CSE497 Fall 2018 Assignment 4.

Assigned: Wednesday, October 9, 2018

Due: Wednesday, October 16, 2018

Maximum: 100 point

Note: This assignment is to be done by an individual student, no team work allowed.

Logistic Regression:

1. Consider a logistic regression model parameterized by $\theta_0 = 14$ and $\theta_1 = -0.14$ and the following sample data.

Samples	X	y
x_1	80	1
x_2	20	1
x_3	120	0

   a) (10%) Calculate the probability that y $= 1$ for each x_i of the data set ($h_{\boldsymbol{\theta}}(x)$).
   b) (10%) Calculate the value of objective (cost) function based on idea of SSE,
   c) (10%) Calculate the value of objective (cost) function based on log of Sigmoid.

Decision Tree:

2. Consider the following data set for binary class problem.

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	T	-
T	T	-
T	F	-

a. (15%) Calculate the information gain (based on entropy) when splitting on A and B. Which attribute would the decision tree induction algorithm choose?
b. (15%) Calculate the gain (based on the Gini index) when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

3. The following table summarizes a data set with three attributes A, B, C and two class labels +, − . Build a two-level decision tree.

A	B	C	+	-
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	10
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	20

a. (15%) According to the classification error rate, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table (i.e., count matrix) and the gains in classification error rate.
b. (15%) Repeat for the two children of the root node.
c. (10%) How many instances are misclassified by the resulting decision tree?