Dr Hari Singh

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST -3 EXAMINATION-2021

B.Tech– 7th Semester (Open Elective)

COURSE CODE: 19B1WCI740 MAX. MARKS: 35

COURSE NAME: Introduction to Statistical Learning

COURSE CREDITS: 3 MAX. TIME: 2 Hours

Note: *All questions are compulsory. Carrying of mobile phone during examinations will be treated as case of unfair means.*

Q1.For the following parts, indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.(2)

(a) The number of predictors p is extremely large, and the number of observations n is small.

(b) The variance of the error terms, i.e. $\sigma^2 = Var(\varepsilon)$, is extremely high.

Q2. Carefully explain the difference between the KNN classifier and KNN regression methods. (2)

Q3. Answer on the following confusion matrix describing whether the person(s) will default or not-default on credit card payments (2)

(a) What is the error rate among individuals who defaulted?
(b) What is the error rate among individuals who did not default?
(c) What is the sensitivity? (Percentage of true defaulters that are correctly identified)
(d) What is the specificity? (Percentage of non-defaulters that are correctly identified)

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

Q4. What are the advantages and disadvantages of k-fold cross-validation relative to the validation set approach? (3)

Q5. Which of the following cross-validated prediction error (C_p, AIC, BIC, or adjusted R^2) would result in selection of smaller model and why? (2)

Q6. How does a smoothing spline fit a smoothing curve to a set of data? Explain. (4)

Q7. Which spline better fits a dataset – Cubic Spline or Natural Cubic Spline? Explain. (4)

Q8. In a study of heart disease on 303 patients, the following statistics are obtained.What is the Gini Impurity for the Chest Pain? (4)

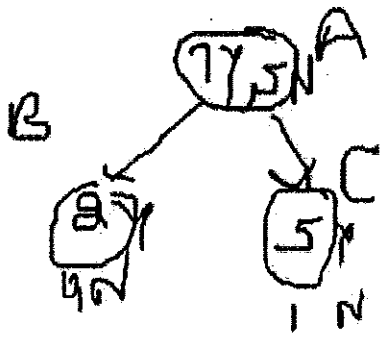
Chest Pain = Yes		*Chest Pain = No*	
Heart Disease = Yes	*Heart Disease = No*	*Heart Disease = Yes*	*Heart Disease = No*
105	*39*	*34*	*125*

Q9. Compute the Information Gain (G) for the following tree where A(7Y, 5N) represents the fact that node A contains 7 Yes's and 5 No's for the given dataset. Similarly, B(2Y, 4N) and C(5Y, 1N) are given. Following information is also provided. *(4)*

Entropy(A)= 0.97

Entropy(B)=0.92

Entropy(C)=0.65



Q10. The Heart Disease dataset predicts whether a patient has heart disease or not on the basis of independent parameters Chest Pain, Blocked Arteries and Patient Weight. Answer the following questions on the following dataset using Adaboost Algorithm.

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease
Yes	Yes	205	Yes
No	Yes	180	Yes
Yes	No	210	Yes
Yes	Yes	167	Yes
No	Yes	156	No
No	Yes	125	No
Yes	No	168	No
Yes	Yes	172	No

(a) Which is the parameter at the root of the decision tree? Explain. *(2)*

(b) Which is the best classification according to the "Amount of Say"? Explain. *(2)*

Q11. What is the significance of slack variables (ε) and tuning parameter (C) in Support Vector Machine? *(4)*