BIO543: Big Data Mining Healthcare

(4th May 2025, End-Sem Exam)

Maximum Marks: 60 **Duration: 75 Minutes**

Instructions: This question paper have two sections, A and B. Attempt any 14 questions from section A, each question carries 2 marks (Total 28 marks). Attempt any 8 questions from section B, each question carries 4 marks (Total 32 marks). Write all answers in answer sheet only.

Section A

1. **What are the full forms of RNN and LSTM?**
   RNN- Recurrent Neural Network
   LSTM- Long Short-Term Memory
2. **Name any two ANN-based techniques used for generating word embeddings**.
   Word2Vec, GloVE, FastText, Context Learning
3. **Name any two major algorithms used in Hidden Markov Models (HMMs).**
   Forward Algorithm, Viterbi Algorithm, Baum-welch Algorithm
4. **List name of four protein language models.**
   ProtBERT, ESM, ProtT5, ProGen
5. **Define the term "Benign" and "Metastatic" in cancer classification.**
   Benign- Localized, don't expand
   Metastatic- Spread to other parts
6. **What is major difference in FASTA and FASTQC format.**
   1. FASTA files start with a > followed by a sequence name, then the sequence while FASTQC has a different sequence identifier.
   2. FASTQC file contains sequence quality score, while FASTA file doesn't.
7. **Write full form of MIAME and GEO.**
   MIAME- Minimum Information About a Microarray Experiment
   GEO- Gene Expression Omnibus
8. **Compute Jaccard similarity between C_1 = [1,1,0,1,1]; C_2 = [1,1,0,1,1]**
   [1, 1, 0, 1, 1] ∩ [1, 1, 0, 1, 1] = [1, 1, 0, 1, 1]
   [1, 1, 0, 1, 1] ∪ [1, 1, 0, 1, 1] = [1, 1, 0, 1, 1]
   JS = 4/4
   JS=1
9. **Calculate cosine similarity between two vectors [6,0,4,0] and [0,2,0,6]**
   CS= A.B/||A|| ||B||
   CS = 0
10. **What is the difference between MinHashing and LSH?**
    - MinHashing is a technique used to approximate the Jaccard similarity between two sets. It converts sets into fixed-size signatures (minhash signatures) that preserve similarity.

- LSH is a broader framework used to efficiently retrieve similar items from large datasets. It uses hash functions designed so that similar items are more likely to hash to the same bucket.

- MinHashing is a component technique, while LSH is a full similarity search framework.

11. What is the key difference between the K-means and K-medoids clustering algorithms?

- In k-means, each cluster is represented by a centroid, which is the mean of all data points assigned to the cluster. The centroid might not necessarily be one of the data points.

- In k-medoids, each cluster is represented by a medoid, which is the most centrally located point in the cluster, i.e., the data point that minimizes the average dissimilarity to all other points in the cluster. Unlike centroids, medoids must be actual data points.

12. Write full form of BFR and CURE algorithms

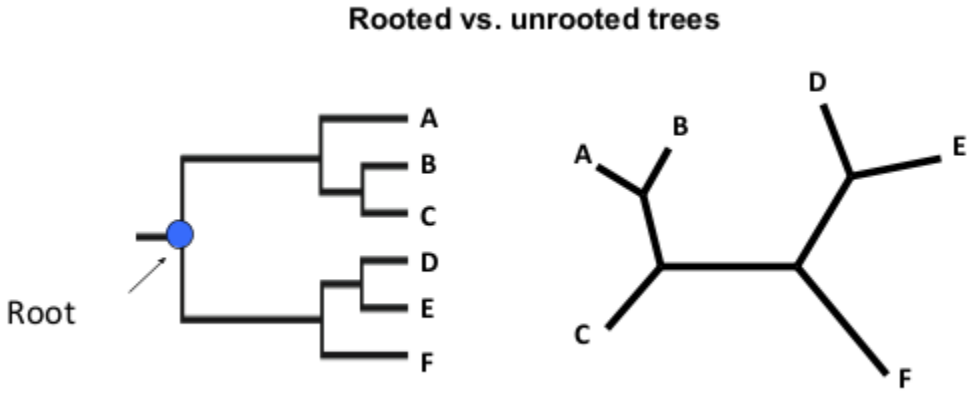
BFR- Bradley-Fayyed-Reina

CURE- Clustering Using Representatives

13. What does BFR algorithm assume about data distribution?

Assumes clusters are normally distributed in each dimension.

