INTERNSHIP ASSIGNMENT FILE3 SOLUTION

1. Among the following identify the one in which dimensionality reduction reduces. a) Performance b) statistics c) Entropy d) Collinearity

Ans. The correct option is d) Collinearity.

Dimensionality reduction refers to the process of reducing the number of variables or features in a dataset while retaining as much relevant information as possible. Collinearity refers to the presence of high correlation between two or more predictor variables in a dataset.

By removing collinearity through techniques such as Principal Component Analysis (PCA), we can reduce the number of variables in the dataset while retaining as much information as possible, and therefore reducing its dimensionality.

Performance, statistics, and entropy are not directly related to dimensionality reduction, although they may be indirectly affected by it.

2. Which of the following machine learning algorithm is based upon the idea of bagging? a) Decision Tree b) Random Forest c) Classfication d) SVM

Ans. The machine learning algorithm based upon the idea of bagging is Random Forest.

3. Choose a disadvantage of decision trees among the following. a) Decision tree robust to outliers b) Factor analysis c) Decision Tree are prone to overfit d) all of the above

Ans. C

4. What is the term known as on which the machine learning algorithms build a model based on sample data? a) Data Training b) Sample Data c) Training data d) None of the above

Ans. C

5) Which of the following machine learning techniques helps in detecting the outliers in data? a) Clustering b) Classification c) Anamoly detection d) All of the above

Ans. c) Anomaly detection.

Anomaly detection is a machine learning technique used to identify patterns in data that do not conform to expected behavior. It is specifically designed to detect outliers or anomalies in the

data. Clustering and classification techniques are not specifically designed for outlier detection, although they can indirectly help identify outliers in some cases.

6) Identify the incorrect numerical functions in the various function representation of machine learning. a) Support Vector b) Regression c) Case based d) Classification

Ans. c) Case-based.

Case-based is not a numerical function representation in machine learning. It is actually a type of machine learning algorithm that works by comparing new problem instances with previously seen examples stored in memory.

The other three options, support vector machines (SVMs), regression, and classification, are all numerical function representations used in machine learning. SVMs are used for classification and regression analysis, regression is used to model the relationship between variables and make predictions based on that relationship, and classification is used to classify data into different classes based on its characteristics.

7) Analysis of ML algorithm needs a) Statistical learning theory b) Computational learning theory c) None of the above d) Both a and b

Ans. d) Both statistical learning theory and computational learning theory.

Statistical learning theory and computational learning theory are two main approaches used to analyze machine learning algorithms.

Statistical learning theory focuses on the study of statistical properties of learning algorithms and their ability to generalize from training data to new, unseen data. It provides bounds on the generalization error of a learning algorithm and helps in selecting appropriate models and tuning hyperparameters.

Computational learning theory, on the other hand, focuses on the computational complexity of learning algorithms and their ability to learn efficiently from data. It provides theoretical guarantees on the running time and space requirements of learning algorithms and helps in designing algorithms that are computationally efficient and scalable.

Both approaches are important for understanding the behavior of machine learning algorithms and for designing new algorithms that can learn from large and complex data sets.

8) Identify the difficulties with the k-nearest neighbor algorithm. a) Curse of dimensionality b) Calculate the distance of test case for all training cases c) Both a and b d) None

Ans. c) Both a and b.

The k-nearest neighbor algorithm has the following difficulties:

a) Curse of dimensionality: This refers to the problem of increased sparsity in high-dimensional data, which can make it difficult to find meaningful nearest neighbors. As the number of dimensions increases, the amount of data required to generalize accurately grows exponentially, making it harder to find neighbors that are similar to a given test case.

b) Calculation of distance: The k-nearest neighbor algorithm requires calculating the distance of a test case from all the training cases, which can be computationally expensive for large datasets. This can also make the algorithm slower as the number of dimensions increases.

These difficulties can make the k-nearest neighbor algorithm less effective or even impractical for some datasets. However, there are techniques such as dimensionality reduction and approximate nearest neighbor search that can help overcome these issues.

9) The total types of the layer in radial basis function neural networks is _____ a) 1 b) 2 c) 3 d) 4

Ans. b) 2.

Radial basis function (RBF) neural networks have two types of layers:

1. Input layer: This layer contains the input variables or features of the data.
2. Radial basis layer: This layer applies a radial basis function to the input data and transforms it into a higher-dimensional space. The radial basis function is a kernel function that measures the similarity between the input data and a set of reference points or centroids. The output of the radial basis layer is a set of weighted distances between the input data and the centroids.

