Dr. Jayashree Ramana.

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT
MID SEMESTER EXAMINATION-2015
B. Tech. VI Semester (Bioinformatics)

COURSE CODE: 10B11BI612 MAX. MARKS: 30
COURSE NAME: Machine Learning for Bioinformatics
COURSE CREDITS: 04 MAX. TIME: 2 HRS

Note: All questions are compulsory.

Section A (Marks: 6)

1. Why is it better to use gain ratio instead of entropy in identification trees?
2. How does the inclusion of m parameter improve the accuracy of Fuzzy K-NN compared to crisp K-NN?
3. How is deterministic splitting of data a disadvantage for classification using identification tree?
4. Explain the difference between liberal and conservative performance in ROC analysis.
5. Why is cross-validation used in machine-learning techniques?
6. What is Hebb's rule in artificial neural networks?

Section B (Marks: 9)

1. Given Initial values: $w_0(0)=-0.05$, $w_1(0)=-0.02$, $w_2(0)=0.02$, and $\eta=0.25$. Show how perceptron can be used to solve the OR Gate. You must use atleast two iterations.

2. Calculate the MCC and likelihood ratio for the following confusion matrix. What do you conclude about the applicability of the method for actual prediction problems (is it a good method)?

		Predicted	
		Negative	Positive
Actual	Negative	25	24
	Positive	24	25

3. Narayana et al. derived the following rules for the prediction of cleavage site in HIV protease. (Their dataset consisted of 114 cleavage substrates and 249 non-cleavage substrates.)

   (a) If position 4 is Phenylalanine then cleavage (35/5)

   (b) If position 4 is Leucine then cleavage (38/9)

   (c) If position 4 is Serine then non cleavage (26/1)

   (d) If position 4 is Tyrosine and position 5 is Proline then cleavage (32/5)

   (e) If position 6 is Glutamate then cleavage (44/8)

   Answer the following w.r.t the above.

   (a) Is it possible to calculate the overall accuracy of the above method using the data given? Justify and calculate it if it is possible.

   (b) Explain the scientific rationale behind developing this method.

   (c) Besides prediction of cleavage and non-cleavage, what are the other inferences from this study?

Section C

(Marks: 15)

1. Consider the following gene expression profiles for four patients. Assume that 0 indicates no gene expression and 3 indicates highest gene expression.

   Cluster the genes as well as patients using Euclidean distance. Discuss the biological interpretations for the analysis. (5)

	Patient 1	*Patient 2*	*Patient 3*	*Patient 4*
Gene 1	3	2	1	0
Gene 2	2	1	0	1
Gene 3	3	1	0	2
Gene 4	1	0	1	0
Gene 5	0	0	3	3

2. Consider the data for serum ferritin as a test for iron deficiency anemia. Plot the ROC. (5)

Serum ferritin (mmol/l)	**# with IDA (% of total)**	**# without IDA (% of total)**
< 15	474	20
15-34	175	79
35-64	82	171
65-94	30	168
> 94	48	1332

3. A group of friends in a particular job planned to study the contribution of various factors to the condition of obesity (having more than normal body weight as determined by the age and height). They measured the expression of a particular gene known to be an important marker of obesity (using DNA test) and also considered the other factors like diet, stress (measured by the number of working hours) and exercise. Can you formulate the identification tree (using entropy) for this problem? (5)

Name	Diet	Gene 1	Stress	Exercise	Condition
Sarah	Low	Heavy	Extreme	no	Obese
Dana	Low	Light	Average	yes	Normal
Alex	Average	Moderate	average	yes	Normal
Annie	Low	Moderate	Average	no	Obese
Emily	High	Heavy	Below average	no	Obese
Pete	Average	Light	Below average	no	Normal
John	Average	Heavy	Extreme	no	Normal
Katie	Low	Moderate	Extreme	yes	Normal