INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Mid-Autumn Semester Examination 2023-24

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. Write-down the algorithmic steps/pseudo code for computing the followings:

(a) Position-specific scoring matrix from a Multiple-Sequence Alignment.
(b) Computing the frequency of occurrence of each amino acid in a window of ±7 from a protein sequence of length 1200.

10+5=15

Marks: [redacted]

---

2. Kaushik has developed one machine learning-based algorithm for the prediction of the protein phosphorylation site from the protein sequence. The algorithm takes a protein sequence as input and assigns '1' if an amino acid can act as protein phosphorylation site; '0' otherwise. For a protein sequence of length 1200, the algorithm correctly predicted 20 phosphorylation sites out of 25 actual phosphorylation sites and correctly predicted 1000 non-phosphorylation sites.

(a) What is the accuracy of the method? Show the confusion matrix.

(b) Dipa suggested using the domain knowledge. What domain knowledge will be useful in this case?

(c) Using the suggestion of Dipa, Kaushik rerun the algorithm where there are only 200 amino acids are considered for the prediction for the given protein sequence. Kaushik did not find any luck on improving the correctly predicted phosphorylation sites, but, correctly predicted non-phosphorylation sites has reduced to 140. How much improvement on accuracy he has achieved as per her domain knowledge? Suggest by analyzing the confusion matrix and any other metric (if you wish to include) do you wish to use domain knowledge or not?

Marks: (2+2)+2+(2+2)=10

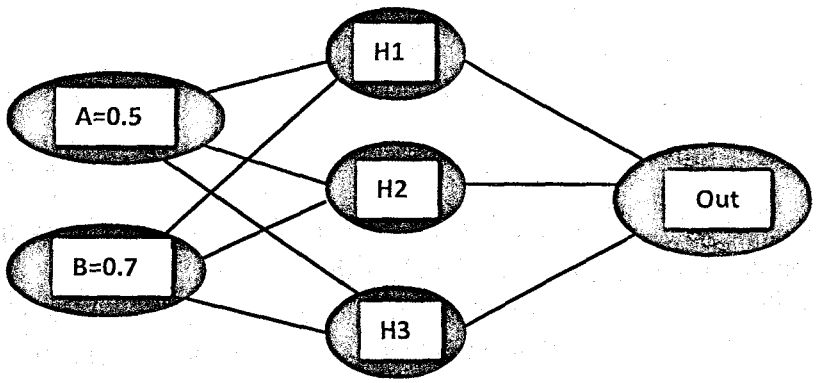
---

3. Consider the following Multiple Sequence Alignment (MSA) of a small stretch of proteins (small stretch is isolated from the original proteins). Clearly, proteins 12AS_1 and 11AS_1 are highly homologous (identical in this case) with each other. In the sequence weight, what will be the effect of including both (12AS_1 and 11AS_1) in the MSA compared to the exclusion of only 11AS_1 from the MSA? Discuss in light of data and fact (Henikoff Weight based analysis) for each of the sequences with and without including 11AS_1 in the MSA.

```
1IND_1|      EAVVTQESALTT
12AS_1|      DRLSPLHSVYVD
11AS_1|      DRLSPLHSVYVD
1PAL_1|      SFAGLKDADVAA
```

Marks: 2.5×4+1×3+1=15

4.



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

Initial edge weights:

AH1=0.1 ; AH2=0.2 ; AH3=0.3
BH1=0.4 ; BH2=0.5 ; BH3=0.6
H1Out=0.7 ; H2Out=0.8 ; H3Out=0.9
Assume that the neurons have a Sigmoid activation function (with λ=1)

(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=0.9) and compute the new edge weights (and hidden node values, if any) by clearly marking your error. Assume learning rate is 0.1.

(c) Perform a further forward pass with the new edge weights and node values as computed in (b) to compute the modified output value.

Note: Please show steps of the calculations. Only final answer will not get any marks. No partial marks if activation function is not applied. Approximate your weights/values till four decimal places for the calculation.

Marks: 5+10+5=20