Portfolio Component: Searching for Similarity

Group Project: Up to 4 people

Sign up your group here:
https://docs.google.com/document/d/1msVrb_U3p27oOngnPL4xF4vZGjP87HQILgJ6hmJEq6s/edit?usp=sharing

Objectives:

- Gain experience with machine learning using similarity models kNN and Decision Trees
- Gain experience with clustering methods kMeans, hierarchical clustering
- Gain experience with dimensionality reduction techniques LDA and PCA

Turn in:

- Print your Rmd notebooks to pdf, upload to your portfolio, and create links on your index page
- Upload your Rmd print-to-pdfs to eLearning

Instructions:

1. Notebook 1 Regression. Find a medium-size data set of at least 10K rows, suitable for regression. Provide a link to the data in markdown.
    a. Divide the data into train/test
    b. Explore the training data statistically and graphically
    c. Perform linear regression, kNN regression and Decision tree regression, then compare the results.
    d. Provide some analysis on why the results were most likely achieved given how the algorithms work.
2. Notebook 2 Classification. Repeat the steps above for a classification data set using Logistic Regression, kNN, and Decision Trees.
3. Notebook 3 Clustering. Using either data set above, or another data set, perform kMeans clustering and Hierarchical clustering. You may need to subset the data for Hierarchical clustering. Research model-based clustering at the link below and implement it in R. Write a paragraph comparing the results of each algorithm and what insights they gave you to this data. (https://www.statmethods.net/advstats/cluster.html)
4. Notebook 4 Dimensionality reduction. Select one of the data sets above to perform both PCA and LDA dimensionality reduction. Try classification or regression on the reduced data and compare the results. How much accuracy was lost with the reduced data?
5. Upload the print-to-pdfs for all notebooks, and create links to them on your index page.
6. Narrative document. Write a 1-2 page narrative discussing:
    a. how kNN and decision trees work for classification and regression
    b. how the 3 clustering methods of step 3 work
    c. how PCA and LDA work, and why they might be useful techiques for machine learning

Caution: All course work is run through plagiarism detection software comparing students' work as well as work from previous semesters and other sources.

Grading Rubric:

Element	Points
Step 1 Regression notebook	50
Step 2 Classification notebook	50
Step 3 Clustering notebook	50
Step 4 Dimensionality reduction notebook	50
Step 5 Narrative	40
Step 6 Create links to the document and code on the index page	10
Total	250

Grading Rubric:

- 90% and above for exceptional work
- 80-89% for good work
- 70-79% for average work
- below 70% for low quality work

Caution: All course work is run through plagiarism detection software comparing students' work as well as work from previous semesters and other sources.