Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES
MSC MATHEMATICAL STATISTICS

END OF SEMESTER EXAM 2018

STA 8302: NON-PARAMETRIC STATISTICS

DATE: 17th August 2018 **TIME:3 Hrs**

ANSWER QUESTION ONE AND ANY OTHER TWO QUESTIONS

QUESTION 1 (30 MARKS)

1. Briefly define the following terms (6 marks)
   a. Hamming distance
   b. Left censoring
   c. Random design regression
2. Show that the Kaplan-Meier and the Nelson-Aalen estimators are equal. (3 marks)
3. An investigation is carried out into the lifestyle of male accountants. A group of 10,000 accountants is selected at random on 1 January 2001. Each member of the sample group supplies detailed personal information as at 1 January 2001 including name, address, date of birth and marital status. The same information is collected as at each 1 January in the years 2002, 2003, 2004 and 2005. The investigation closes in 2005. A PhD student wishes to use the data from this investigation for her thesis on the mortality of married men. Describe the ways in which the available data for this investigation are censored (6 marks)
4. Briefly discuss the following density estimation techniques.
   a. Penalized likelihood Estimators (3 marks)
   b. K-Nearest Neighbors (2 marks)
   c. Kernel Density estimation (2 marks)
5. Let $\{X_i, Y_i\}_{i=1}^{n} = \{(2,6), (8,13), (4,2), (3,1), (6,5)\}$, compute the K-NN estimate $M_k(x)$ for $x = 4 \text{ and } k = 3$ (4 marks).
6. Derive the basis functions for a quadratic Bezier curve and hence or otherwise write down the quadratic Bezier curve function (4 marks)

QUESTION 2

a. Using orthogonal series estimators, with trigonometric basis functions, find the estimator of $\hat{f}(2)$, for $x \in 1,3,4,5$, assuming the smoothing parameter $m = 2$ and $j = 0,1,2$. (9 marks)

b. Discuss three ways you would reduce the dimension of a given data set. (6 marks)

QUESTION 3

a. A data analyst at a leading bank discovers that out of the 30 customers who received business loans in the past month, 15 defaulted. The analyst observed that 2 out of a total of 10 women defaulted while 13 out of 20 males defaulted. On investigating the employment status of all these 30 individuals, 6 out of 14 employed individuals defaulted while 9 out of 16 unemployed individuals defaulted. You are to obtain a decision tree for the above data, determine where the node split will take place using:

i. Gini Index (3 marks)
ii. Entropy (4 marks)
iii. Reduction in variance (4 marks)
iv. Chi-square (4 marks)

QUESTION 4

a. The following data is obtained by a botanist on different plant species.

Height	Leaf area	Stem Radius	Species
10	100	3	M
25	250	5	M
9	90	2	Z
3	55	1	M
11	90	4	Z

Using the K-NN classification technique with Euclidean distance, classify the species with Height=7, Leaf area=75 and Stem radius=1 (9 marks)

b. Discuss the random forests method of classification (6 marks)

QUESTION 5

1. An engineer in a car manufacturing plant experiments on new spark plugs to observe the duration until burning out. Some spark plugs cause engine vibration and are replaced before completely burning out. The following times (in days) until burning out (no asterisk) or replacement before burning out (asterisk) were observed.

17, 13, 15*, 7*, 21, 18*, 5, 18, 6*, 22, 19*, 15, 4, 11, 14*, 18, 10, 10, 8*, 17

a. Define $n, m, t_j, d_j, c_j, and\ n_j$ for these data, assuming that censoring occurs just after the failures were observed (4 marks)

b. Calculate the Nelson-Aalen estimate of $F(t)$ (7 marks)

c. Using Greenwood's formula, estimate $\text{var}\left[\tilde{F}(13)\right]$ (4 marks)