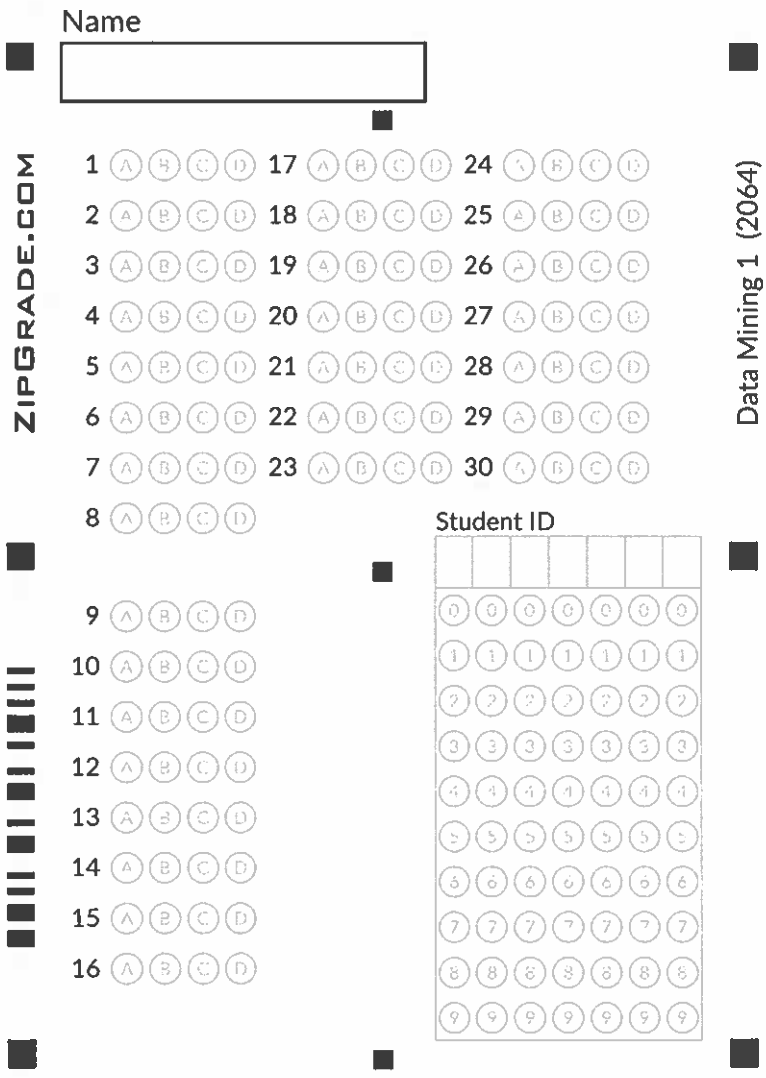

Name
ZIPGRADE.COM
Data Mining 1 (2064)
1 A B C D 17 A B C D 24 A B C D
2 A B C D 18 A B C D 25 A B C D
3 A B C D 19 A B C D 26 A B C D
4 A B C D 20 A B C D 27 A B C D
5 A B C D 21 A B C D 28 A B C D
6 A B C D 22 A B C D 29 A B C D
7 A B C D 23 A B C D 30 A B C D
8 A B C D
Student ID
9 A B C D
10 A B C D
11 A B C D
12 A B C D
13 A B C D
14 A B C D
15 A B C D
16 A B C D

CPSC 552 –Final Exam – Spring 2024

Problem #1: If you are going to use PCA for visualization of TCGA-PANCAN cancer dataset used in some of the assignments, then how many dimensions will you reduce the data to:

a) 5-10 b) 1 c) 100-500 d) 2 or 3

Problem #2: What is the number of dimensions, rows and classes in the TCGA-PANCAN dataset?

a) 1202, 905, 7 b) 20000, 500, 10 c) 20531, 801, 5 d) 1535, 450, 7

Problem #3: Which of the following algorithms are more popular for visualization of genomics data, and for dimensionality reduction.

a) UMAP – genomics visualization, PCA - dimensionality reduction
b) TSNE – genomics visualization, UMAP – dimensionality reduction
c) PCA - genomics visualization, SVD – dimensionality reduction
d) SVD – genomics visualization, PCA – dimensionality reduction

Problem #4: Which of the following statements is false:

a) Deep learning approaches work better than classical machine learning algorithms such as random forests or SVMs when the available training data is large.
b) Deep CNNs are prone to overfitting if trained for too many epochs.
c) Some times PCA can be used before using a CNN for classification to remove redundant features.
d) PCA for face recognition is better than Deep CNN based approaches.

Problem #5: For visualizing the TCGA-PANCAN cancer dataset, which technique performed the worst and best in your visualization assignment.

a) Worst: ISOMAP best: TSNE b) worst: PCA best: UMAP
c) worst:TSNE best: UMAP d) worst: MDS best: TSNE

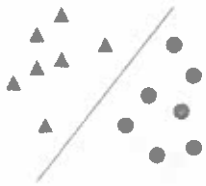
Problem #6: Which activation function is used in the development of the Logistic Regression Algorithm?

a) Tanh b) RELU c) Sigmoid d) Linear

Problem #7: Which of the following statements is false?

