# INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

## End-Autumn Semester Examination 2024-25

**Date of Examination: 21/11/2024** **Session (FN/AN): FN** **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. [1+1+1+1+1+1+1+1+1+1=10]
   a. Pooling layers in CNNs are used only to enhance the resolution of the input data.
   b. Backpropagation is the process by which weights in a neural network are updated during training the model.
   c. The median filter is a non-linear filter used to remove Gaussian noise from images.
   d. The Benjamini-Hochberg method controls the family-wise error rate (FWER).
   e. TPM normalization is preferred over FPKM for analysing gene expression data.
   f. Convolution with stride reduces the spatial dimensions of the input without any information loss.
   g. SVM tries to maximize the margin between the two classes.
   h. Logistic regression can be used when the dependent variable is categorical with one category.
   i. Radial kernel calculates relationship between observation in infinite dimension.
   j. Overfitting occurs when a ML model performs better on the training data than for testing/validation data.

2. (a) Consider a data set of 4 individual where 2 people (Patient a and b) are diagnosed with Colorectal cancer (CRC) and 2 (Patient c and d) are healthy. If the log odds value of candidate data point is -3.1 what will be the candidate probability of sample 'd' being healthy? Also calculate the log of likelihood of overall probability of CRC.
   (b) For a linear regression model, calculate the R2 value, when SS(fit) is "0". [4+2=6]

3. Fill the missing values in table 2 with one time with attribute mean and another time with attribute mean for each class and compare two methods. [2+2=4]

| Patient condition | Dosage used |
|---|---|
| Good | 23 |
| Moderate | 45 |
| Severe | 78 |
| Severe | 94 |
| Good | 12 |
| Moderate | 67 |
| Severe | 95 |
| Good | 20 |
| Severe | ? |
| Moderate | 54 |
| Moderate | ? |
| Good | ? |

4. Suppose you are building a decision tree to classify whether a student will pass or fail an exam based on study hours and attendance. [Note: For decision making nodes use mean value for the given data]



22 Nov 2024, 4:36 am

a. Calculate the gini imputies for each input.
b. Build the whole decision tree. [3+3=6]

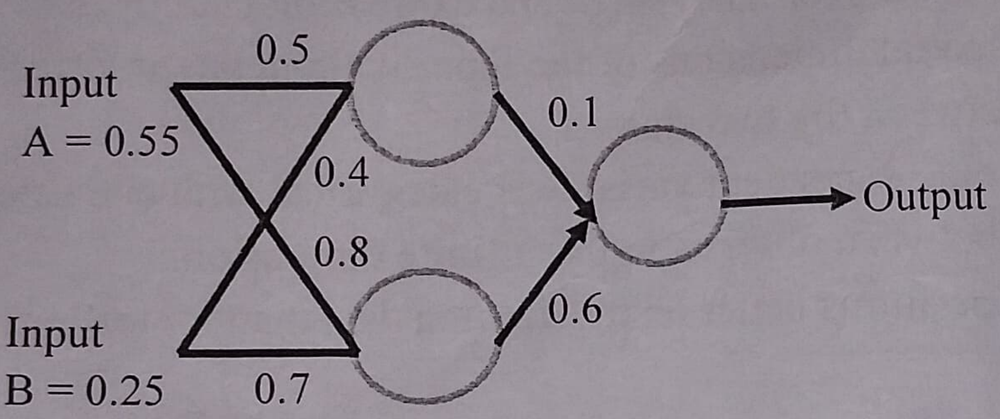
| Student | Study Hours | Attendance (%) | Pass/Fail |
|---|---|---|---|
| A | 2 | 50 | Fail |
| B | 3 | 72 | Fail |
| C | 5 | 70 | Pass |
| D | 7 | 80 | Pass |
| E | 4 | 65 | Pass |
| F | 6 | 85 | Pass |

5. Given an input matrix (image) and a filter (kernel), perform convolution and max pooling operations.
Input Matrix:

$$\begin{bmatrix} 12 & 34 & 56 & 78 & 90 \\ 23 & 45 & 67 & 89 & 12 \\ 34 & 56 & 78 & 90 & 23 \\ 45 & 67 & 89 & 12 & 34 \\ 56 & 78 & 90 & 23 & 45 \end{bmatrix}$$

Filter (Kernel): $\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ Padding: None [3+2=5]

6. Assume that the neurons have a sigmoid activation function ($\lambda = 1.2$)




(a) Perform a forward pass on the network. [3+5+4 = 12]
(b) Perform a reverse pass (training) once (target=0.7).
(c) Perform a further forward pass and calculate current error.

