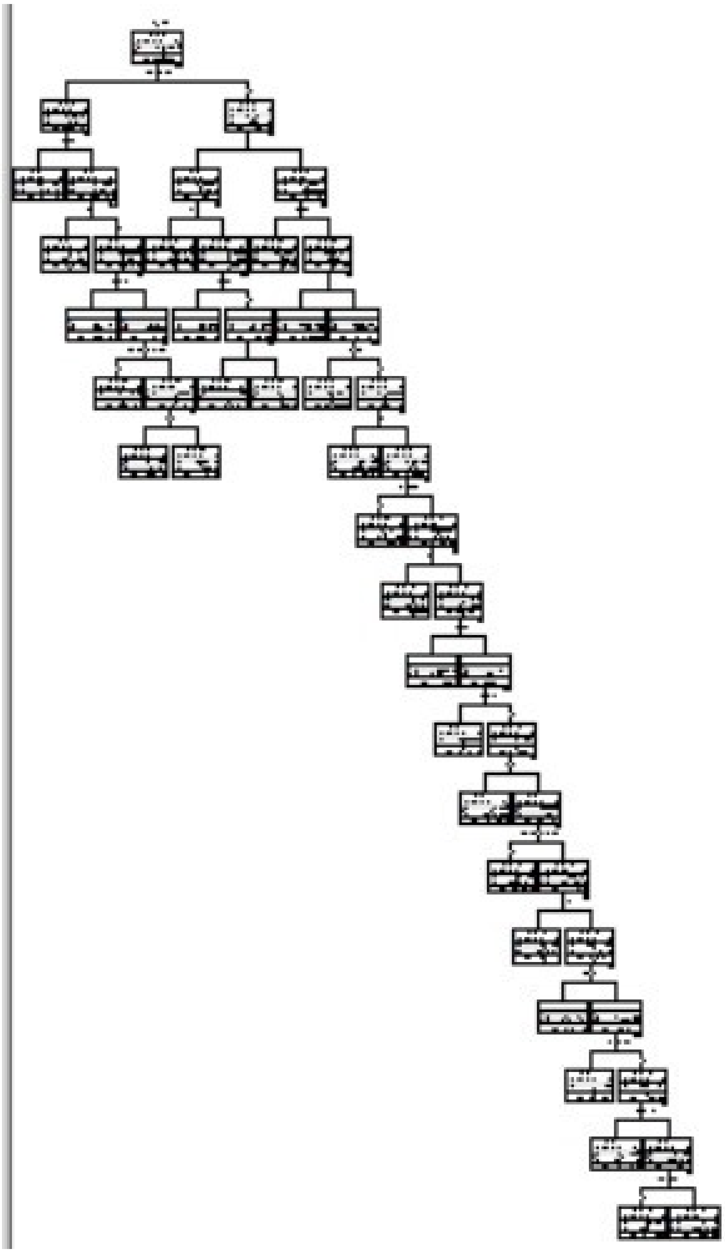
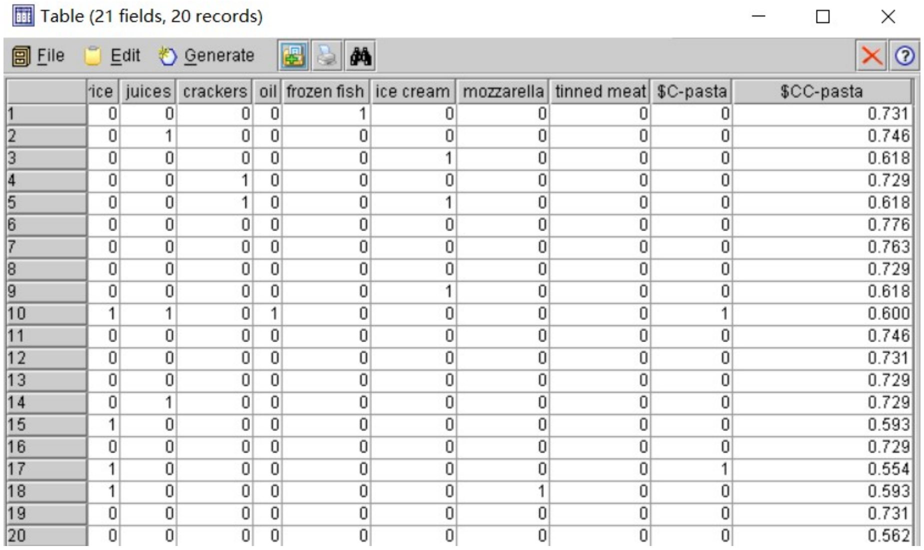
