Problem Set 4 - Data Reduction & Feature Selection

NB

1) Which programming languages to use?

We recommend to use Python for this task, but if you find working library alternatives for the algorithms we use in this assignment in R, you are free to work with that as well.

2) What libraries/packages to use?

You are free to choose any appropriate libraries (good choice would be pandas, numpy, scikit-learn).

3) How to summarize my work?

The best way is to create an Jupyter/R notebook with code and explanations for each strategy. In case you are not familiar with these tools, you can create a Python/R scripts and write explanations as comments. However, we strongly recommend you to use Jupyter/R notebooks, as those are #1 tools in applied data analysis nowadays.

Please do not include large datasets in the archive with your notebook(s)!

4) Useful links

1. Q/A on Dimentionality Reduction Techniques
2. The Ultimate Guide to 12 Dimensionality Reduction Techniques
3. Reducing Dimentionality
4. PCA in details
5. Why, How and When to apply Feature Selection

Tasks

1) [5pt] Dimentionality reduction

1.1. Download the Gisette Data Set from UCI ML repository **(dataset)**.

1.2. Read the description of the broblem and try to solve it using Logistic Regression (without dimentionality reduction). Save it's performance (time to train) and accuracy on `test` dataset. Is this approach efficient? Would it work better with some other ML model?

1.3. Apply PCA algorithm to reduce dimentionality of your data. Use $n = (5, 10, 20, 50, 100)$ as a number of components.

1.3.1. Visualize how do the models with given ns differ with regard to the amount of variance they explain.

1.3.2. Visualize decomposed components for $n = 5$. What can you tell about the amount of variance each additional component explains?

1.3.3. For each dataset obtained from PCA with given n train a Logistic Regression model. Save their performance and accuracy on `test` dataset.

1.4. Apply Factor Analysis technique to reduce dimentionality of your dataset. Use $n = (3, 5, 10, 20, 50)$ as a number of factors.

1.4.1. Visualize factors you obtained for $n = 3$.

1.4.2. For each datase obtained train a Logisitc Regression model. Save their performance and accuracy on `test` dataset.

1.5. Compare performance/accuracy of original model with models trained on datasets obtained using PCA and FA. What are the pros/cons of each approach? What are the main use-cases for those algorithms? (5-6 sentences)

1.6.

2) [5pt] Feature selection

2.1. Download the Spambase Data Set from UCI ML repository **(dataset)**.

2.2. Train a regular Logistic Regression model using original attributes. Save its accuracy on `train` and `test` sets.

2.3. Apply Forward Stepwise Selection technique to find the subset of attributes which minimizes an estimate of the expected predition error. Visualize the process of this selection (subset size on x-axis).

2.3.1. Train a Logistion Regression model using the features found with FSS. Save its accuracy on `train` and `test` datasets.

2.4. Apply Backward Stepwise Selection technique to find the subset of attributes which minimizes an estimate of the expected predition error. Visualize the process of this selection (subset size on x-axis).

2.4.1. Train a Logistion Regression model using the features found with BSS. Save its accuracy on `train` and `test` datasets.

2.5. Compare two approaches (FSS and BSS). Did they found the same subset? If not, explain why it could have happend?

2.6. Use Decision Trees to find important features. Visualize relative feature importance.

2.6.1. Train a Logistic Regression model using the features found with DT. Save its accuracy on `train` and `test` datasets.

2.7. Apply the following "manual" techniques to detect redundant features: missing value ratio, low variance and high correlation filters.

2.7.1. Remove features you found (in 2.7.) from the dataset and train a Logistic Regression model using the rest of the data. Save its accuracy on `train` and `test` datasets.

2.8. Compare the results of different methods of feature selection. Write pros/cons of each of them.