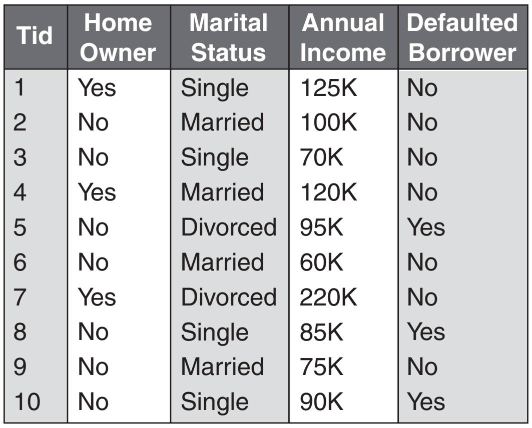
DS 5220 / CS6140-02 - Fall 2017 Olga Vitek

Homework 5

Each part of the problems 5 points
Due on Blackboard before midnight on Tuesday November 28, 2017.

1. *[Note:] This is a writing exercise, it does not intend to use a computer.* Use the dataset below to build a classification tree, to predict whether a loan applicant will repay her loan obligations, or will default.



Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

   (a) Build the maximal tree using Gini index to quantify the impurity of candidate splits

   (b) Create a sequence of sub-trees using cost-complexity pruning and re-substitution error

   (c) Describe in pseudo-code the algorithm that you will use to find the optimal subtree.

2. Consider the dataset "South African Heart Disease" dataset, which was also used in Homeworks 3 and 4. You can use the entire dataset (i.e., no need to partition it into training and validation)

   - Implement a gradient descent algorithm to fit logistic regression that predicts `chd` as function of the continuous predictors only, by viewing it as a single-layer neural network.
   - Compare the parameter estimates, and the training set predictive accuracy, to the results from an existing implementation, e.g. in R.

3. *[Note:] This is a writing exercise, it does not intend to use a computer.* Consider a 5-layer neural network for a binary classification with two predictors. The layers 1, 2, 3, 4 and 5 have respectively 3, 4, 4, 2 and 1 nodes. The last layer of the network is the output layer.

(a) Draw a diagram of the network.

(b) State the total number of parameters in the network.

(c) State the equations for forward propagation in this network, for one observation, in vector/matrix notation. Make sure to clearly define each element of the equation, and its dimensions.

(d) State the equations for forward propagation in this network, for multiple observations, in vector/matrix notation. Make sure to clearly define each element of the equation, and its dimensions.

(e) Repeat (a) and (b) for a same network design, but for a 3-class classification problem.