BIO543: Big Data Mining Healthcare

(2nd March 2025, Mid-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 do DNA and RNA stand for?**

   DNA- Deoxyribonucleic Acid

   RNA- Ribonucleic Acid

2. **Arrange the following in ascending order (from small to large): Cell, Tissue and Macromolecules**

   Macromolecules < Cell < Tissue

3. **Is a virus a living or non-living organism? Explain why.**

   Virus is a non-living. They do not have their own biological machinery to replicate.

4. **Name two Mobile apps for Health & Telemedicine.**

   1mg, Aarogya Setu, AIIMS-WHO CC ENBC etc.

5. **Which database was used to create the dataset for developing PPRINT2?**

   BioLiP and PRIDB

6. **What are the full forms of the databases IEDB and PRRDB?**

   IEDB- Immune Epitope Database

   PRRDB- Pattern Recognition Receptor Database

7. **In which year and computer generation was the concept of the microprocessor introduced?**

   1971, 4th generation

8. **Name two procedural programming languages.**

   C, Fortran and Pascal

9. **In Python, if li = [9, 8, 5, -7], what are the values of li[2] and li[-1]?**

   li[2] = 5

   li[-1] = -7

10. **What is sequential data? Provide two examples.**

    It refers to ordered collections of elements or events where the order of occurrence carries significance. Examples: DNA sequence, Time series data.

11. **Calculate the Euclidean and Minkowski distances between the points (5, 3) and (8, 7).**

    Euclidean:

    $$d=\sqrt{(8-5)^2+(7-3)^2}$$

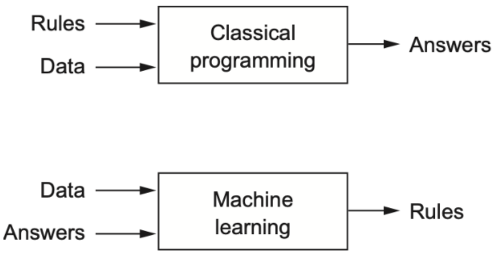
    $$d= 5$$

    Minkowski: at p=1

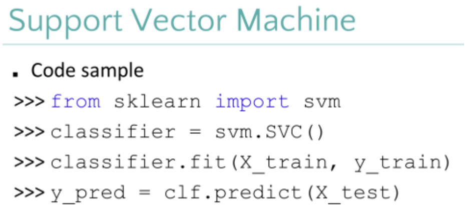
    $$d = (|8-5|^1 + |7-3|^1)$$

    $$d= 7$$

12. **Graphically illustrate the concept of classical programming and machine learning.**



13. Write a Python code to train an SVM classifier using sklearn.



Support Vector Machine

- Code sample

```
>>> from sklearn import svm
>>> classifier = svm.SVC()
>>> classifier.fit(X_train, y_train)
>>> y_pred = clf.predict(X_test)
```

14. Provide the formula for calculating Inverse Document Frequency (IDF).

IDF= log [Nd/(1 + Nt)] , where Nd is total documents, Nt is number of documents contain term t.

15. Name any two structure-based features in the software Pfeature.

Fingerprints, smiles, surface accessibility and secondary structure

16. What is the full form of "ACID" in the context of RDBMS properties?

(Atomicity, Consistency, Isolation, Durability)

17. Name any two vector types used in Mahout.

Dense vector, random access sparse vector and sequential access sparse vector.

