Math506-241-KFUPM-Dr. Jamal

Name/ID: ***Enjoy it*** 😊

Questions 1-15: True (A) of False (B)

Q1: Length, weight, and density are examples of numerical type and interval measurement scales.

Q2: An example of a numerical measurement level is Discrete.

Q3: Variables arising from a counting process are called continuous.

Q4: Categorical data, where ordering is important, are nominal.

Q5: The First Quartile is equal to the 25^{th} percentile.

Q6: The most common measure of central tendency that is less sensitive to outliers is the Mean.

Q7: When a sample data has a perfect normal distribution, then Mean=Median.

Q8: Measures of variation include Range, Variance, Standard Deviation, and Coefficient of Variation.

Q9: A data analyst should make sure that the test data set is balanced to avoid inaccuracies.

Q10: Data Dredging is an essential step in the EDA phase.

Q11: Resampling is a technique used for balancing the training set.

Q12: When the model is highly nonlinear, a data analyst may need 90% of the data for training the model.

Q13: Data science methodology follows the concept of the statistical inference approach.

Q14: High correlations between predictors in linear regression should be avoided.

Q15: The solution of the normal equation is not unique when the coefficient matrix has linearly dependent columns.

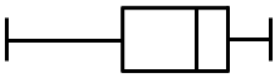
Q16: The measure that shows the variation relative to the mean is the

a) Standard Deviation b) Variance c) Mode d) Coefficient of Variation

Q17: If Median=200, Mode=200, and Mean=300, then the shape of distribution is

a) Symmetric b) Left-Skewed c) Right-Skewed d) Ordinal

Q18: The following Box Plot is



a) Left-Skewed b) Right-Skewed c) Symmetric d) Nominal

Q19: A normally distributed sample covers 99.7% within how many standard deviations from the mean?

a) 1.5 b) 2 c) 2.5 d) 3

Q20: Of invoices with errors, the percentage of the Small Amount is

a) 1 b) 100 c) 25 d) 16.7

Q21: The proportion of invoices with no errors is

a) 50% b) 75% c) 83.3 d) 500

	No Errors	Errors
Small Amount	300	100
Medium Amount	200	0

Q22: A training set has a binary target with 250-Yes records, and 700-No records. How many Yes-records do you resample in order to have the Yes-records 35% of the rebalanced training set.

a) 120 b) 124 c) 127 d) 136

Q23: $H(X) = -\Sigma_j \, p_j log_2(p_j)$ is called the

a) Enropy of X b) Gini Index c) Information Gain d) Heaviside function

Q24: The minimum value of $H(X) = -\Sigma_j \, p_j log_2(p_j)$ is achieved when values of p_j equal to

a) 1 or 2 b) 0 or 1 c) 0.5 or 2 d) 1 or infinity

Q26: Random Forests determines the final classification of a record by considering

a) Multiple trees b) More nodes c) More leaf nodes d) An optimal root node

Q27: A Modeling Technique, that takes into account a number of models' votes for classification, is called

a) C5.0 b) Ensemble c) Decision Trees d) CART

Q28: Which object sets the number of decision trees in Random Forests to be 20?

a) n_estimators=20 b) num_trees=20 c) trees_forests=20 d) criterion=20

Q29: Given: TP=30, FP=20, TN=40, and FN=10, then the Precision of All_Positive_Model is

a) 10% b) 40% c) 50% d) 60%

Q30: Given: FP=10, FN=20, TAP=160, and TAN=140, then Specificity×Sensitivity =

a) 75.2% b) 81.3% c) 85.7% d) 90.1%

Q31: Use the data-driven cost matrix to calculate the profit per record.

TN=40; cost=0$	FP=30; cost=3$
FN=10; profit=1$	TP=10; profit=20$

a) 1.20$ b) 1.33$ c) 1.50$ d) 1.56$

Q32: Model _____ makes sure that the model's results are consistent between the training and test data sets

a) preparation b) evaluation c) validation d) reduction

Q33: Which of the measures combines precision and recall in a single measure?