7. Answer the following questions.
(a) 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 = 5, b = 20, and the polynomial coefficient is 1/4, calculate the 2-dimensional relationship between the observation 'a' and 'b' using a polynomial kernel.
(b) Suppose we have two different dosage levels of a medication for treating diabetes, i.e., x=2 and y=3. Using the Radial Basis Function (RBF) kernel, calculate the influence of x on y with $\gamma$=1.4; which scales the influence of the distance between the two dosage levels. [3+3=6]

8. (a) Consider an input image file that has been converted into a matrix of size 14 X 14 along with a filter of size 3 X 3 with a Stride of 1. Determine the size of the convoluted matrix.
(b) Suppose we have an artificial neural network having three inputs with corresponding weights, i.e., w1 = 4; w2 = −2; and w3 = 2. The step function gives the activation of the unit: $\varphi(v) = 1$ if $v > 0$ otherwise, 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 | 0 | 0 |
| $X_2$ | 1 | 1 | 0 | 1 |
| $X_3$ | 0 | 1 | 1 | 1 |

(i) Calculate the weighted sum for each pattern.


22 Nov 2024, 4:36 am

(ii) Calculate the output value y of the given neural network [2+3+2=7]

9. (a) Calculate the Fold-change in expression level of all the genes and identify the least impacted gene?

| Gene ID | Coverage in diseased | Coverage in healthy |
|---|---|---|
| Gene A | 150 | 200 |
| Gene B | 80 | 110 |
| Gene C | 200 | 190 |
| Gene D | 210 | 120 |

(b) If PHRED/ quality Score (Q) is 25, 32, 42; then calculate the Percentage Accuracy.
(c) If the Quality cut-off for SNP base (Q) is 10 and Cut-off for number of reads (C) is 5, then calculate the probability of error/false call.
(d) Calculate the coverage ratio (CR), if the coverage in healthy sample is 4 times than the coverage in diseased sample. [3+3+1+1=8]

10. Answer the following:
(a) Calculate Familywise Error Rate (FWER) for the following data.
alpha level for an individual test = 9% and Number of comparisons = 24
(b) Do BW transformation for the sequence: AGATGATT
(c) Calculate the False Discovery Rate (FDR) for the given data: Out of the 70 cancer patients, the classifier correctly predicts for 40 patients. Out of the 50 healthy individuals, the classifier predicts wrong for 20 cases.
(d) If probabilities of base call error are 0.01, 0.5; then calculate the respective quality scores.
[2+2+2+2=8]

11. Do the following calculations with the provided data sets?

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

| Gene Name | Gene Lengths | Sample I | Sample II | Sample III |
|---|---|---|---|---|
| Gene-I | 6 kb | 810 | 524 | 923 |
| Gene-II | 25 kb | 210 | 316 | 594 |
| Gene-III | 16 kb | 2464 | 3076 | 1208 |
| Gene-IV | 10 kb | 876 | 750 | 954 |
| Gene-V | 20 kb | 435 | 1430 | 1868 |

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

| Gene Name | Gene Lengths | Sample I | Sample II | Sample III |
|---|---|---|---|---|
| Gene-I | 26 kb | 1326 | 4806 | 2044 |
| Gene-II | 16 kb | 496 | 2454 | 1896 |
| Gene-III | 8 kb | 1077 | 1822 | 1960 |
| Gene-IV | 14 kb | 156 | 794 | 880 |
| Gene-V | 34 kb | 612 | 1203 | 1500 |

[4+4=8]

3667 11079 8280

22 Nov 2024, 4:36 am