IIMT2641 Introduction to Business Analytics

Assignment 5

Due: 3 May 2023, 00:00

For this assignment, you ONLY need to answer either Q1 or Q2

Q1 In this problem, we will practice building CART models with a continuous outcome, using the dataset StateData.csv which has data from 1970s on all fifty US states. A description of the variables in the dataset is given in Table 1.

Variable	Description
Population	Population estimate of the state in 1975.
Income	Per capita income in the state in 1974.
Illiteracy	Illiteracy rates in 1970, as a percentage of the state's population.
LifeExp	The life expectancy in years of residents of the state in 1970.
Murder	The murder and non-negligent manslaughter rate per 100,000 population in 1976.
HighSchoolGrad	The high-school graduation rate in the state in 1970.
Frost	The mean number of days with minimum temperature below freezing from 1931 to 1960 in the capital or a large city of the state.
Area	The land area (in square miles) of the state.
Longitude	The longitude of the center of the state.
Latitude	The latitude of the center of the state.
Region	The region (Northeast, South, North Central, or West) that the state belongs to.

Table 1: Variables in the dataset StateData.csv.

(a) Let us start by building a linear regression model. Randomly split the dataset into a training set (70%) and a test set (30%).

(i) First, build a linear regression model to predict LifeExp using the following several variables as the independent variables: Population, Murder, Frost, Income, Illiteracy, Area, and HighSchoolGrad. Use the training dataset to build the model. What is the R^2 of the model on the test set?

(ii) Now, build a linear regression model to predict LifeExp the following four variables as the independent variables: Population, Murder, Frost, and HighSchoolGrad. Again, use the training dataset to build the model. What is the R^2 of the model on the test set?

(iii) Compare these two models. What are we achieving by removing independent variables? What is the equivalent procedure in a CART model?

(b) Now, build a CART model to predict LifeExP using the following seven variables as the independent variables: Population, Murder, Frost, Income, Illiteracy, Area, and HighSchoolGrad. Set the parameter minbucket to be 5. Make sure that you are building a regression tree, and not a classification tree, by setting the argument method to “anova” instead of “class”.

(i) Plot the trees. Which of the independent variables appear in the tree? Do you find the linear regression model or the CART model easier to interpret?
(ii) Compute the predicted life expectancies for the test dataset using the CART model, and calculate the R^2 of the predictions.

(c) Now, build a random forest model to predict LifeExP using the same seven variables as the independent variables. Set the parameter nodesize to 5. Compute the predicted life expectancies for the test dataset using the random forest model, and calculate the R2 of the predictions.

(d) Which of the four models you built do you think is the best model, if out-of-sample accuracy is the most important. How about if interpretability is the most important?

Q2 Document clustering, or text clustering, is a very popular application of clustering algorithms. A web search engine, like Google, often returns thousands of results for a simple query. For example, if you type the search term “jaguar” into Google, over 400 million results are returned. This makes it very difficult to browse or find relevant information, especially if the search term has multiple meanings, like this one. If we search for “jaguar”, we might be looking for information about the animal, the car, or the Jacksonville Jaguars football team.

Clustering methods can be used to automatically group search results into categories, making it easier to find relevant results. The two most common clustering algorithms used for document clustering are Hierarchical and K-means.

In this problem, we will be clustering articles published on Daily Kos, an American political blog that publishes news and opinion articles written from a progressive point of view. The file DailyKos.csv contains data on 3,430 news articles or blogs that have been posted on Daily Kos. These articles were posted in 2004, leading up to the United States Presidential Election. The leading candidates were incumbent President George W. Bush (Republican candidate) and Senator John Kerry (Democratic candidate). Foreign policy was a dominant topic of the election, specifically, the 2003 invasion of Iraq.

There are 1,545 variables in this dataset. Each of the variables in the dataset is a word that has appeared in at least 50 different articles (1,545 words in total). For each document, or observation, the variable values are the number of times that word appeared in the document. (This approach is called bag of words in text analytics.)

(a) Start by building a Hierarchical Clustering model to cluster documents using all of the variables in the dataset. Indicate which distance metrics you used for distances between the observations and distances between the clusters.

(i) Building a hierarchical clustering model will probably take a significant amount of time on this dataset. Why?

(ii) Plot the dendrogram of your hierarchical clustering model. Using the dendrogram and thinking about this particular application, which number of clusters would you recommend? Keep in mind that document clustering would most likely be used by Daily Kos to show readers categories to choose from when trying to decide which articles to read.

(iii) Assign each observation to a cluster, using the number of clusters you recommended in the previous subproblem. How many observations are in each cluster?

(iv) Split your dataset into a dataset for each cluster, using your cluster assignments. Then, find the six most frequent words in each cluster. If you name the first HC1 in R, this can be done with the command: tail(sort(colMeans(HC1))). Describe each cluster. Is there a cluster that is mostly about the Iraq war? Is there a cluster that is mostly about the democratic party? It might be helpful to know that in 2004, Howard Dean was one of the candidates for the Democratic nomination for the President of the United States, John Kerry was the candidate who won the democratic nomination, and John Edwards was the running mate of John Kerry (the Democratic Vice President nominee).

(b) Now cluster the documents using K-means clustering. Choose the same number of clusters that you recommended for Hierarchical clustering.

(i) How many observations are in each cluster? Is your answer the same as it was with Hierarchical clustering? Why or why not?

(ii) Just like you did for Hierarchical clustering, split your dataset into a dataset for each K-means cluster, and analyze the most frequent words in each cluster. Are the clusters similar to the Hierarchical clusters? Can you find a similar Hierarchical cluster for each K-means cluster? Keep in mind that the order of the clusters (which cluster is labeled as 1, which cluster is labeled as 2, etc.) is meaningless — for example, Hierarchical cluster 3 might be very similar to K-means cluster 1.