Seat No. 379372

Department of Statistics
School of Mathematical Sciences
Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon
M.Sc. (Statistics) Term End Examination
ST-404 (A): Data Mining

N. B. (1) Do not write anything on the question paper except your seat number.
(2) All questions are compulsory and carries equal marks.

Max. Marks: 60 Time: 3 hours

Q.1. Attempt any three of the following. [04 marks each]

a) For each of the following scenarios, state it is an example of supervised or unsupervised learning. Explain your answers.

i) A recommendation system on an online retailer that recommends to users what to buy based on their past purchasing history.

ii) A system in a credit card company that captures fraudulent transactions.

b) What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

c) Consider the model $Y = f(X) + \varepsilon$. In the context of statistical learning, discuss in detail how to estimate f.

d) Define the following terms in the context of statistical learning. Underfitting, overfitting, Bayes classifier and test error.

e) Provide a sketch of typical (squared) bias, variance, training error, test error and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one. Explain why any of the two curves has the shape displayed in your sketch.

Q.2. Attempt any three of the following. [04 marks each]

a) Explain the differences between the K-nearest neighbors (KNN) classifier and KNN regression methods.

b) Discuss in brief logistic regression and linear discriminant analysis techniques. Provide a brief comparison between these two techniques.

c) Suppose that we take a data set, divide it into equally-sized training and test sets and then try out two different classification procedures. First, we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next, we use 1-nearest neighbors and get an average error rate (average over both test and training data sets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why? Also, discuss the meaning of train and test error rate in this example.

d) What are the key components of a confusion matrix and how are they interpreted? How would you calculate precision, recall and F_1 score using a confusion matrix?

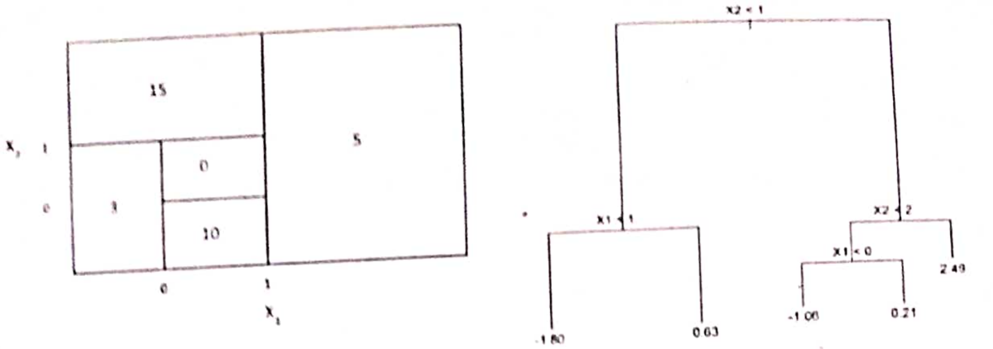
e) Explain the step-by-step procedure for obtaining bootstrap confidence intervals for the population median.

Q.3. Attempt any three of the following. [04 marks each]

a) Discuss the following cross-validation approaches with its advantages and disadvantages.

i) Validation set approach ii) k- Fold cross-validation

b) Sketch the tree corresponding to the partition of the predictor space illustrated in the left-side panel of the following graph. The numbers inside the boxes indicate the mean of Y within each region. Create a diagram similar to the left-hand panel of following figure, using the tree illustrated in the right-hand panel of the same figure. You should divide up the predictor space into the correct regions and indicate the mean for each region.



c) Discuss in detail the two important steps in the process of building a regression tree.

d) Which of the following statements is/are correct? Justify your answer.

i) Boosting algorithm gives more importance to misclassified observations.

ii) Increasing number of trees in bagging can lead to overfitting.

iii) Out-of-bag (OOB) error in Random Forest can be used to estimate training error.

e) Discuss the advantages and disadvantages of decision trees.

Q.4. Attempt any two of the following. [06 marks each]

a) We are given $n = 7$ observations in $p = 2$ dimensions. For each observation, there is an associated class label.