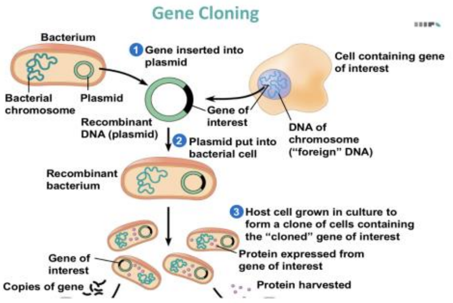
18. **What are the full forms of GD and SGD in the context of SVM for big data?**

GD: Gradient Descent

SGD: Stochastic Gradient Descent

Section B

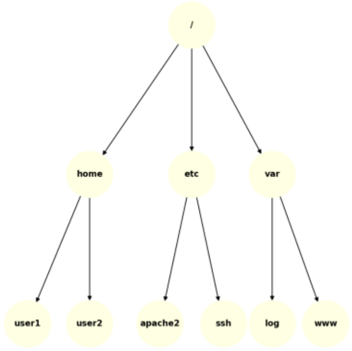
1. Graphically illustrate the process of bacterial gene cloning.


Gene Cloning
Bacterium
1 Gene inserted into plasmid
Cell containing gene of interest
Bacterial chromosome
Plasmid
Gene of interest
Recombinant DNA (plasmid)
DNA of chromosome ("foreign" DNA)
2 Plasmid put into bacterial cell
Recombinant bacterium
3 Host cell grown in culture to form a clone of cells containing the "cloned" gene of interest
Protein expressed from gene of interest
Gene of interest
Copies of gene
Protein harvested

2. **In ThpDB, proteins/peptides are grouped based on their mode of activity. Name these groups.**

- Group 1: Therapeutics with enzymatic or regulatory activity.
- Group 2: Therapeutics with special targeting activity.
- Group 3: Vaccines.
- Group 4: Diagnostic agents

3. **Draw the Linux directory structure, showing three directories and two subdirectories under each.**



4. **Graphically illustrate the concepts of virtual machines and Docker containers.**

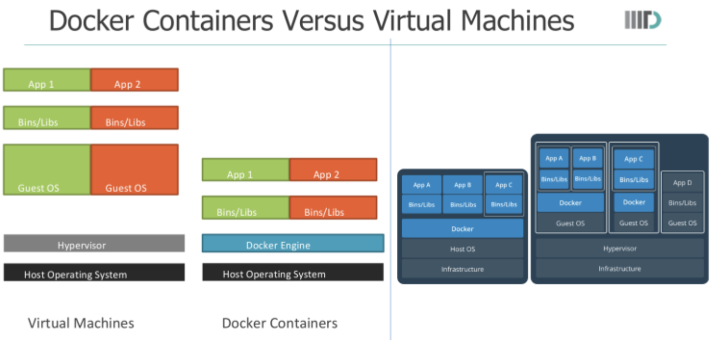
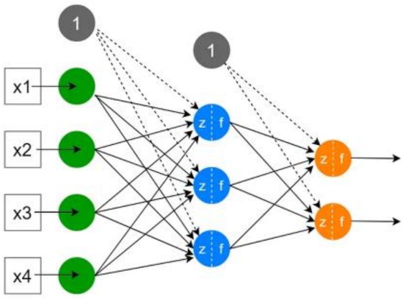


Fig 1 **Fig 2**

--> Either Fig 1 or 2

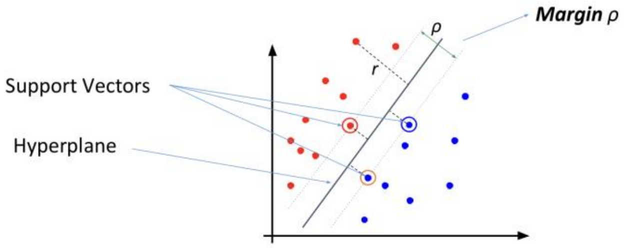
5. **Draw a neural network with 4 input units, 2 output units, and one hidden layer having 3 hidden units.**

Bias unit



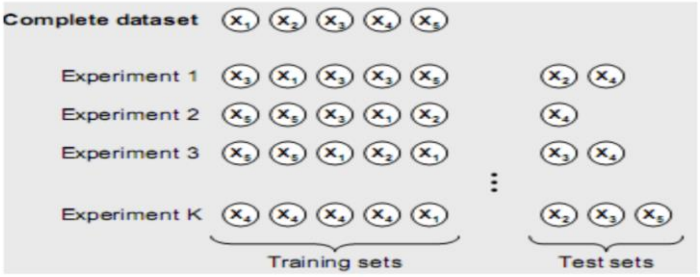
Input **Hidden** **Output**

6. **Graphically represent the concept of SVM, labelling support vectors, the hyperplane, and the margin.**