There is no hidden layer or output layer in RBF networks, as the output is generated directly from the radial basis layer.

10) Which of the following is not a supervised learning a) PCA b) Naïve bayes c) Linear regression d) KMeans

Ans. d) KMeans.

KMeans is not a supervised learning algorithm, it is an unsupervised learning algorithm used for clustering.

PCA (Principal Component Analysis), Naïve Bayes, and Linear Regression are all supervised learning algorithms. PCA is used for dimensionality reduction and feature extraction, Naïve Bayes is used for classification and probabilistic modeling, and Linear Regression is used for

modeling the relationship between a dependent variable and one or more independent variables.

11) What is unsupervised learning? a) Number of groups may be known b) Features of groups explicitly stated c) Neither feature nor number of groups is known d) None of the above

Ans. c) Neither feature nor number of groups is known.

Unsupervised learning is a type of machine learning in which the goal is to discover patterns, relationships, and structure in data without any prior knowledge or labels. In unsupervised learning, there is no target variable or known output, and the algorithm has to learn from the input data by finding patterns or clusters that are not explicitly labeled or categorized.

Unlike supervised learning, where the goal is to predict a target variable based on labeled training data, unsupervised learning is used for tasks such as clustering, dimensionality reduction, anomaly detection, and generative modeling. In clustering, the algorithm groups similar data points together based on their characteristics, while in dimensionality reduction, the algorithm reduces the number of input features while retaining the most important information. In anomaly detection, the algorithm identifies data points that are significantly different from the rest of the data, while in generative modeling, the algorithm learns to generate new data that is similar to the training data.

12) Which of the following is not a machine learning algorithm? a) SVM b) SVG c) Random Forest Algorithm d) None of the above

Ans. b) SVG.

SVG is not a machine learning algorithm, it is a file format for scalable vector graphics.

The other options, SVM (Support Vector Machine) and Random Forest Algorithm, are both machine learning algorithms used for classification and regression tasks. SVM is a linear and non-linear classifier that separates data points into different classes by finding the best possible hyperplane in a high-dimensional space, while Random Forest Algorithm is an ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting.

13) _____ is the scenario when the model fails to decipher the underlying trend in the input data a) Overfitting b) Underfitting c) Both a and b d) None of the above

Ans. b) Underfitting is the scenario when the model fails to decipher the underlying trend in the input data. Overfitting, on the other hand, occurs when the model learns the noise in the

training data too well and becomes too complex, resulting in poor generalization performance on new, unseen data.

14) Real-Time decisions, Game AI, Learning Tasks, Skill acquisition, and Robot Navigation are applications of ..... a) Reinforcement learning b) Supervised learning c) Unsupervised Learning d) None of the above

Ans. a) Reinforcement learning is commonly used in Real-Time decisions, Game AI, Learning Tasks, Skill acquisition, and Robot Navigation. In reinforcement learning, an agent learns to make a sequence of decisions in an environment to maximize a cumulative reward signal. This makes it particularly useful in situations where an agent must learn through trial and error to achieve a long-term goal.

15) What is called the average squared difference between classifier predicted output and actual output? 55) What is called the average squared difference between 55classifier a) Mean relative error b) Mean squared error c) Mean absolute error d) Root mean squared error

Ans. b) Mean squared error (MSE) is the average squared difference between the predicted output of a classifier and the actual output. MSE is a commonly used loss function in regression problems, where the goal is to minimize the difference between predicted and actual values.

Mean relative error (MRE), mean absolute error (MAE), and root mean squared error (RMSE) are also commonly used metrics to measure the performance of a classifier or regression model, but they are not equivalent to the MSE.

16) Logistic regression is a ..... regression technique that is used to model data having a ..... outcome. a) Linear, binary b) Linear, numeric c) Nonlinear, binary d) Nonlinear, numeric

Ans. a) Logistic regression is a linear regression technique that is used to model data having a binary outcome, i.e., an outcome that takes one of two possible values, such as true/false or yes/no. In logistic regression, a linear function of the input features is transformed using the logistic (sigmoid) function to produce a probability estimate of the positive class. The model is trained using a binary cross-entropy loss function to minimize the difference between the predicted probabilities and the true labels.

