INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Mid-Autumn Semester 2022-23

Duration: 2 hrs **Full Marks: 60**

Subject No: ES60011 Subject: Application Of Machine Learning In Biological Systems

Department/Center/School: Energy Science and Engineering

Specific charts, graph paper, log book etc., required: None

Special Instructions (if any): *(1) Answer all the questions. (2) In case of reasonable doubt, make practical assumptions and write that on your answer script. (3) The parts of each question must answered be together. (4) Calculator is allowed.*

---

1. Justify the following statements stating whether it is true/false. Justification is must.

(a) Cross validation is required to verify the correctness of the result.
(b) Data balancing is one of the key to the success of any machine learning method.
(c) Protein secondary structure can be predicted from the protein sequence only.
(d) For protein phosphorylation site prediction problem, data balancing can be done considering some (which?) biological insight.

Marks: 3+3+3+3=12

---

2. TP=10000 ; FP=400; FN=300; TN=1000

A machine learning algorithm generates above mentioned result.

(a) Compute the accuracy of the method.

(b) Compare the accuracy with F1-score. F1 score [F=2×(P ×R)/(P+R)] is the harmonic mean of precision and recall.

Marks: 3+5=8

---

3. Consider the following Multiple Sequence Alignment of proteins as output from Clustal Omega software. Compute the Henikoff Weight for each of the sequences

```
7PUB_10|      QKMLQRKVTCFQ
12AS_1|       DRLSPLHSVYVD
11AS_1|       DRLSPLHSVYVD
6HIV_56|      YFIVKRCTLYFS
```

Marks: 2.5×4=10

---

4. (a) Write down the steps of Principal Component Analysis in connection with feature analysis and dimension reduction.

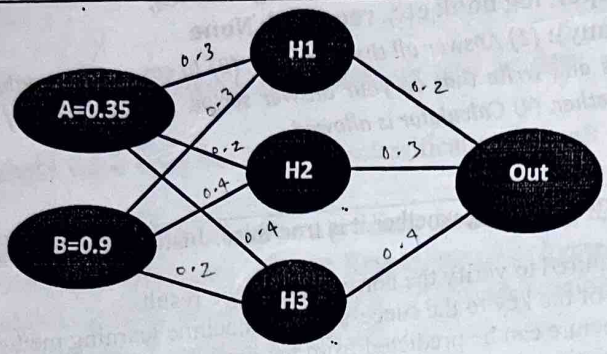
(b) Compute the Eigen values for the following covariance matrix. Next compute the Eigen vector corresponding to the highest Eigen value.

$$\begin{pmatrix} 2 & 2 \\ 5 & -1 \end{pmatrix}$$

$\begin{bmatrix} 0.707 \\ 0.707 \end{bmatrix}$ $\lambda_1 = 4$ $\lambda_2 = -3$

Marks: 4+(4+2)=10

5.



Input Layer **Hidden Layer** **Output Layer**

Initial edge weights:

AH1=0.3 ; AH2=0.2 ; AH3=0.4

BH1=0.3 ; BH2=0.4 ; BH3=0.2

H1Out=0.2 ; H2Out=0.3 ; H3Out=0.4

Assume that the neurons have a Sigmoid activation function

(a) Perform a forward pass on the network and compute the values at each hidden and output node.

(b) Perform a reverse pass (training) once (target=0.5) and compute the new edge weights by clearly marking your error.

(c) Perform a further forward pass with the new edge weights as computed in (b) to compute the modified output value. Comment on the result.

Error 1 → −0.129

Error 2 → −0.121

Please show steps of the calculations. Only final answer will not get any marks.