Obs.	1	2	3	4	5	6	7
X_1	3	2	4	1	2	4	4
X_2	4	2	4	4	1	3	1
Y	Red	Red	Red	Red	Blue	Blue	Blue

Sketch the observation and answer the following.

i) Sketch the optimal separating hyperplane. Describe the classification rule for the maximal margin classifier. It should be something along the lines of "Classify to Red if $\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$ and classify to Blue otherwise." Provide the values for β_0, β_1, and β_2. On your sketch, indicate the margin for the maximal margin hyperplane. Indicate the support vectors for the maximal margin classifier.

ii) Argue that a slight movement of the seventh observation would not affect the maximal margin hyperplane.

b) Compare and contrast the construction of the maximal margin classifier and the support vector classifier. Highlight the key differences in their approaches and objectives.

c) Attempt the following.

i) Define support vector machine (SVM) and discuss commonly used SVM kernels.

ii) How can SVMs be applied to multi-class classification tasks? Discuss the techniques used for multi-class SVM classification.

Q.5. Attempt any two of the following. [06 marks each]

a) Attempt the following.

i) What are principal components in PCA? How are they computed from the dataset?

ii) How do you determine the optimal number of principal components to retain in PCA?

b) Discuss in detail hierarchical clustering.

c) Perform the K-means clustering manually, with $K = 2$ on a small example with $n = 6$ observations and $p = 2$ features. The initial cluster assignments for each observation are provided in the table below. The observations are as follows.

Obs.	1	2	3	4	5	6
X_1	1	1	0	5	6	4
X_2	4	3	4	1	2	0
#	1	2	2	1	2	2

#: Initial cluster assignment

Compute the centroid for each cluster. Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation. Repeat these required steps until the assignment of observations to clusters stop changing.

Seat No. 387419

Department of Statistics

School of Mathematical Sciences

Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon

M.Sc. (Statistics) Term End Examination

ST-404 (A): Data Mining

N. B. (1) Do not write anything on the question paper except your seat number.

(2) All questions are compulsory and carries equal marks.

Max. Marks: 60 Time: 3 hours

Q.1. Attempt any four of the following. [03 marks each]

a) Define the following terms in the context of statistical learning. Irreducible error, underfitting and overfitting.

b) Consider the model $Y = f(X) + \varepsilon$. In the context of statistical learning, discuss why to estimate f.

c) What is the difference between 'train error' and 'test error'? Which one of them is better indicator of prediction ability of the model and why?

d) Discuss the trade-off between prediction accuracy and model interpretability.

e) Describe the differences between a parametric and nonparametric statistical learning approach.

f) For each of the following scenarios, state if it is an example of supervised or unsupervised learning. Explain your answers.

i) A recommendation system on a social network that recommends potential friends to a user.

ii) A recommendation system on an online retailer that recommends to users what to buy based on their past purchasing history.

Q.2. Attempt any three of the following. [04 marks each]

a) Discuss the comparison of linear regression over K-nearest neighbors regression.

b) Explain the concept of odds and solve the following.

i) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?

ii) Suppose that an individual has a 16% chance of defaulting on her credit card payment. What are the odds that she will default?

c) Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First, we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next, we use 1-nearest neighbors and get an average error rate (average over both test and training data sets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why? Also, discuss the meaning of train and test error rate in this example.

d) Discuss in detail the linear discriminant analysis with one predictor as a classification technique.

e) Define the following terms.

Confusion matrix, precision, F_1 score, ROC curve.

Q.3. Attempt any three of the following. [04 marks each]

a) Discuss the following resampling methods in the context of statistical learning.

i) k- Fold cross-validation   ii) Bootstrap

b) Provide a detail explanation of the algorithm that is used to build a regression tree.

c) Explain the similarities and differences between the bagging of trees and random forest models. In which way, does the random forest algorithm improve the model based on bagging and how?

d) Explain the following terms in the context of trees.

Gini index and Entropy.

e) A small dataset with eight users, their age and their engagement with our developed App. The engagement is measured in the number of days when they opened the App in one week.