17) You are given reviews of few netflix series marked as positive, negative and neutral. Classifying reviews of a new netflix series is an example of A. supervised learning B. unsupervised learning C. semisupervised learning D. reinforcement learning

Ans. The task of classifying reviews of a new Netflix series into positive, negative, or neutral categories is an example of supervised learning.

In supervised learning, the algorithm learns to map inputs (in this case, the text of the reviews) to outputs (the sentiment category of the review) based on labeled examples (the reviews that are already marked as positive, negative, or neutral).

The algorithm is trained on this labeled data to learn the underlying patterns and relationships between the input data and the output labels. Once the model has been trained on the labeled data, it can then be used to predict the sentiment category of new, unseen reviews.

18) Following is powerful distance metrics used by Geometric model A. euclidean distance B. manhattan distance C. both a and b D. square distance

Ans. The answer is C. both Euclidean distance (option A) and Manhattan distance (option B) are powerful distance metrics used by geometric models.

Euclidean distance measures the straight-line distance between two points in a geometric space, while Manhattan distance measures the distance between two points along the axes of the space. Both of these metrics are commonly used in machine learning and data analysis to measure the similarity or dissimilarity between observations or variables.

Square distance (option D) is not a standard distance metric, but it may refer to the squared Euclidean distance or the squared Manhattan distance. However, squared distances are not always useful for distance-based algorithms as they may distort the relative distances between points.

19) Which of the following techniques would perform better for reducing dimensions of a data set? A. removing columns which have too many missing values B. removing columns which have high variance in data C. removing columns with dissimilar data trends D. none of these

Ans. Dimensionality reduction techniques aim to reduce the number of features or variables in a dataset while retaining most of the relevant information.

Out of the given options, removing columns with high variance in data (option B) is likely to perform better for reducing the dimensions of a dataset. This is because high variance in a column suggests that it contains a lot of information and may be useful for modeling, but it may also contribute to overfitting and decrease the performance of some models. Removing such columns can help in reducing the complexity of the model while still retaining most of the important information.

Removing columns with too many missing values (option A) may also be useful for reducing dimensions, but this technique may result in a loss of information if the columns with missing values contain important information.

Removing columns with dissimilar data trends (option C) may not be a good option for dimensionality reduction as it is not clear what is meant by "dissimilar data trends" and it is not a common criterion for removing columns.

Therefore, the correct option is B. removing columns which have high variance in data.

20) Supervised learning and unsupervised clustering both require which is correct according to the statement. A. output attribute. B. hidden attribute. C. input attribute. D. categorical attribute

Ans. Both supervised learning and unsupervised clustering require input attributes (option C).

In supervised learning, the algorithm learns from labeled examples in which both the input attributes and the output attributes are provided. The goal is to learn a mapping from input attributes to output attributes, which can then be used to make predictions on new, unseen examples.

In unsupervised clustering, the algorithm seeks to identify patterns or groupings in a dataset based solely on the input attributes. Since there are no output attributes to guide the clustering, the algorithm must rely solely on the input attributes to identify meaningful patterns or groupings in the data.

Hidden attributes (option B) are not directly observable and cannot be used as input for machine learning algorithms. Categorical attributes (option D) are a type of input attribute that take on discrete values or categories, but they are not the only type of input attribute used in machine learning.

21) What is the meaning of hard margin in SVM? (A) SVM allows very low error in classification (B) SVM allows high amount of error in classification (C) Underfitting (D) SVM is highly flexible

Ans. The meaning of the hard margin in SVM (Support Vector Machines) is that it allows for a very low error in classification.

In SVM, the goal is to find a hyperplane that separates the data into two classes with the largest possible margin between them. The margin is the distance between the hyperplane and the closest data points from both classes, also known as support vectors.

In a hard margin SVM, the goal is to find a hyperplane that perfectly separates the data without any misclassifications. This means that the margin is at its maximum, and any new data point that falls on the correct side of the hyperplane is classified with confidence. However, if there is

any overlap or noise in the data, a hard margin SVM can result in overfitting or poor performance.

In contrast, a soft margin SVM allows for some misclassifications in order to find a more flexible hyperplane that can handle noisy or overlapping data.

Therefore, option (A) SVM allows very low error in classification is the correct meaning of the hard margin in SVM.

