Data Challenge 6

Data Science I 2020

Due 12/16/2020

For this assignment you will be applying what we have learned about unsupervised learning to the AirBnB data. Specifically we will be exploring the structure of the AirBnB listings using the detailed listings file (listings.csv.gz) with a clustering algorithm of your choice and principal component analysis (PCA). You will first generate a set of features for unsupervised learning from the detailed listings file (listings.csv.gz). Then you will apply either k-means clustering or hierarchical clustering as well as dimension reduction with PCA. Lastly you will explore and interpret the clusters / dimension reduction and discuss any insights gleaned from the unsupervised learning exercises. The final product from this Data Challenge will be a report detailing your findings.

- **Abstract** Summarize the entire analysis, with emphasis on the key results.
- **Introduction** Introduce the data and provide the necessary background information for the analysis you will perform.
- **Methods** Describe the features (X's) that you used in your model. Which features from the data did you use? Why? Briefly describe the unsupervised clustering methods you will be using and the analysis you will be performing to explore and interpret the results.
- **Results** Present the results from the clustering and PCA. Show at least 3 figures with your results. Please note that an excessive number of figures will make the report difficult to read.
- **Discussion** Interpret the clusters and results from the dimension reduction. What insights are gleaned from the analysis? What are the limitations of the analysis? What are the next steps?
- **Conclusion** A high level summary of your findings for the project.

Just as in Data Challenge 5, we will access the AirBnB data at http://insideairbnb.com/get-the-data.html. You can pick any city of your choosing for this analysis --- please only pick one city.

You will submit this assignment as an html file on Canvas. Please write a formal report where you can suppress the code (Hint: Do some googling to find ways to suppress code in a Jupyter Notebook! There are a lot of code snippets that allow you to press a button to show / hide code). Submit your .ipynb file to a GitHub repository and include a link to the repository at the top of the html that you submit on Canvas. Make the repository private and invite the instructor and both of the TAs to your repository.

The rubric for grading is shown below. The assignment is worth **100 pts**.

- **Code Style (10 pts)** Is code organized well and commented?
- **Submission (10 pts)** Was the data challenge submitted as an html document on Canvas? Did the html document contain code and written explanations of results? Does the html document look aesthetically pleasing (ability to hide / show code)? Did the homework contain a link for a GitHub repository? Did the repository contain the code for the assignment?
- **Feature Generation and Justification (10 pts)** Were features thoughtfully and properly engineered (for example dummy variables for categorical features)? Was justification for the features used provided and compelling?
- **Clustering (15 pts)** Was clustering done properly? How was the number of clusters chosen? Were variables scaled when appropriate?
- **Principal Component Analysis (15 pts)** Was PCA done properly? How were the number of PCs chosen? Were variables scaled when appropriate?
- **Synthesis of Unsupervised Clustering Results (20 pts)** How were the clusters and dimension reduction results synthesized and explored? Were meaningful insights gleaned from the analysis?
- **Written Report (20 pts)** Are the required sections included in the report? How well are the methods and results communicated?