## HW2

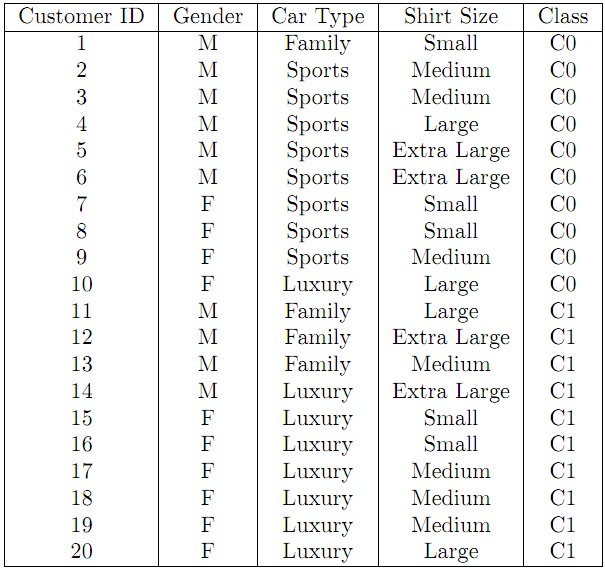
**Due Date: Nov. 11, 2020**

**Submission requirements:**

**Please submit your solutions to our class website.**

**Part I: written part：**

1.



1. Compute the Information Gain for Gender, Car Type and Shirt Size.
2. Construct a decision tree with Information Gain.
3. (a) Design a multilayer feed-forward neural network (one hidden layer) for the data set in Q1. Label the nodes in the input and output layers.
4. Using the neural network obtained above, show the weight values after one iteration of the back propagation algorithm, given the training instance “(M, Family, Small)". Indicate your initial weight values and biases and the learning rate used.

3. Classify the unknown sample *Z* based on the training data set in Q1:

*Z* = (Gender = F, Car Type = Family, Shirt Size = Large). What would a naïve Bayesian classifier classify Z?

4. Suppose that the data mining task is to cluster the following ten points (with(x, y, z) representing location) into three clusters:

A1(4,2,5), A2(10,5,2), A3(5,8,7), B1(1,1,1), B2(2,3,2), B3(3,6,9), C1(11,9,2),C2(1,4,6), C3(9,1,7), C4(5,6,7)

The distance function is Euclidean distance. Suppose initially we assign A1,B1,C1 as the center of each cluster, respectively. Use the K-Means algorithm to show only

1. The three cluster centers after the first round execution
2. The final three clusters

**Part II: lab part**

**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).

**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：

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.**