CSE 5160 Machine Learning (Fall 2020)

Assignment #1 (Due on Oct 2nd, 2020)

All assignments are to be submitted to Blackboard. Please note that the due time of each assignment is at 11:55 pm (Blackboard time) on the due date. Please make sure to "submit" after uploading your files. Please do not attach unrelated files. You will not be able to change your files after deadline.

1. [20 marks] Manually train a decision tree based on the following dataset.
   a) Please show the detailed process of how to select an attribute to split the instance set at every node. Gini index should be used to measure the impurity (10 marks).
   b) Draw the final decision tree (10 marks).

Employed	MaritalStatus	AnnualIncome	Approved
Yes	Single	High	Yes
No	Married	High	Yes
No	Single	Low	Yes
Yes	Married	High	Yes
No	Divorced	Average	No
No	Single	Low	Yes
Yes	Divorced	High	Yes
No	Single	Average	No
No	Married	Low	Yes
No	Single	Average	No

2. [20 marks] Manually train a decision tree based on the following dataset.

   c) Please show the detailed process of selecting an attribute to split the instance set at every node. Gini index should be used to measure the impurity (15 marks).

   d) Draw the final decision tree and classify the test instance X= (Outlook = rainy, Temperature = hot, Humidity = high, Windy = FALSE) (5 marks).

Outlook	**Temperature**	**Humidity**	**Windy**	**Play**
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

3. [30 marks] (K-Nearest Neighbors) Please show the detailed process of using K-Nearest Neighbor classifier to predict the test instance X= (Speed = 5.20, Weight = 500) is qualified or not, by setting k = 1, 3, and 5, respectively (20 points).

   Before using KNN classifier, please use Min-max normalization (KNN.pdf page 16) to preprocess the attribute values and plot the preprocessed training data set on a 2d plane (Speed - X axis and Weight - Y axis, the instances of class no are labeled by − and the instances of class yes are labeled by + in the plot) (10 points).

ID	**Speed**	**Weight**	**Qualified**
1	2.50	600	no
2	3.75	800	no
3	2.25	550	no
4	3.25	825	no
5	2.75	750	no
6	4.50	500	no
7	3.50	525	no
8	3.00	325	no
9	4.00	400	no
10	4.25	375	no
11	2.00	200	no
12	5.00	250	no
13	8.25	850	no
14	5.75	875	yes
15	4.75	625	yes
16	5.50	675	yes
17	5.25	950	yes
18	7.00	425	yes
19	7.50	800	yes
20	7.25	575	yes

4. [30 marks] (Naive Bayes Classifier) Please show the detailed process of classifying the test instance X = (HM = No, MS = Divorced, AI = 120000) based on the following data sets.

For the continuous attribute **AnnualIncome,** you may use discretization to convert the attribute as binary attribute by setting threshold 91000 or use probability density estimation to estimate the conditional probabilities. (Please refer to note NaiveBayes.pdf page 14)

HomeOwner (HO)	**MaritalStatus (MS)**	**AnnualIncome (AI)**	**Defaulted**
Yes	Single	125000	No
No	Married	100000	No
No	Single	70000	No
Yes	Married	120000	No
No	Divorced	95000	Yes
No	Single	60000	No
Yes	Divorced	220000	No
No	Single	85000	Yes
No	Married	75000	No
No	Single	90000	Yes