Dr. Tony Diana
DATA 602 Introduction to Machine Learning
Homework I Week 8

Unsupervised Models

Use the wine data to conduct Principal Component Analysis. The URL is https://archive.ics.uci.edu/ml/'machine-learning-databases/wine/wine.data

- Load the dataset. What are the features?
- Create a DataFrame of given wine dataset. Standardize the features.
- Split the dataset into training and test sets (30%).
- Use the linalg.eig function from NumPy to obtain the eigenpairs of the Wine covariance matrix. Print the eigenvalues.
- Compute and plot the individual explained variance and the cumulative explained variance (range from 1 to 14). Using the NumPy cumsum function, calculate the cumulative sum of explained variances, which you will then plot via Matplotlib's step function. Explain the outcomes.
- Sort the eigenpairs by decreasing order of the eigenvalues. Make a list of (eigenvalue, eigenvector) tuples. Sort the (eigenvalue, eigenvector) tuples from high to low. Print the projection matrix W.
- Using the projection matrix, transform a sample x onto the PCA subspace (the principal components 1 and 2). Transform the entire 124 x 13-dimensional training dataset onto the two principal components by calculating the matrix dot product.
- Visualize the transformed Wine training set into a two-dimensional scatterplot.
- Method number 2. Import ListedColormap. Set up marker generator and color map. ---
- Plot the decision surface and plot the class samples.
- Import LogisticRegression and PCA. Visualize the decision regions (hint: X_train_pca, y_train, classifier=lr).