Aaron Hamilton
Conrado Arroyo
Leona Liu
MATH-514

Project 2 EXTRA Credit:

Written Questions (Complete these in the Assignment Short Answer, located in your assignment page)
(Short answer, 2-3 sentences each)

- **How many principal components do you think are necessary for the cat faces dataset? Which measurement of the number of features, using the eigenvectors or cross-validation, do you think would be likely to give better results for answering this question? Connect your answer to what you know about supervised learning.**
    - When looking at the plot of explained variance over the range of different numbers of principal components, there is a dropoff approximately after the fourth component is introduced, meaning that four principal components are necessary for the cat faces dataset. Outside of this “elbow method”, a way to ensure that the number of principal components maintained will be the best option is to do cross validation. Because there is potential for the model to only be adjusted to the current set of data it was given, doing cross validation is important to ensure that the model is not simply well adjusted to what it is currently handling but is able to work well on new data sets as well. Also, because the approach we are trying to take is supervised learning, it could be better to use cross-validation as opposed to eigenvectors—which are more useful in unsupervised ML techniques.

- **Compare the results you obtained from the cat faces to the human faces in the original demo. What could be done to the cat faces dataset to get results that look more like the results using human faces.**
    - It looks like using Cluster centers or FA on the cat faces could provide some level of accuracy better than PCA. It also looks like the cat faces somewhat blend into the background of the picture, so Independent Components might be a good choice for maximizing independence between subcomponents. Centering the cat faces might help a bit, though they look relatively centered already.

- **In the review data, you plotted the data by label along with the principal components. You also looked at the top topics extracted from the text of the reviews. Which of these two do you think is a more meaningful representation of the data?**

    - It seems like the top topics extracted from reviews would be a better representation of the data in this specific case, mostly because of the unstructured appearance of the first two PCs. Even though the PCA plot with labeled features does look like it provides some level of accuracy for identification, with higher-numbered labels having lower PC scores than lower-numbered ones, overall it seems like the ability to extract meaning from the review data could be more easily interpretable and useful.

- **Did embedding the data using t-Distributed Stochastic Neighbor Embedding (t-SNE) give any additional information about the review data? Do you see more structure in the data with this nonlinear model than the one using PCA?**
    - Using t-SNE produced much more clear results about the data than the PCA; whereas the two components generated by PCA were seemingly unstructured, when we applied t-SNE we can see the resulting scatterplots have a much more linear structure. Overall the use of t-SNE produced better dimensionality reduction results than PCA. Although it is relatively intense with computations, and it is complicated to tune all of the hyperparameters, the fact that it was able to handle outliers well and preserve local structure for the data points allowed it to produce a better result than PCA.