MATH 3190 Homework 7

Regularization, Cross-validation, Dimension reduction

Due 4/15/2022

In this homework add K-nearest neighbor and dimension reduction features to your Shiny Apps. Please do the following and upload changes to your packages on GitHub:

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.

   (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.

   (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).

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

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.