14. Draw a rooted and unrooted tree



15. Write full form of UPGMA and NJ method

UPGMA- Unweighted Pair Group Method

NJ- Neighbor Joining

16. Write formula for weighted moving average or autocorrelation.

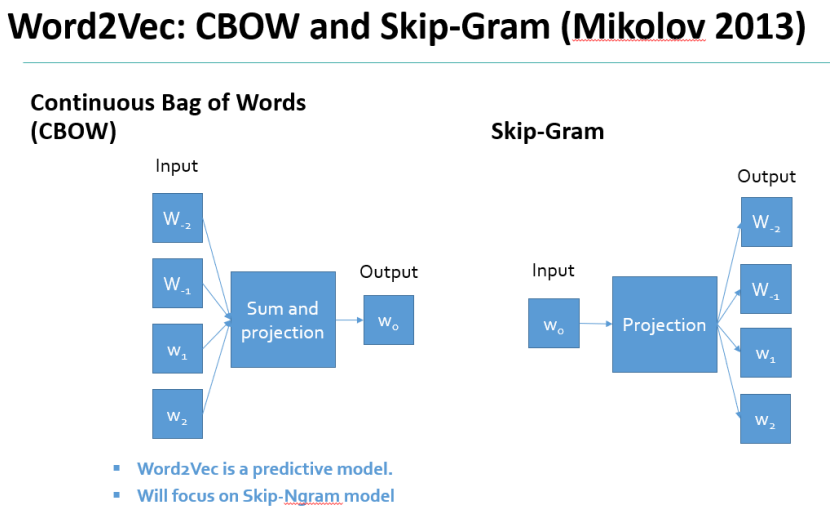
$$x_t = \sum_{i=1}^{N} w_i \times x_{t-i}$$

$$\sum_{i=1}^{N} w_i = 1$$

17. **Which variable was used as the dependent variable in RBpred models?**
    Leaf blast severity was used as the dependent variable
18. **Write difference between univariate and multivariate time series.**
    Univariate: One independent variable
    Multivariate: Multiple independent varibales

Section B

1. **Illustrate the CBOW and Skip-gram models of Word2Vec using diagrams.**

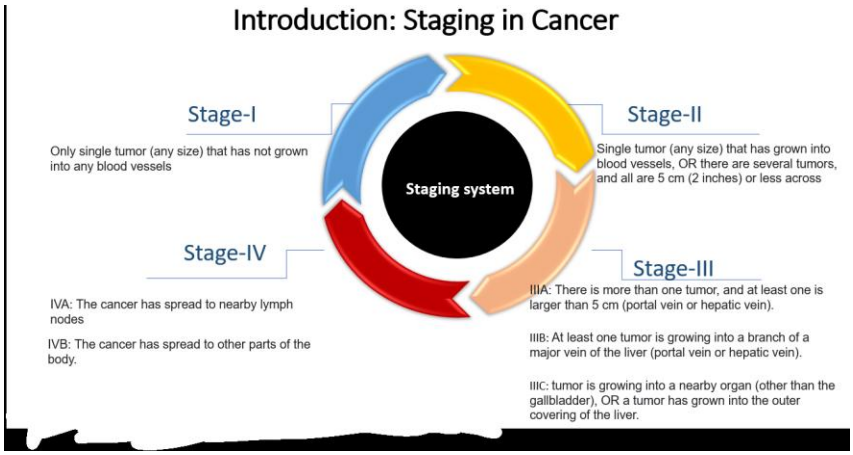


2. **Provide a tabular comparison between GPT and BERT models, minimum four key differences.**