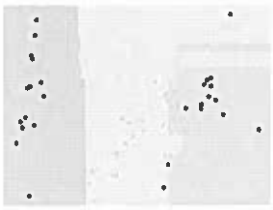
a) Logistic Regression is used for binary classification with a non-linear decision boundary.
b) Logistic Regression uses a Sigmoid function in its development.
c) Logistic Regression and SVM can be adapted to non-linear decision boundaries by using the kernel trick
d) Logistic Regression and SVM can be adapted to multi-class classification by the one-vs-rest technique

Problem #8: For the following classification problem, which technique was most likely used in coming up with the decision boundary?



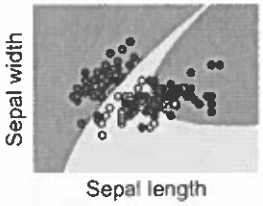
a) SVM b) Logistic Regression c) UMAP d) NN with 5 Neurons and RELU

Problem #9: The following classification and decision boundary is most likely created by:



a) Decision Tree b) SVM c) NN with Sigmoid d) Random Forest

Problem 10: For the following classification and the decision boundary, the most likely technique used was?

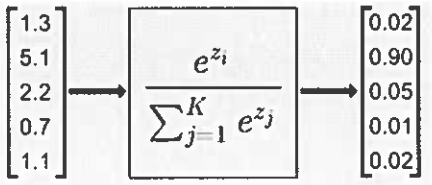


a) T-SNE b) UMAP c) Logistic Regression d) SVM with kernel

Problem #11: Which of the following statements is false:

a) Decision trees can do non-linear classification and use non numerical data
b) Random Forests can do non linear classification
c) Random Forests create smooth decision boundaries.
d) Random Forests perform better than Decision trees

Problem #12: The function producing the following transformation is called:



a) Softmax b) Sigmoid c) UMAP d) Tanh

Problem #13: Suppose you are given data with five features i.e., height, weight, age, blood pressure, and glucose level. You are given data for 500 persons with some having diabetes, some prediabetes, and the rest no diabetes. If you were to create a simple two layer Neural Network model for this dataset, what will be size of input, number of neurons in hidden layer, and number of neurons in output layer.

a) 3, 4, 5 b) 5, 1, 5 c) 5, 4, 3 d) 5, 5, 1

Problem #14: What loss function will you use in the previous problem for predicting diabetes?

a) Mean Square Error b) Mean Absolute Error c) Cross Entropy d) MMSE

Problem #15: Assignment 12 was related to?

a) Cancer prediction b) Recommender System c) Diabetes Prediction d) Stock Prediction

Problem #16: What was the loss function used in the AutoEncoder based Recommender System?

a) Mean Square Error b) Mean Absolute Error c) Cross Entropy d) MMSE

Problem #17: For an AutoEncoder based Recommender System, if there are 100 users and 20 products, the size of the input and the output in the AutoEncoder will be:

a) 100, 20 b) 20, 20 c) 100, 100 d) 20, 100

Problem #18: Assignment 13 was related to?

a) VAE b) Recommender System c) Diabetes Prediction d) GNN

Problem #19: Which of the following is used as a generative model to generate new data:

a) Deep Neural Network b) CNN c) VAE d) GNN

Problem #20: For a Variational Auto Encoder for the PANCAN cancer dataset, the size of the input, the size of latent space, and the output could be:

a) Input: 21532 latent: 10,10 output: 20531 b) Input: 100 latent: 5,10 output: 100
c) Input: 500 latent: 10,10 output: 500 d) Input: 1000 latent: 100,10 output: 1000

Problem #21: Suppose you have two data distributions D1, and D2 as:

D1= [0.1, 0.1, 0.3, 0.4, 0.1] D2= [0.1, 0.1, 0.3, 0.4, 0.1]

The KL divergence between D1 and D2 will be

a) 1 b) 0 c) ∞ d) -1

Problem #22: Which of the following uses KL divergence in the development of its loss function:

a) GNN b) CNN c) AutoEncoder d) VAE

Problem #23: Which of the following statements is false

a) Deep CNNs are better suited for Vision type of data
b) GNN stands for Generative Neural Network
c) VAEs create a distribution for each class in the latent distribution
d) PCA can be combined with Neural Networks for classification

Problem #24: For the PANCAN cancer dataset, which technique may yield the best classification accuracy?

a) CNN b) PCA followed by an NN c) VAE d) GNN

Problem #25: How many assignments were given in the CPSC 552 course?:

a) 10 b) 12 c) 13 d) 14

Problem #26: In the following code:

self.conv1 = nn.Conv1d(1, 6, 15), the 1, 6, and 15 indicate:

a) Input channels, feature maps, convolution kernel size
b) feature maps, convolution kernel size, output channels
c) Input channels, convolution kernel size, feature maps
d) convolution kernel size, input channels, output channels

Problem 27: For the CORA citation dataset that you used in one of the assignments, the number of features and number of classes is:

a) 1433, 7 b) 7, 1433 c) 1000, 784 d) 784, 1000

Problem #28: For analyzing social media data, the best architecture usually is:

a) CNN b) PCA followed by CNN c) GCN d) GAT

Problem #29: For the following code, what will be the value of z:

self.conv1 = GATConv(num_features=100, 30, heads=8)
self.conv2 = GATConv(**z**, 20, 8)

a) 100 b) 240 c) 8 d) 30

Problem #30: Which of the following statements is false

a) GNN can be used to do either node classification, or entire graph classification.
b) For drug discovery, GNNs can model the atomic structure of molecules.
c) GCNs use both CNNs and GNNs
d) Pytorch Geometric library makes programming of GNNs quite straightforward and easy.