22) Increase in which of the following hyper parameter results into overfit in Random forest? (1). Number of Trees. (2). Depth of Tree, (3). Learning Rate (A) Only 1 (B) Only 2 (C) 2 and 3 (D) 1,2 and 3

Ans. The hyperparameter that can result in overfitting in Random Forest is the depth of tree. Therefore, option (B) Only 2 is the correct answer.

Increasing the depth of the tree in Random Forest can lead to overfitting because it allows the model to fit the training data too closely, including noise and outliers. This means that the model will have high accuracy on the training data but may perform poorly on new, unseen data.

Increasing the number of trees can help reduce overfitting in Random Forest by increasing the diversity of the model and reducing the impact of individual trees.

Learning rate is not a hyperparameter for Random Forest. It is a hyperparameter for boosting algorithms such as AdaBoost and Gradient Boosting.

23) Below are the 8 actual values of target variable in the train file: [0,0,0, 0, 1, 1,1,1,1,1],

What is the entropy of the target variable?

(A) -(6/10 log(6/10) + 4/10 log(4/10))

(B) 6/10 log(6/10) + 4/10 log(4/10)

(C) 4/10 log(6/10) + 6/10 log(4/10)

(D) 6/10 log(4/10) – 4/10 log(6/10)

Ans. The entropy of a binary target variable with two possible outcomes (0 or 1) can be calculated as follows:

Entropy = -p(0) log2 p(0) - p(1) log2 p(1)

where p(0) is the proportion of instances with outcome 0, and p(1) is the proportion of instances with outcome 1.

In this case, there are 6 instances with outcome 0 and 4 instances with outcome 1. Therefore:

p(0) = 6/10 = 0.6 p(1) = 4/10 = 0.4

Plugging these values into the entropy formula, we get:

Entropy = -(0.6 log2 0.6 + 0.4 log2 0.4)

Simplifying this expression, we get:

Entropy = -(0.6 * (-0.73696) + 0.4 * (-1.32193))

Entropy = -(-0.44218 + (-0.52877))

Entropy = 1.04595

Therefore, the entropy of the target variable is approximately 1.04595.

The closest answer choice to this value is (A) -(6/10 log(6/10) + 4/10 log(4/10)).

24) Lasso can be interpreted as least-squares linear regression where (A) weights are regularized with the l1 norm (B) weights are regularized with the l2 norm (C) the solution algorithm is simpler

Ans. Lasso can be interpreted as least-squares linear regression where weights are regularized with the l1 norm. Therefore, option (A) is the correct answer.

In Lasso regression, the objective is to minimize the sum of squared errors between the predicted and actual values, subject to a constraint on the sum of the absolute values of the weights (i.e., the l1 norm). This constraint promotes sparsity in the weight vector, resulting in a simpler and more interpretable model that includes only the most important features.

In contrast, Ridge regression is a variant of linear regression where the weights are regularized with the l2 norm. This regularization term penalizes large weights and tends to distribute the weight values more evenly across all features, without necessarily forcing any of them to be exactly zero.

Therefore, Lasso and Ridge regression differ in the type of regularization they apply to the weight vector.

25) Consider the problem of binary classification. Assume I trained a model on a linearly separable training set, and now I have a new labeled data point that the model properly categorized and is far away from the decision border. In which instances is the learnt decision boundary likely to change if I now add this additional point to my previous training set and re-train? When the training model is, (A) Perceptron and logistic regression (B) Logistic regression and Gaussian discriminant analysis (C) Support vector machine (D) Perceptron

Ans. If the trained model is a perceptron or logistic regression, the decision boundary is likely to change if a new data point is added to the training set. Therefore, option (A) Perceptron and logistic regression is the correct answer.

Perceptron and logistic regression are linear models that rely on finding a hyperplane that separates the two classes in the feature space. Adding a new data point to the training set can cause the hyperplane to shift or rotate, potentially changing the decision boundary.

In contrast, support vector machine (SVM) and Gaussian discriminant analysis (GDA) are also linear models that can handle linearly separable data. However, they are less sensitive to small changes in the training set and are more robust to overfitting.

In summary, the decision boundary is more likely to change when a new data point is added to the training set for linear models such as perceptron and logistic regression, and less likely to change for models such as SVM and GDA.

