MATH 3190 Homework 7

Devin Warner

Due 4/15/2022

These answers work in tandem with my KMeans and College Basketball RShiny apps which can be found in my *devinwkmeans* and *devinwcbb* packages respectivly. To view the Shiny Apps downland the packages and run the functions *run_kmeans_shiny()* for the KMeans Shiny and *run_cbb_shiny()* for the Basketball Shiny. Please reach out if you have trouble running the apps.

1. (50 points) Here we will modify your K-means package and Shiny App to include K nearest neighbor, principle components, and umap functions. Please do the following:

   (a) **Kmeans:** Change your K-means algorithm to allow for the users choice of 1 or more input variables (currently using 2), still allowing for the user to choose K and display the two dimensions of choice (as is the case currently). Change K-means plotting function (and thus the app) to display iris species in different colors and classification group using shapes (add a legend to this plot). Apply cross-validation to identify the the optimal value for K for the iris dataset.

   On the K-Means tab, users now have the option to select radio buttons of the variables they would like to use while clustering. The plot still only shows two variables at a time, but the results change depending on what variables are being clustered. A warning is given if a variable is being plot that is not also being clustered by.

   The user can still change the number of clusters they would like. We determined in class that we couldn't run cross validation on k-means, and so after some data exploration I have decided (and its pretty intuitive) that the optimal value of K is 3 for this dataset.

   (b) **K nearest neighbors:** Write a function that plots the classification results of a K nearest neighbors algorithm for the users choice of 1 or more input variables and the user's choice of K. Plot points in two dimensions (user's choice) and display iris species in different colors and classification group using shapes (add a legend). Add this to the K-means Shiny App. Apply cross-validation to identify the the optimal value for K for the iris dataset.

   The kNN tab on the Shiny app includes the changes requested in this question. In the background, the `iris` dataset is split into training and testing data. What is plotted is the predicted `Species` of the testing observations using (user's choice) k-nearest neighbors, as well as the actual `Species` of the of the testing observation.

   (c) **Dimension reduction:** write a function that applies dimension reduction methods (PCA, UMAP) to a dataset and plots a user's choice of reduced components in two dimensions (UMAP only provides two), and color the points based on iris species (add a legend).

   The Dimension Reduction tab on the Shiny app includes both Principle Component and UMAP methods and includes the functionality requested in the question.

(d) Which methods would you prefer for classification or analysis for the iris dataset?

Personaly I would use PCA and take the first two components, and then a kNN for classification. I like how PCA groups each of the observations into their respective `Species`, and the kNN appears to be very accurate for prediction. I think this method will give us our best results.

2. (50 points) For your basketball dataset, `mutate` or `summarize` a new dataset that contains the following for each team: average points scored (total, home, away), average points allowed (total, home, away), score difference (total, home, away), winning percentage (total, home, away) new columns for each team (may need to use a log or logistic transformation), conference (get help from Akhil), whether or not they participated in the tournament, and any other relevant statistic or summary measure may think of (if you come up with something good, share with the class!). Do the following:

   (a) Fit a LASSO model to predict factors that predict final winning percentage (might have to use a log or logistic transformation on the percentage). Exclude home and away winning percentage. Identify a "best" value for λ and interpret your model.

   (b) Use PCA and UMAP to provide a two-dimensional map for all of these variables except conference and tournament participation. Try to interpret the PCA rotations. In the plot, do you see any patterns (e.g. conference? tournament appearance?). Do these reductions work better than the individual vairable alone? Add a dimension reduction feature to your basketball Shiny App.