GPT vs BERT Models

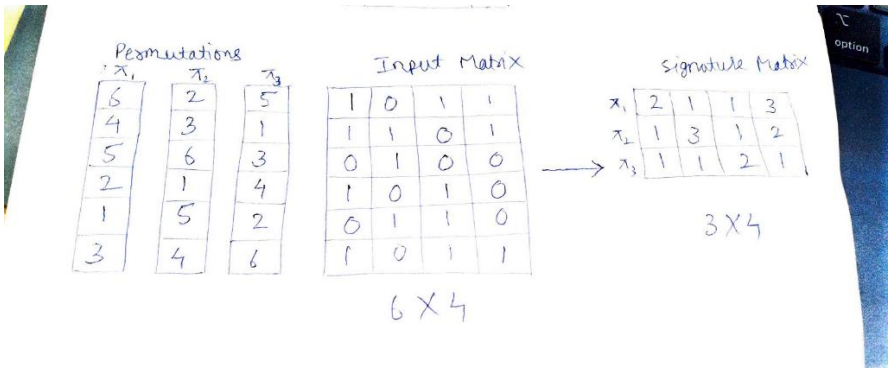
Feature	GPT (Generative Pre-trained Transformer)	BERT (Bidirectional Encoder Representations from Transformers)
Architecture	Decoder-only Transformer	Encoder-only Transformer
Training Approach	Unidirectional (left-to-right)	Bidirectional (uses both)
Objective	Predicts the next word	Predicts masked words
Applications	Text generation (chatbots, summarization)	Text understanding (search engines)
Output Type	Generates coherent text responses	Provides contextual embeddings

3. **Explain all stages of cancer.**



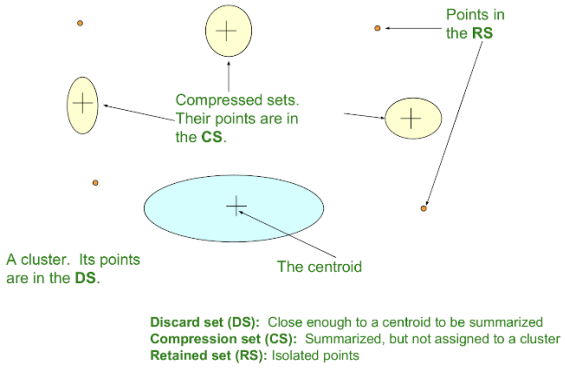
4. Create a min-hash signature matrix of 3×4 from raw matrix 6×4

Can be multiple answers. This is just an example:



5. Describe DS, CS & RS set of BFR, show these sets graphically

- **Discard set (DS):**
  - Points close enough to a centroid to be summarized
- **Compression set (CS):**
  - Groups of points that are close together but not close to any existing centroid
  - These points are summarized, but not assigned to a cluster
- **Retained set (RS):**
  - Isolated points waiting to be assigned to a compression set



6. Describe major steps/pass used in CURE algorithm

2 Pass algorithm. Pass 1:

- **0) Pick a random sample of points that fit in main memory**
- **1) Initial clusters:**
  - Cluster these points hierarchically – group nearest points/clusters
- **2) Pick representative points:**
  - For each cluster, pick a sample of points, as dispersed as possible
  - From the sample, pick representatives by moving them (say) 20% toward the centroid of the cluster

Pass 2:

- Now, rescan the whole dataset and visit each point ***p*** in the data set
- **Place it in the "closest cluster"**
  - Normal definition of "closest": Find the closest representative to ***p*** and assign it to representative's cluster