26) Assume you've discovered multi-collinear features. Which of the following actions do you intend to take next? (1). Both collinear variables should be removed. (2). Instead of deleting both variables, we can simply delete one. (3). Removing correlated variables may result in information loss. We may utilize penalized regression models such as ridge or lasso regression to keep such variables. (A) Only 1 (B) Only 2 (C) Either 1 or 3 (D) Either 2 or 3

Ans. When multicollinearity is detected among the features, there are several actions that can be taken. Option (D) Either 2 or 3 is the correct answer as both options are valid approaches.

1. Both collinear variables should be removed: This is one option to address multicollinearity. Removing one or both of the collinear variables can reduce the redundancy in the data and improve model performance. However, this approach may result in a loss of information, especially if both variables are important predictors.
2. Instead of deleting both variables, we can simply delete one: This is another option to address multicollinearity. Removing one of the collinear variables can reduce the redundancy in the data without losing too much information. The choice of which variable to remove can be based on domain knowledge or statistical significance.
3. Utilize penalized regression models such as ridge or lasso regression to keep such variables: This is another approach that can be used to address multicollinearity. Penalized regression models such as ridge or lasso regression can be used to impose a penalty on the coefficients of the collinear variables, effectively reducing their impact on the model. This approach can help retain important information from the collinear variables while reducing their redundancy.

Ultimately, the choice of which approach to take will depend on the specific situation and the goals of the analysis.

27) A least squares regression study of weight (y) and height (x) yielded the following least squares line: y = 120 + 5x. This means that if the height is increased by one inch, the weight should increase by what amount? (A) increase by 1 pound (B) increase by 5 pound (C) increase by 125 pound (D) None of the above

Ans. According to the given least squares line, y = 120 + 5x, where y represents weight and x represents height. The coefficient 5 indicates that for a one-unit increase in x (i.e., for a one-inch increase in height), y (i.e., weight) is expected to increase by 5 units.

Therefore, the answer is (B) increase by 5 pound.

28) The line described by the linear regression equation (OLS) attempts to ____? (A) Pass through as many points as possible. (B) Pass through as few points as possible (C) Minimize the number of points it touches (D) Minimize the squared distance from the points

Ans. The line described by the linear regression equation (OLS) attempts to minimize the squared distance from the points.

In other words, the OLS linear regression line is the line that best fits the given data by minimizing the sum of the squared differences between the predicted values and the actual values. This means that the line is chosen to minimize the overall distance between the line and the data points, with the distance being measured as the vertical difference between the predicted and actual values.

Therefore, the correct answer is (D) Minimize the squared distance from the points.

29) For two real-valued attributes, the correlation coefficient is 0.85. What does this value indicate? (A) The attributes are not linearly related (B) As the value of one attribute increases the value of the second attribute also increases (C) As the value of one attribute decreases the value of the second attribute increases (D) The attributes show a curvilinear relationship

Ans. A correlation coefficient of 0.85 for two real-valued attributes indicates that there is a strong positive linear relationship between the two attributes.

A correlation coefficient is a measure of the strength and direction of the linear relationship between two variables. The value of the correlation coefficient ranges from -1 to +1. A value of +1 indicates a perfect positive linear relationship, while a value of -1 indicates a perfect negative linear relationship. A value of 0 indicates no linear relationship between the two variables.

Therefore, the correct answer is (B) As the value of one attribute increases the value of the second attribute also increases.

30) Which neural network architecture would be most suited to handle an image identification problem (recognizing a dog in a photo)? (A) Multi Layer Perceptron (B) Convolutional Neural Network (C) Recurrent Neural network (D) Perceptron

Ans. The most suited neural network architecture to handle an image identification problem like recognizing a dog in a photo is a Convolutional Neural Network (CNN).

Convolutional Neural Networks are specifically designed for image recognition tasks and have become the state-of-the-art approach for image classification problems. They have several layers of convolutional, pooling, and activation functions that enable them to learn and extract features from images. The convolutional layers perform feature extraction by applying filters to the input image and generating feature maps. The pooling layers then downsample the feature maps, reducing the spatial dimensions and providing invariance to small translations in the image. Finally, the fully connected layers perform classification based on the learned features.

On the other hand, Multi Layer Perceptron (MLP) is a basic feedforward neural network that can be used for various tasks, including image classification, but they are not specialized for image processing. Similarly, Recurrent Neural Networks (RNNs) are designed for handling sequential data, such as natural language processing, and may not be suitable for image recognition tasks. Perceptron is a single layer neural network and is not suitable for complex image recognition problems.

Therefore, the correct answer is (B) Convolutional Neural Network (CNN).