a) Accuracy b) Sensitivity c) F_1 d) SSE

Q34: Using the given table, the value of net_A is equal to

	W0A=0.4
X1=1	**W1A=0.3**
X2=2	**W2A=0.5**

a) 1.3 b) 1.7 c) 1.8 d) 2.1

Q35: Given $net_B = 3$, then the output of node B is equal to

a) 0.81 b) 0.95 c) 0.77 d) 0.99

Q36: Let X be standardized data, then the sigmoid value of the mean $\bar{X}$, that is $f(\bar{X}) = \frac{1}{1+e^{-\bar{X}}}$ is equal to

a) 0 b) 1 c) 0.5 d) 0.68

Q37: The behavior of the sigmoid function near $x = 0$ is almost

a) Linear b) Highly Nonlinear c) Constant d) Curvilinear

Q38: Clustering refers to the grouping of records of

a) Similar objects b) Different objects c) Mixed objects d) Normalized objects

Q39: When using Clustering, there is

a) a target variable b) no target variable c) no data variation d) no data skewness

Q40: The number of clusters in the K-Means Clustering Algorithm is

a) initially 2 b) initially 3 c) initially 5 d) set by the user

Q41: The initial cluster center locations, in K-Means, are

a) random b) the mean c) 0 d) σ

Q42: The cluster centroids, in K-Means, are calculated using the

a) median b) mode c) mean d) σ

Q43: A way of predicting an output variable from one or more input variables is

A: correlation	C: regression
B: PCA	D: MSE

Q44: Data Visualization is done using

A: matplotlib only	C: seaborn only
B: Graphix	D: both matplotlib and seaborn

Q45: In linear regression models, an iterative algorithm that is used to minimize the least squares error is

A: Normal Equation	C: Intercept
B: PCA	D: Gradient Descent

Q46: The linear regression coefficients (W) can be obtained by the solution of the normal equation, $W =$

A: $(X^T X)^{-1} X^T Y$	C: $X^{-1} Y$
B: $X^T Y$	D: $(X^T X)^{-1} W$

Q47: Which of the following transforms a categorical attribute ‘Education’ using one hot encoding?

A: df['Education'].replace({1:'A',2:'B',3:'C'})	C: df['Education'].one_hot({1:'A',2:'B',3:'C'})
B: pd.get_dummies(df.columns=['Education'])	D: pd.one_hot(df,columns=['Education'])

Q48: A logistic regression model is trained using

A: model.fit(X_test, y_test)	C: model.fit(X_train, y_train)
B: model.train(X_train, y_train)	D: model.solve(X_train, y_train)

Q49: Ridge and Lasso Models are imported using

A: sklearn.penalized_model	C: sklearn.linear_model
B: sklearn.regulazor_model	D: sklearn.reduced_model

Q50: In the code: regr = Ridge(alpha=450), the value of alpha sets the

A: regularization coefficient	C: intercept parameter
B: maximum number of iterations	D: number of sample records

Q51: Which code that is used for building training and testing sets is

A: training_testing_builder (X, y, test_size=0.30)	C: train_test_split (X, y, test_size=0.30)
B: training_testing_spliter (X, y, test_size=0.30)	D: training_testing_spliting (X, y, test_size=0.30)

Q52: MSE or SSE can be used in

A: Lasso Model	C: Linear Regression Model
B: Ridge Model	D: All of the Above

Q53: Classification problems are a type of

A: Supervised Learning	C: PCA Models
B: Unsupervised Learning	D: All of the Above

Q54: Descriptive analysis uses

A: multivariate plots	C: correlations
B: regressions	D: gradient descent iterations

Q55: The correlation of two attributes determines

A: which one should be the target variable	C: the intercept
B: the right choice of the z-score	D: the strength of their linear relationship

Q56: Which of the following is the strongest correlation score?

A: 0.1	C: - 0.95
B: 0.90	D: 0.83

Q57: Which of the following sentences is True about the principal component analysis?

A: It is used for data visualization and exploration.	C: It projects the input data into a lower dimensional. linear space
B: It is used to reduce the number of attributes.	D: All of the Above.

Q58: Which of the following is True about linear regression?

A: It is used for prediction.	C: It defines a relationship between independent and dependent variables.
B: It defines the dependent variable as a linear function of independent variables.	D: All of the Above.

Q59: In a single input, single output regression $y = \beta_0 + \beta_1 x$, the parameters (β_0, β_1) refer to

A: $(y - intercept, slope)$	C: $(x - intercept, correlation)$
B: $(x - intercept, MSE)$	D: (pca_1, pca_2)

Q60: Given that the correlation of the input variable and output variable is negative, then the regression equation has

A: a positive intercept.	C: a positive slope.
B: a negative intercept.	D: a negative slope.