Age	10	20	30	40	50	60	70	80
Engagement	7	5	7	1	2	1	5	4

To construct regression tree, we decided to split the data with respect to Age. Hence, calculate the MSE for each of the different cutoff values of age as 15, 35 and 65 with the objective is to decide the optimal cutoff value of age as a splitting criterion.

Q.4. Attempt any two of the following. [06 marks each]

a) Explain the following terms.

Hyperplane and classification using separating hyperplane.

b) Discuss the comparative differences between the construction of maximal margin classifier and support vector classifier.

c) Attempt the following.

i) What is a support vector machine (SVM)? What are the different types of SVM kernels commonly used?

ii) Can SVM used for multi-class classification? Discuss.

Q.5. Attempt any two of the following. [06 marks each]

a) Attempt the following.

i) What are the practical issues arises in K-means clustering and hierarchical clustering?

ii) Define linkage and its types. Discuss its used in hierarchical clustering.

b) What are principal components? How principal components are constructed.

c) What is the main goal of clustering? Discuss the details of K-means clustering and its algorithm.

Department of Statistics, School of Mathematical Sciences

Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon

Course: M. Sc. (Statistics) ST-404(A): Data Mining

Internal Test -I

Date: 09.03.2024 Time: 04.00 pm – 05.00 pm

1) Attempt any three. (Each carries 2 marks). (6)

a) Explain the difference between underfitting and overfitting in the context of statistical learning. How can these phenomena be addressed?

b) Explain the bias-variance trade-off. Provide a sketch of typical bias and variance curves on a single plot as we go from less flexible statistical learning methods towards more flexible approaches.

c) Define K-nearest neighbors classifier. Discuss the impact of different values of K on 'train error' and 'test error'.

d) Describe two real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

2) Attempt any three. (Each carries 3 marks). (9)

a) Discuss the differences between supervised and unsupervised learning problems, highlighting their respective objectives.

b) Discuss in brief logistic regression and linear discriminant analysis techniques. Provide a brief comparison between these two techniques.

c) Discuss in detail how to assess the performance of different classifier models using a confusion matrix and related measures.

d) Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First, we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next, we use 1-nearest neighbors and get an average error rate (average over both test and training data sets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Provide justification.

Department of Statistics, School of Mathematical Sciences
Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon
Course: M. Sc. (Statistics) ST-404(A): Data Mining
Internal Test -II

Date: 29.04.2024 Time: 04.00 pm – 05.00 pm

1) Attempt any three. (Each carries 2 marks). (6)

a) Define the following.

i) Hyperplane ii) k- fold cross validation

b) Which of the following statements is correct? Justify your answer.

i) Random Forest decreases the variance of the final predictions.

ii) Boosting rarely leads to overfitting because the trees fitted are independent of each other.

c) Define Gini index and Entropy with their importance in the context of classification trees.

d) Discuss the concept of proportion of variance explained (PVE) by principal components. Also, describe the method for determining optimal number of principal components.

2) Attempt the following. (Each carries 3 marks). (9)

a) Suppose we wish to invest a fixed sum of money in two financial assets that yield returns X and Y respectively. We will invest a fraction α of our money in X and will invest the remaining $1-\alpha$ in Y. Prove that $\alpha = \frac{\sigma_Y^2-\sigma_{XY}}{\sigma_X^2+\sigma_Y^2-2\sigma_{XY}}$ does indeed minimize $Var(\alpha X+(1-\alpha)Y)$. Suppose 100 bivariate sample data for (X,Y) is given. Describe the steps required to compute bootstrap estimate of α and its estimate of standard error based on 1000 bootstrap samples.

b) Explain the differences between Maximum Margin Classifier, Support Vector Classifier and Support Vector Machine.

c) Suppose that we have four observations, for which we compute a dissimilarity matrix given by,

$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}$$

The dissimilarity between the first and second observations is 0.3 and the dissimilarity between the second and fourth observations is 0.8.

i) On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observations using complete linkage.

ii) Suppose that we cut the dendrogram obtained in (i) such that two clusters result. Which observations are in each cluster?