Homework 7: Unsupervised Learning

Overview

Due Sunday by 11:59 pm.

Fork the `problem-set-7` repository

k-Means Clustering "By Hand"

You fielded an experiment and collected observations for 10 respondents across two features. The data are:

```
input_1 = c(5,8,7,8,3,4,2,3,4,5)
```

```
input_2 = c(8,6,5,4,3,2,2,8,9,8)
```

After inspecting your data, you suspect 3 clusters likely characterize these data, but you'd like to check your intuition. Perform k-means clustering "by hand" on these data, initializing at `k = 3`. Be sure to set the seed for reproducibility. Specifically:

1. (5 points) Imitate the k-means random initialization part of the algorithm by assigning each observation to a cluster at random.
2. (5 points) Compute the cluster centroid and update cluster assignments for each observation iteratively based on spatial similarity.
3. (5 points) Present a visual description of the final, converged (stopped) cluster assignments.
4. (5 points) Now, repeat the process, but this time initialize at `k = 2` and present a final cluster assignment visually next to the previous search at `k = 3`.
5. (10 points) Did your initial hunch of 3 clusters pan out, or would other values of `k`, like 2, fit these data better? Why or why not?

Application

`wiki.csv` contains a data set of survey responses from university faculty members related to their perceptions and practices of using Wikipedia as a teaching resource. Documentation for this dataset can be found at the UCI machine learning repository. The dataset has been pre-processed for you as follows:

- Include only employees of UOC and remove `OTHER*`, `UNIVERSITY` variables
- Impute missing values
- Convert `domain` and `uoc_position` to dummy variables

Dimension reduction

6. (15 points) Perform PCA on the dataset and plot the observations on the first and second principal components. Describe your results, e.g.,
   - What variables appear strongly correlated on the first principal component?
   - What about the second principal component?
7. (5 points) Calculate the proportion of variance explained (PVE) *and* the cumulative PVE for all the principal components. **Approximately how much of the variance is explained by the first two principal components?**
8. (10 points) Perform *t*-SNE on the dataset *and* plot the observations on the first and second dimensions. Describe your results.

Clustering

9. (15 points) Perform k-means clustering with $k = 2, 3, 4$. Be sure to scale each feature (i.e.,mean zero and standard deviation one). *Plot* the observations on the first and second principal components from PCA and *color-code* each observation based on their cluster membership. *Discuss* your results.

10. (10 points) Use the elbow method, average silhouette, and/or gap statistic to identify the optimal number of clusters based on k-means clustering with scaled features.

11. (15 points) Visualize the results of the optimal $\hat{k}$-means clustering model. **First** use the first and second principal components from PCA, and color-code each observation based on their cluster membership. **Next** use the first and second dimensions from t-SNE, and color-code each observation based on their cluster membership. **Describe your results. How do your interpretations differ between PCA and t-SNE?**