Marks: 5+10+5=20

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

End-Autumn Semester Examination 2022-23

Date of Examination: 16/11/2022 **Session (FN/AN): AN** **Duration: 3 hrs.** **Full Marks: 80**

Subject No.: ES60011 **Subject: Application of Machine Learning in Biological Systems**

Department/Center/School: School of Energy Science and Engineering

Specific charts, graph paper, log book etc., required

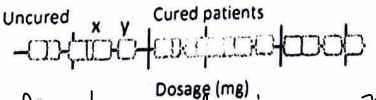
Special Instructions (if any): *(1) Answer all questions, (2) In case of reasonable doubt make practical assumptions and write that on your answer script, (3) The parts of each question must be answered together, (4) Calculator is allowed*

---

1. Mention whether the following statements are true or false and justify your answer with a proper explanation.
   (a) The threshold value does not allow misclassification for soft margin classifiers in the Support vector machine
   (b) Backpropagation learning is based on gradient descent along the error surface
   (c) Random forest algorithm uses learning Rate as one of its hyper-parameter.
   (d) RPKM normalization method is better than TPM normalization to analyze the gene expression data.
   (e) RNA-seq technique examines the expression of all the genes in different tissues..
   (f) The False Discovery Rate (FDR) can control the number of false positives. [2+2+2+2+2+2= 12]

2. Calculate Quality Score and probability of Error for folowing bases during sequencing.
   (a) If the probabilities of base call error are 0.2, 0.015, 0.23 and 0.045, then calculate the respective quality score.
   (b) If the quality score is 15, 20, 26 and 34, calculate the probability of base call error. [3+3=6]

3. (a) What is a polynomial kernel? Let's consider a and b represent the observed number of cured and uncured patients with respect to the drug dosage in milligrams. If a = 3, b = 11, and the polynomial coefficient is ½, calculate the 2-dimensional relationship between the observation 'a' and 'b' using a polynomial kernel.
   (b) Suppose we have two dosage measurements of a drug for heart disease *i.e.*, x and y. If x = 5 and y = 6, what will be the influence value of x and y on each other? Consider the value γ = 2 for scaling the influence with radial kernel. [3+3=6]



4. (a) What is Convolutional Neural Network (CNN)? Mention the usage of CNN
   (b) Explain different layers in Convolutional Neural Network (CNN)
   (c) Consider an input image file that has been converted into a matrix of size 12 x 12 along with a filter of size 3 x 3 with a Stride of 1. Determine the size of the convoluted matrix.
   (d) When we use "valid padding" CNN? Considering an input matrix with 10 x 10 dimensions and a (2 x 2) filter, what will be the dimensions of the output matrix after convolution using valid padding? [2+2+2+2=8]

5. (a) Write the BTW transformation for the sequence: ATCCGATC
   (b) Calculate the False Discovery Rate (FDR) for the following data set:

False positive = 9 and True positive = 65

(c) Calculate Familywise Error Rate (FWER) for the following data:
alpha level for an individual test = 14%; Number of comparisons = 19

$1-(1-\alpha)$

[2+2+2=6]

6. Below table represents the drug effectiveness corresponding to the provided dosage
(a) Considering the threshold of the dosage < 3, calculate the squared residuals (SSR) of the dosage
(b) Calculate the optimal threshold value of the dosage for considering it as the root node in the regression tree.

[2+6=8]

Dosage	Drug effectiveness
2	0
8	12
10	25
13	35
16	98
17	100

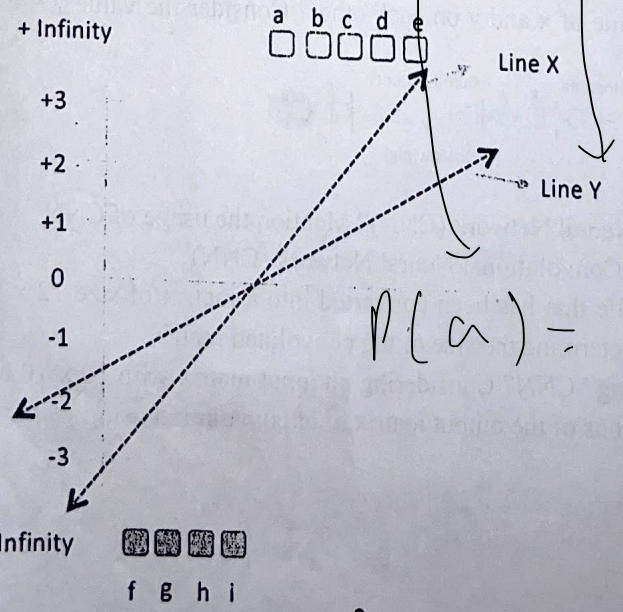
7. Consider a data set with 9 mice where 5 are obese and 4 are not obese. We have calculated the log(odds) of obesity for each candidate date by fitting line **X**.

