INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

End-Autumn Semester Examination 2023-24

Date of Examination: 17/11/2023 **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) CNN can learn multiple layers of feature representations of an image by applying filters, or transformations.

(b) The training accuracy increases as the size of the tree grows (assuming no noise).

(c) The False Discovery Rate (FDR) can control the number of false positives.

(d) RPKM and FPKM are the better normalization method than TPM to analyze the gene expression data.

(e) Polynomial Kernel can determine the relationship between observations up to infinite dimensions.

[2+2+2+2+2 = 10]

2. Calculate Quality Score and probability of Error for folowing bases during sequencing.

(a) Consider 117 observations where you have 90 true positives and 27 false negatives. Calculate the sensitivity values of the following observations? $\frac{TP}{TP+FN}$

(b) Consider a set of observations where you have 41 true positives value and 31 false positives value. Calculate the precision for following the observation. $\frac{TP}{TP+FP}$

(c) While calculating the error in gene prediction rate by a sequencing analysis using Bonferroni correction, If the value of $\alpha = 0.1$ and the number of predicted genes are 1000. What will be the Bonferroni corrected p value?

[2+2+2 = 6]

3. Do as directed.

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

Gene ID	Coverage in diseased	Coverage in healthy
Gene A	100	200
Gene B	50	130
Gene C	200	90
Gene D	220	140

(b) If PHRED/ quality Score (Q) is 20, 30, 40, then calculate the Percentage Accuracy. $(1 - 10^{-Q/10}) \times 100$

(c) If the Quality cut-off for SNP base (Q) is 10 and Cut-off for number of reads (C) is 7, then calculate the probability of error/false call. $10^{-\frac{QC}{10}}$

(d) Calculate the coverage ratio (CR), if the coverage in healthy sample is 8 times than the coverage in diseased sample.

[3+2+1+1 = 7]

4. Answer the following questions.

(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 = 2, b = 10, and the polynomial coefficient is ½, calculate the 2-dimensional relationship between the observation 'a' and 'b' using a polynomial kernel.

(b) Let's consider a perceptron having weights corresponding to the three inputs have the following values: w1 = 3; w2 = −2; and w3 = 1 and the activation of the unit is given by the step-function: $\varphi(v) = 1$ if $v \geq 0$ otherwise 0. Calculate the weighted sum for each pattern.

Pattern	P_1	P_2	P_3
Q_1	1	0	1
Q_2	1	1	0
Q_3	0	1	1

(c) Consider an input image file that has been converted into a matrix of size 10 x 10 along with a filter of size 2 x 2 with a Stride of 1. Determine the size of the convoluted matrix.

[3+4+3 = 10]

5. Do as directed.

(d) If the probability of base call error is 0.3, 0.6, and 0.049, then calculate the respective quality score.

(e) If the quality score is 46, 59 and 24, calculate the probability of base call error.

(f) Write the BW transformation for the following sequences.

AATTGACT

GGCTTAGT

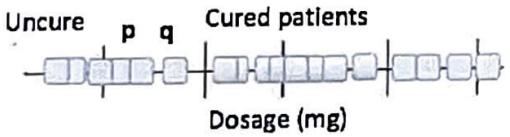
(g) Calculate the Familywise Error Rate (FWER) for the following data.

alpha level for an individual test = 24% and Number of comparisons = 39 [3+3+4+2 = 12]

6. Answer the following questions.

(a) Suppose we have two dosage measurements of a drug for cancer, *i.e.*, p and q. If p = 7 and q = 8, what will be the influence value of p and q on each other?

Consider the value $\gamma = 2$ for scaling the influence with radial kernel



(b) When we use "valid padding" CNN? Considering an input matrix with 12 x 12 dimensions and a (3 x 3) filter, what will be the dimensions of the output matrix after convolution using valid padding?

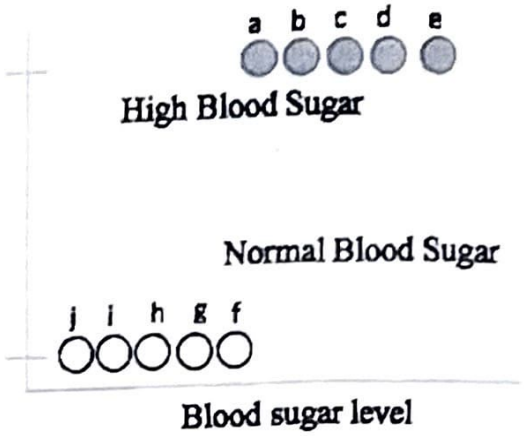
(c) Let's consider the chance of getting lung cancer from cigarette smoking is 7 to 3. Thus, the probability of getting cancer from cigarette smoking is 7/10. Calculate the log (odds) for the above observation. [2+2+1 = 5]