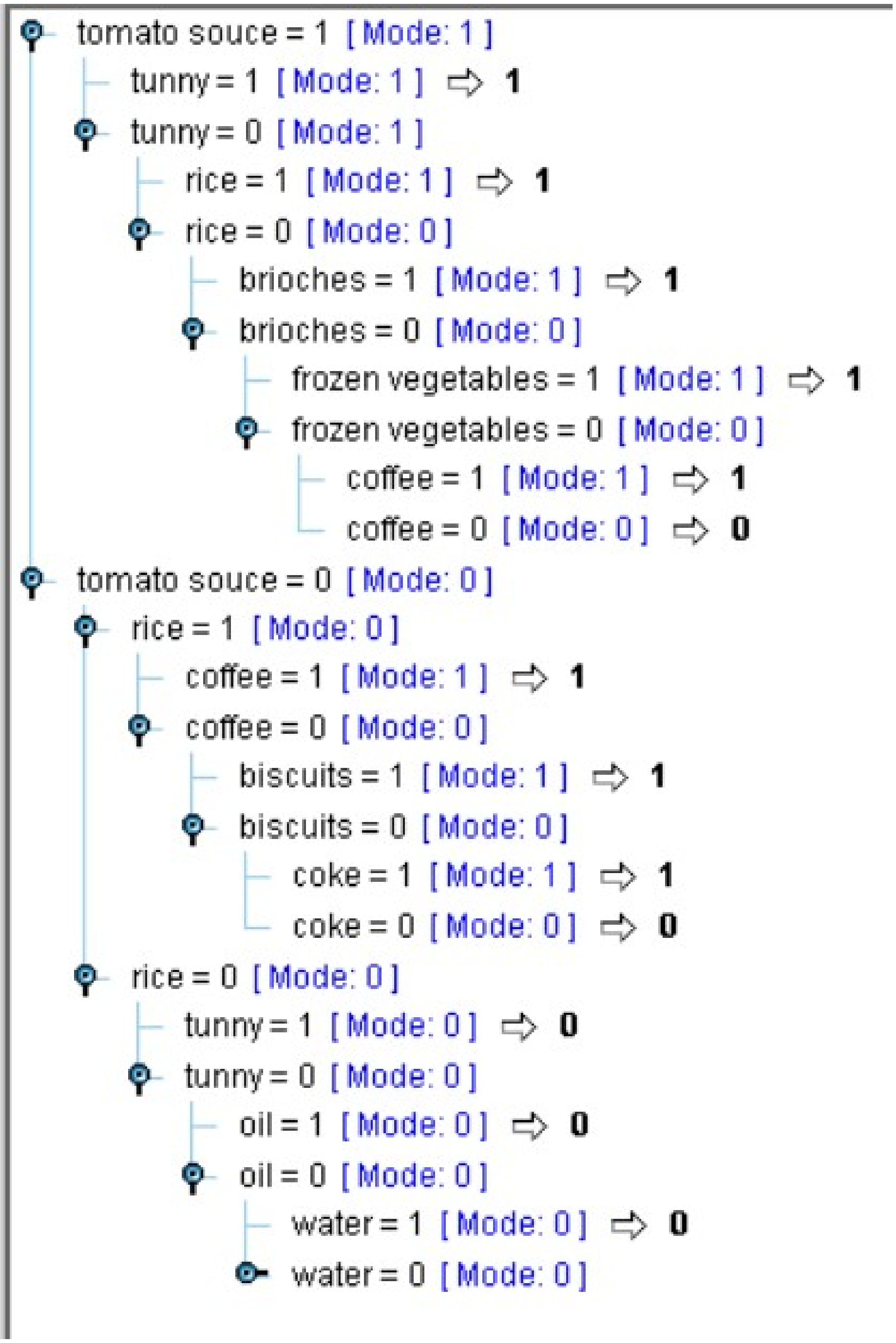