The log (odds) of each candidate data point for line X is as follows:
log(odds) of **a** = +.3, log(odds) of **b** = +1.2, log(odds) of **c** = +2, log(odds) of **d** = +2.9
log(odds) of **e** = +3.8, log(odds) of **f** = -1.8, log(odds) of **g** = -1.2, log(odds) of **h** = -0.7, log(odds) of **i** = -0.1

Next, we rotate the line (**Y**) and calculate the log(odds of obesity) for all the candidate data.
The log (odds) of each candidate data point for line Y is as follows:
log(odds) of **a** = +0.2, log(odds) of **b** = +0.5, log(odds) of **c** = +0.8, log(odds) of **d** = +1.1
log(odds) of **e** = +1.6, log(odds) of **f** = -1.1, log(odds) of **g** = -0.9, log(odds) of **h** = -0.5, log(odds) of **i** = -0.2

(a) What will be the average probability of the mice being obese
(b) Calculate the likelihood of the obese mice when fitting line X
(c) Calculate the log(likelihood) of all the given data points when fitting line Y
(d) Which line can be considered the best fitting line for the above scenario and why?

[2+2+2+4=10]



$P(a) = \frac{e^{\log odds}}{1+e^{\log odds}}$

8. Here we have provided a dataset of 6 patients with colon cancer and its associated symptoms as observed in these patients.

(a) Calculate the total Gini impurity value of **Fatigue** symptoms for separating the patients with or without colon cancer.

(b) Which symptom can be considered the root node for constructing the classification tree and why?

Symptoms			Disease
Diarrhea	**Loss of Appetite**	**Fatigue**	**Colon Cancer**
Yes	No	No	No
No	Yes	Yes	Yes
Yes	Yes	Yes	Yes
No	No	Yes	No
Yes	No	Yes	Yes
No	No	No	Yes

[3+5=8]

9. Suppose we have an artificial neural network having three inputs with corresponding weights, i.e., $w_1 = 2$; $w_2 = -4$; and $w_3 = 1$.

The step function gives the activation of the unit: $\varphi(v) = 1$ if $v \geq 0$ otherwise, 0

Consider each of the following input patterns and answer the following questions:

Pattern	P_1	P_2	P_3	P_4
X_1	1	0	1	1
X_2	0	1	0	1
X_3	0	1	1	1

(a) Calculate the weighted sum for each pattern.

(b) Calculate the output value y of the given neural network

(c) What is the role of the activation functions in Neural Networks?

[4+2+2=8]

10. (a) Normalize the sequence read count genomic data using RPKM normalization.

Gene Name	Gene Lengths	Sample_1	Sample_2	Sample_3
Gene-1	30 kb	'3216	1560	1250
Gene-2	14 kb	1452	960	852
Gene-3	18 kb	745	486	856
Gene-4	24 kb	1258	1469	751
Gene-5	28 kb	2548	2984	1589
Gene-6	19 kb	879	1546	1852
Gene-7	6 kb	450	145	94
Gene-8	25 kb	2546	2879	1548
Gene-9	9 kb	198	247	254
Gene-10	4 kb	56	98	105

(b) Normalize the sequence read count genomic data for four samples using TPM normalization.

Gene Name	Gene Lengths	Sample_1	Sample_2	Sample_3
Gene-1	40 kb	2100	1240	952
Gene-2	25 kb	1254	540	643
Gene-3	16 kb	950	964	785
Gene-4	24 kb	1120	1840	647
Gene-5	9 kb	582	498	399
Gene-6	7 kb	325	256	478
Gene-7	5 kb	100	350	296
Gene-8	10 kb	507	800	777
Gene-9	4 kb	150	156	198
Gene-10	9 kb	165	289	335

$$\left(\frac{\text{old dim} - \text{filtersize} - 2}{\text{stride}}\right) + 1$$

[4+4=8]

$3 \times 4 \quad 2 \times 2$

$4 - 2 + 1$

$3 - 2 + 1$

2×3