7. Calculate the following questions.

(a) Consider a data set of 10 individual male patients. 5 of them are diagnosed with high blood sugar levels, and 5 individuals have normal blood sugar. The log (odds of high blood sugar) for each candidate data point is as follows:

log(odds) of a = +0.8, log(odds) of b = +1.5, log(odds) of c = +2.9, log(odds) of d = +3.5, log(odds) of e = 4.5, log(odds) of f = -3.1, log(odds) of g = -2.2, log(odds) of h = -1.2, log(odds) of i = -0.8, log(odds) of j = -0.6

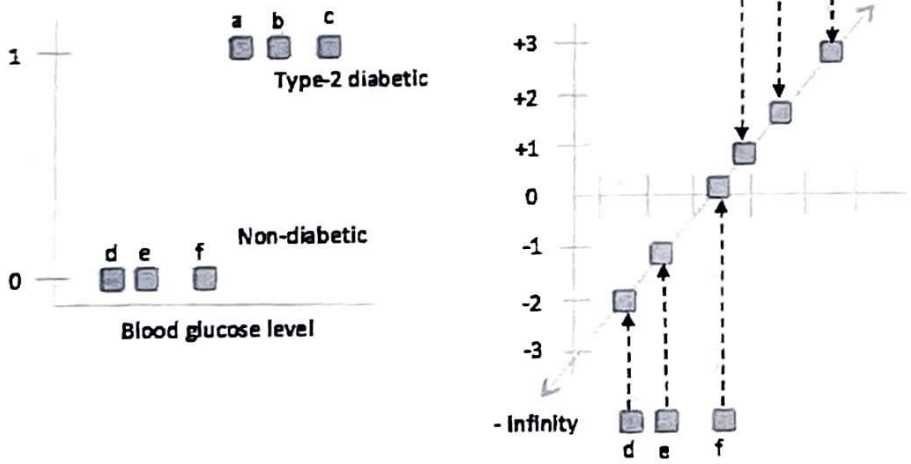
Calculate the log-likelihood of the entire data set



Hint: Calculate the probability of each individual having high BP followed by estimating the individual likelihood for each candidate's data.

(b) Consider a data set of 6 individual where 3 people are diagnosed with type 2 diabetics (T2D) and 3 are non-diabetic control. Below given plot shows the probability of having T2D on the y-axis and blood glucose level on the x axis. Now, we have transformed the probability of T2D into log(odds of T2D) and draw the candidate best fitting line.

If the log odds value of candidate data point is -1.9 what will be the candidate probability of sample 'd' being non-diabetic? Also calculate the log of likelihood of overall probability of T2D.



(c) For a linear regression model, calculate the R2 value, when SS(fit) is "0". [4+4+2 = 10]

8. 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	18 kb	745	486	856
Gene-II	9 kb	198	247	395
Gene-III	28 kb	2548	2984	987
Gene-IV	17 kb	958	745	1023
Gene-V	13 kb	987	1205	1925

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

Gene Name	Gene Lengths	Sample I	Sample II	Sample III
Gene-I	31 kb	1140	4452	2256
Gene-II	12 kb	540	2643	1920
Gene-III	9 kb	964	1785	1823
Gene-IV	11 kb	184	647	990
Gene-V	28 kb	498	1399	1612

(b) Below table represents the drug effectiveness corresponding to the provided dosage.

Dosage	Drug effectiveness
2	0
8	11
10	15
13	16
16	98
22	100

(i) Considering the threshold of the dosage < 3, calculate the squared residuals (SSR) of the dosage.

(ii) Considering the threshold of the dosage < 9, calculate the squared residuals (SSR) of the dosage.

[4+4+4 = 12]

9. Here we have provided a dataset of 8 patients with T2DM (Type 2 Diabetes Mellitus) and its symptoms as observed in these patients.

Blurry vision	Weight Loss	Extreme Thirst	T2DM
Yes	No	No	Yes
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	Yes	Yes	Yes
No	No	Yes	No
Yes	Yes	No	Yes
Yes	No	Yes	No
No	No	Yes	Yes

(a) Calculate the total Gini impurity value of 'Weight Loss' symptoms for separating the patients with or without T2DM.

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

[4+4 = 8]