7. Present a tabular comparison between Clustering and Phylogenetics, minimum four differences

Phylogenetic Trees vs Clustering

Feature	Clustering	Phylogenetics
Objective	Group similar entities	Infer evolutionary relationships and ancestral lineages
Time Dimension	Absent	Present, include divergence times
Biological Clock	Not used	Often uses molecular clock
Tree Interpretation	Dendrogram branches show similarity, not ancestry or time	Tree branches can reflect evolutionary time or mutation rates
Branch Length Meaning	Arbitrary or similarity-based	Reflect genetic change
Rooting	Often unrooted	Rooted to infer common ancestors
Assumptions	Minimal; purely distance/similarity	Evolution, clock models, rate variation
Examples	Clustering gene expression	Reconstructing species trees

8. Briefly describe Homologs, Paralogs and Orthologs, and illustrate using diagram

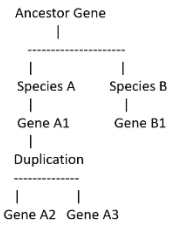
Homologs, Paralogs and Orthologs

Homologs: Genes or proteins in different species that share a common ancestry.
- Have similar sequences, structures, or functions
- Homologs can arise from orthologs or paralogs

Paralogs: Genes that arise from a gene duplication within a species' genome.
- Having similar in sequence but may have diverged in function over time.
- Related but distinct functions within the same organism.

Orthologs: Genes in different species that evolved from a common ancestral
- Retain similar functions across different organisms, differences in sequence
- Play equivalent roles in different species
- Important for inferring evolutionary relationships and gene function

Homologs, Paralogs and Orthologs



Gene A1 and **Gene B1** are **orthologs**: they diverged due to a speciation event.

Gene A1, **A2**, and **A3** are **homologs** (general term for genes with shared ancestry).

Gene A2 and **Gene A3** are **paralogs**: they diverged within the same species due to a gene duplication event.

9. Show exponential smoothing method by example table, write formula

Univariate time series

Week	Temp	Fore
1	25	25
2	27	25
3	30	25.40
4	32	26.32
5	33	27.45
6	35	28.56
7	36	29.85
8	36	31.08
9	?	32.06
10	?	

Exponential smoothing method

$X_t = F_t = F_{t-1} + \alpha (X_{t-1} - F_{t-1})$
OR
$F_t = \alpha X_{t-1} + (1- \alpha) F_{t-1}$

where is α is smoothing function ($0 < \alpha < 1$). F_t is forecasted value for even t .

If α *is 0.2*

$F_1 = X_1 = 25$

$F_2 = 0.2 * X_1 + (1 - 0.2) * F_1 = 0.2 * 25 + 0.8 * 25 = 25 = X_1$

$F_3 = 0.2 * 27 + 0.8 * 25 = 25.40$

$F_{10} = 0.2 * F_9 + 0.8 * F_9 = F_9$

10. Write about RBpred, including input variables and ML techniques for building models.

RB-Pred, a web-based server, is an attempt to forecast leaf blast severity based on the weather variables, which may help farmers and plant pathologists in the timely prediction of rice blast in their areas and ultimately, in their decision-making process.

DataSet:

One year historical data for the year 2000
Four years of data (2001 to 2004) from NATP project
Data was collected from five different locations.
Data on meteorological variables such as temperature max, temperature min, relative humidity max, relative humidity min, rainfall and rainy days per week were recorded daily
Weekly averages of these weather variables were calculated
Severity of disease also measured from leave of rice

Development of models

- **Multiple regression**

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + \ldots\ldots\ldots + b_n X_n$$

where, a = intercept, b_n = slope of line (the partial regression coefficient value), and X_n = independent variable.

- **Leaf blast severity was used as the dependent variable**
- **Weekly average of various weather variables 1 week prior to disease assessment viz. max temp (X1), min temp(X2),**
- **Artificial neural network (ANN), using SNNS 4.2 & MATLAB**
  - **ANN, feed-forward backpropagation neural network (BPNN)**
  - **Generalized regression neural network (GRNN)**
- **Support vector machine (SVM) using SVM_light**
  - **Support vector regression**
- **Evaluation of models**
  - **cross-year models**
  - **cross-location models**