ML Foundations
HW 3: PCA¡ K-Means, and Data Validation

Item(s) to turn in: jupyter notebook, classifiers.py, feature_reduction.py and pdfs of each file (notebook, and python files). Note: You should zip all code/notebook files and turn in that zip file. Pdfs should be submitted separately and not zipped.

Approved Libraries: numpy, pandas, matplotlib,seaborn, sklearn for validation/ support vector machines

---

1. Your first goal is to fill in the KMeans classifier and test it. (10 pts)

   For testing:

   - Create two numpy arrays (resting, stressed) with 1,000 samples each: $\mu_{rest} = [60, 10], \Sigma_{rest} = [[20, 100], [100, 20]]$. $\mu_{stress} = [100, 80], \Sigma_{stress} = [[50, 20], [20, 50]]$.
   - Run your k-means clustering algorithm on this data.
   - Plot the resulting centroids of each clusters and the data (plt.scatter is a useful function).
   - Test your algorithm against Sklearn's KMeans algorithm

   Short Answer:

   1. Is k-means guaranteed to provide you a unique solution?

2. Your next goal is to fill in the PrincipleComponentAnalysis class. (10 pts)

   For testing:

   - Read in the iris dataset (pd.read_csv("https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data", header=None)
   - Drop all nans and only look at the columns: "sepal_length","sepal_width", 'petal_length','petal_width','species'
   - Run your PCA on this code with a threshold of 0.95
   - Print out your resulting projection matrix
   - Test your algorithm against Sklearn's PCA

   Short-Answer

1. Why is it important to standardize your data before PCA?

3. Now we are going to put things together using the pokemon_dataset.csv. Your goal is to classify a pokemon as a legendary or non-legendary using a Support Vector Machine from Sklearn. You are not allowed to use the pokemon's name or name2 as features.

   You will be graded on the following:

   1. Data Exploration (3 pts).
   2. Preprocessing (3 pts).
   3. Classifier Validation and Hyperparameter tuning (4 pts).

   Notice that you are not graded on overall performance. You also must include a markdown cell discussing your entire approach and design decisions you made (e.g., which kernel function and why).