图 9: decision tree total



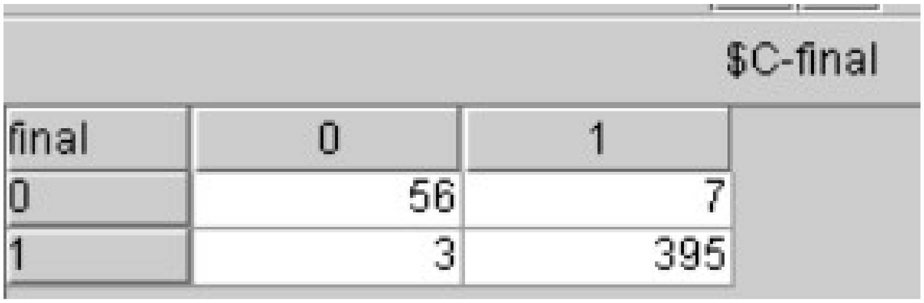
	ice	juices	crackers	oil	frozen fish	ice cream	mozzarella	tinned meat	$C-pasta	$CC-pasta
1	0	0	0	0	1	0	0	0	0	0.731
2	0	1	0	0	0	0	0	0	0	0.746
3	0	0	0	0	0	1	0	0	0	0.618
4	0	0	1	0	0	0	0	0	0	0.729
5	0	0	1	0	0	1	0	0	0	0.618
6	0	0	0	0	0	0	0	0	0	0.776
7	0	0	0	0	0	0	0	0	0	0.763
8	0	0	0	0	0	0	0	0	0	0.729
9	0	0	0	0	0	1	0	0	0	0.618
10	1	1	0	1	0	0	0	0	1	0.600
11	0	0	0	0	0	0	0	0	0	0.746
12	0	0	0	0	0	0	0	0	0	0.731
13	0	0	0	0	0	0	0	0	0	0.729
14	0	1	0	0	0	0	0	0	0	0.729
15	1	0	0	0	0	0	0	0	0	0.593
16	0	0	0	0	0	0	0	0	0	0.729
17	1	0	0	0	0	0	0	0	1	0.554
18	1	0	0	0	0	0	1	0	0	0.593
19	0	0	0	0	0	0	0	0	0	0.731
20	0	0	0	0	0	0	0	0	0	0.562

图 10: prediction



```
tomato souce = 1 [Mode: 1 ]
    tunny = 1 [Mode: 1 ] ⇨ 1
    tunny = 0 [Mode: 1 ]
        rice = 1 [Mode: 1 ] ⇨ 1
        rice = 0 [Mode: 0 ]
            brioches = 1 [Mode: 1 ] ⇨ 1
            brioches = 0 [Mode: 0 ]
                frozen vegetables = 1 [Mode: 1 ] ⇨ 1
                frozen vegetables = 0 [Mode: 0 ]
                    coffee = 1 [Mode: 1 ] ⇨ 1
                    coffee = 0 [Mode: 0 ] ⇨ 0
tomato souce = 0 [Mode: 0 ]
    rice = 1 [Mode: 0 ]
        coffee = 1 [Mode: 1 ] ⇨ 1
        coffee = 0 [Mode: 0 ]
            biscuits = 1 [Mode: 1 ] ⇨ 1
            biscuits = 0 [Mode: 0 ]
                coke = 1 [Mode: 1 ] ⇨ 1
                coke = 0 [Mode: 0 ] ⇨ 0
    rice = 0 [Mode: 0 ]
        tunny = 1 [Mode: 0 ] ⇨ 0
        tunny = 0 [Mode: 0 ]
            oil = 1 [Mode: 0 ] ⇨ 0
            oil = 0 [Mode: 0 ]
                water = 1 [Mode: 0 ] ⇨ 0
                water = 0 [Mode: 0 ]
```

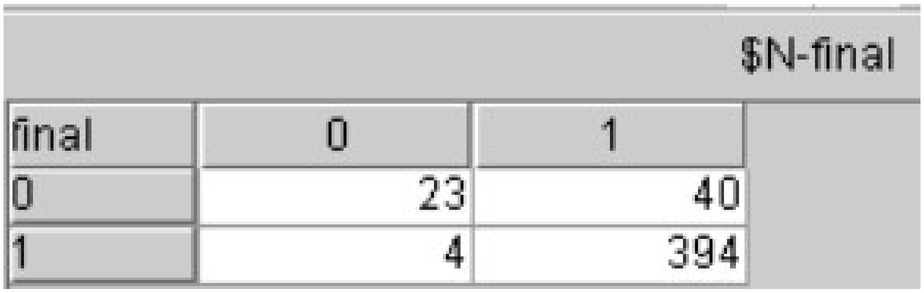
图 11: rule



$C-final

final	0	1
0	56	7
1	3	395

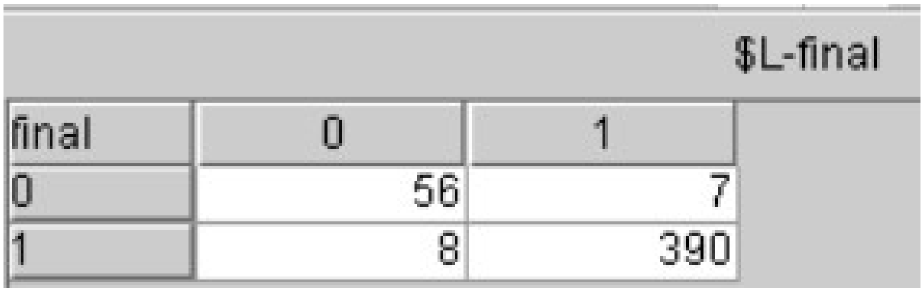
图 12: matrix of decision tree



$N-final

final	0	1
0	23	40
1	4	394

图 13: matrix of neural network



$L-final

final	0	1
0	56	7
1	8	390

图 14: matrix of logistic regression