Data Mining and Machine Learning

CAP6673
Dr. T.M. Khoshgoftaar
Exam, Spring 2021

Name: _____ Z# _____

Please answer all questions. Show all your work. Good luck!

1. For a discriminant model that classifies high risk and low risk program modules, what is Type I error (false positive), and what is a Type II error (false negative)?

2. What are the differences and similarities (if any) between multiple linear regression models and logistic regression models.

3. Define (or answer) the following terms (questions):

   (a) Resubstitution, subsequent project, data splitting, and cross-validation methods for evaluation of software quality models.

   (b) efficiency (precision) and effectiveness (recall) for classification models.

(c) For linear regression models among the selection methods: greedy, M5, no selection

   i. Which method will use the most number of independent variables

(d) An overfitted numeric prediction or classification model.

4. Suppose we have used a decision tree algorithm to build a classification tree model for detection of Fault-Prone (FP) and Not Fault-Prone (NFP) program modules or objects. We have encoded FP=1 and NFP=0. Also, assume that we are interested in a particular leaf node (NODE t) in the tree model with mean value of 0.25. We would like to classify an object from the test data set which falls in this leaf node (NODE t). Predict this object as FP or NFP for the following cases.

   (a) $\zeta\ (or\ c) = 0.30$

   (b) $\zeta\ (or\ c) = 2.5$

   Show all your work!

5. Given the following FIT and TEST data sets:

FIT data set:
==========

Program Modules	Actual no. of Faults (Y)	Fault-Prone (FP) Not Fault-Prone (NFP)	LOC Indep. Var X
A	1	NFP	23
B	1	NFP	25
C	0	NFP	21
D	2	NFP	28
E	1	NFP	22
F	2	NFP	30
O	2	NFP	32
G	3	FP	34
H	5	FP	35
I	9	FP	40
J	12	FP	55
P	8	FP	38

TEST data set:
==========

K	?	?	20
L	?	?	48
M	?	?	38
N	?	?	28

Where (?) means that we do NOT know its value and have to predict/estimate.

(a) Use CBR with Euclidean Distance as similarity function and unweighted average of THREE most similar cases to predict the number of faults in modules K, L, M, and N. SHOW ALL YOUR WORK!

(b) Use CBR with Euclidean Distance as similarity function and Majority Voting method with THREE most similar cases to classify modules K, L, M, and N as FP or NFP when C = 0.75. SHOW ALL YOUR WORK!

(c) Use CBR with Euclidean Distance as similarity function and Clustering method with THREE most similar cases to classify modules K, L, M, and N as FP or NFP when C = 1.25. SHOW ALL YOUR WORK!

6. Calculate the following performance metrics for the confusion matrix shown below.

* True positive rate.
* True negative rate.
* False positive rate.
* False negative rate.
* Precision.
* Recall
* Accuracy.
* Error rate.

		Actual Class	
		Positive	Negative
Predicted Class	Positive	30	5
	Negative	10	85