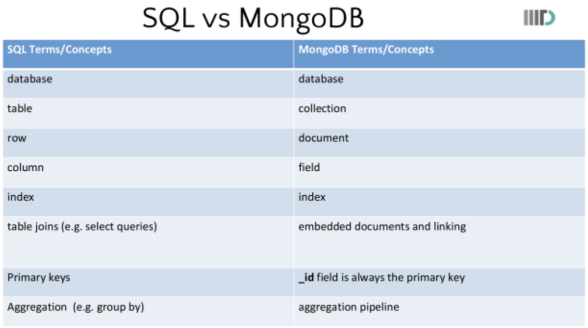
7. **Illustrate the concept of the bootstrapping technique with a diagram.**

- Bootstrapping Technique
  - Statistical technique for estimating distribution



Complete dataset	X_1	X_2	X_3	X_4	X_5			
Experiment 1	X_3	X_1	X_3	X_3	X_5	X_2	X_4	
Experiment 2	X_5	X_5	X_3	X_1	X_2	X_4		
Experiment 3	X_5	X_5	X_1	X_2	X_1	X_3	X_4	
⋮								
Experiment K	X_4	X_4	X_4	X_4	X_1	X_2	X_3	X_5
	Training sets					Test sets		

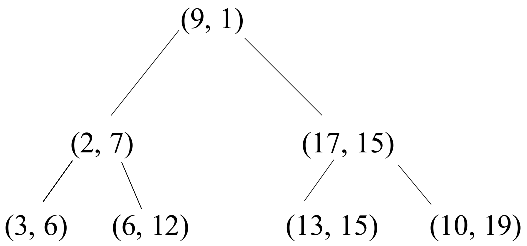
8. Create a comparison table of SQL and MongoDB terms/concepts with four key terms.



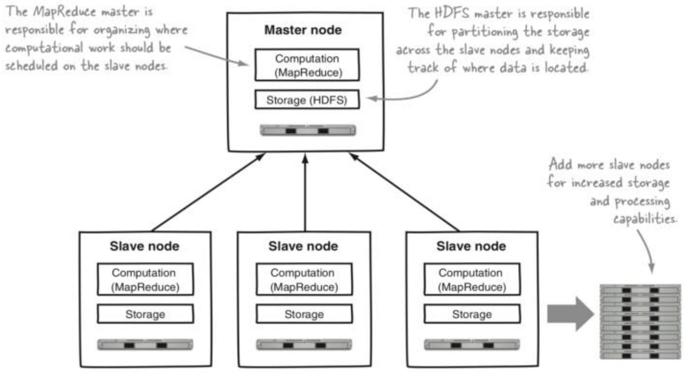
SQL vs MongoDB

SQL Terms/Concepts	MongoDB Terms/Concepts
database	database
table	collection
row	document
column	field
index	index
table joins (e.g. select queries)	embedded documents and linking
Primary keys	**_id** field is always the primary key
Aggregation (e.g. group by)	aggregation pipeline

9. Construct a KD-Tree for the following points: (3,6), (17,15), (13,15), (6,12), (9,1), (2,7), (10,19), (3,6), (17,15), (13,15), (6,12), (9,1), (2,7), (10,19).



10. Draw a high-level Hadoop architecture diagram, labelling the master node, slave node, MapReduce, and storage.



The MapReduce master is responsible for organizing where computational work should be scheduled on the slave nodes.

The HDFS master is responsible for partitioning the storage across the slave nodes and keeping track of where data is located.

Add more slave nodes for increased storage and processing capabilities.

Figure 1.2 High-level Hadoop architecture