INT368:MACHINE LEARNING-I

L:2 T:0 P:2 Credits:3

Course Outcomes: Through this course students should be able to

CO1 :: examine linear regression using SLR and MLR

CO2 :: analyze the different logistic regression techniques using evaluating parameters

CO3 :: develop text classification models using naive bayes

CO4 :: discuss the different type of clustering techniques

CO5 :: compare the k means and hierarchal clustering technique

CO6 :: construct regression and clustering models for real time problems

Unit I

Linear Regression Part 1: SLR : Introduction to Machine Learning, Regression Line, Best Fit Line, Strength of Simple Linear Regression, Assumptions of Simple Linear Regression, Reading and Understanding the Data, Hypothesis Testing in Linear Regression, Building a Linear Model, Residual Analysis and Predictions, Linear Regression using SKLearn

Unit II

Linear Regression Part 2: MLR : Motivation When One Variable Is Not Enough, Moving from SLR to MLR: New Considerations, Multicollinearity, Dealing with Categorical Variables, Model Assessment and Comparison, Feature Selection, Reading and Understanding the Data, Data Preparation, Initial Steps, Building the Model, Residual Analysis and Predictions, Variable Selection using RFE

Unit III

Logistic Regression : The Classification Function, Odds and Log Odds, Binary Variable Distribution, Maximum Likelihood Function (MLE), Understanding Gradient Descent Optimization, Minimizing the Log Loss Function, Gradient Descent Optimization of the log loss, Evaluation Metrics for Classification Problems, Multivariate Logistic Regression - Telecom Churn Example, Data Cleaning and Preparation, Building Your First Model, Feature Elimination using RFE, Confusion Matrix and Accuracy, Manual Feature Elimination, Metrics Beyond Accuracy: Sensitivity & Specificity, Sensitivity & Specificity in Python, ROC Curve, ROC Curve in Python, Finding the Optimal Threshold, Model Evaluation Metrics - Exercise, Precision & Recall, Making Predictions

Unit IV

Naives Bayes : Conditional Probability and Its Intuition, Bayes' Theorem, Naive Bayes -With One Feature, Conditional Independence in Naive Bayes, Deciphering Naive Bayes, Introduction - Naive Bayes for Text Classification, Document Classifier - Pre Processing Steps, Document Classifier - Worked out Example, Laplace Smoothing, Quick Introduction to Bernoulli Naive Bayes, Python Lab - Education Or Cinema ?, Python Lab - SMS Spam Ham Classifier : Multinomial, Python Lab - SMS Spam Ham Classifier : Bernoulli, Comprehension - Naive Bayes for Text Classification

Unit V

Clustering Part 1: K-Means : Understanding Clustering, Segmentation, Types of Clustering and Clustering Methods, Distance Measure, Hopkins Statistics, K-means algorithm, Cost Function, Optimal K, Pros and Cons of K-Means, K-means++, Data Understanding and Data Cleaning, Data Preparation - I, Making the Clusters, Optimal Number of Clusters, Cluster Analysis, Other Behavioural Segmentation Types

Unit VI

Clustering Part 2: Hierarchical Clustering : Hierarchical Clustering Algorithm, Building a Dendrogram, Types of Linkages, Pros and Cons of Hierarchical Clustering, Hierarchical Clustering in Python, Bisecting K-means

List of Practicals / Experiments:

List of practical

- Write a code which works exactly like normalize(data).
- Based on the given cost function, write a code for implementing it.

- You’re given two lists, the first of which contains the name of some people and the second contains their corresponding ‘response’. These lists have been converted to a dataframe.Now, the values that the ‘response’ variable can take are ‘Yes’, ‘No’, and ‘Maybe’. Write a code to map these variables to the values ‘1.0’, ‘0.0’, and ‘0.5’.
- The equation given below represents a model used to predict if a child is likely to choose commerce as their major in high school. The features of the model are x1 = marks in 10th class (use percentage value directly, e.g., for 60% marks, use 60), x2 = a boolean value representing whether some family member is a CA.?p=1/1+e-(0.005x1+0.5x2)?For a particular child scoring 80% marks in class ?10th and their elder brother being a CA, find the probability of the child opting for commerce (Solve th
- Consider the minimisation of the function L(w1,w2)=w21+w22. We have already performed one iteration starting from w0=(5,-4) which moves to w1=(4,-3.2). Continue one more iteration with the same learning rate a=0.1. What is the value of W2? (Write a code to calculate W2)
- A logistic regression model is trained on the following data setx1x2x3y10-0.510550.301 During the gradient descent, the weights calculated in the 20th iteration are w1 = 0.04, w2 = -1, w3 = 0.4. Calculate the weights of the 21st iteration (assume the learning rate of 0.2, and use the batch-gradient algorithm). (Write a code to solve the problem)
- Using the weights: w1=-0.025;w2=-0.957;w2=0.322?, predict the class of a data point with the following features:?x1=40; x2=0.6; x3=1?Assume the threshold to be 0.6. Write the code to find ?s and then based on the threshold classify the data point.
- Which among accuracy, sensitivity, and specificity is the highest for the model below? Actual/PredictedNot ChurnChurnNot Churn8040Churn3050 Write a code to find accuracy, sensitivity, and specificity and then input the given values.
- There is a measure known as F1-score, which essentially combines both precision and recall. It is basically the harmonic mean of precision and recall, and its formula is given by:?F=2× (precision×recall)/(precision+recall?)The F1-score is useful when you want to look at the performance of precision and recall together.Write a code to calculate the F1-score for the model below.Actual/PredictedNot ChurnChurnNot Churn400100Churn50150
- Bag A contains 3 Red and 4 Green balls, and bag B contains 4 Red and 6 Green balls. One bag is selected randomly, and a ball is drawn from it. If the ball drawn is found Green, find the probability that the bag was chosen was A. (Write a code for Bayes Theorem and input the given values)
- Consider points, A(7,50) and B(23,34). Write a code to compute the Euclidean and Manhattan distance between the two. [Round off the answer to two decimal places]
- Write a code defining Hopkins statistic function which checks whether the dataset is suitable for clustering or not when a dataframe is passed through it.
- Given a dataset here about the batting figures of batsmen in ODI matches.Choose the number of clusters as four. Does ?SR Tendulkar fall in the same cluster as Virat Kohli?
- Given a dataset here about the batting figures of batsmen in ODI matches.Based on the clustering, given that the clusters formed are (high SR, high Ave) - A, (low SR, low Ave) - B, (High SR, Low Ave) - C, (Low SR, High Ave) - D. Who all fall in group A?

Text Books:

1. MACHINE LEARNING USING PYTHON by MANARANJAN PRADHAN, U DINESH KUMAR, WILEY

References:

1. PYTHON MACHINE LEARNING - SECOND EDITION by SEBASTIAN RASCHKA, PACKT PUBLISHING