Applied Machine Learning

Mini Project – K-means Clustering and Principal Component Analysis(PCA)

Due Date: Refer to NYU Class

Part I. K-means clustering

Introduction

K-means is one of the widely used unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result.

Dataset

In this assignment, you will get familiar with generating dataset by yourself using the third party library.

Requirements

1. Use sklearn library to generate the synthetic data for k-means clustering.
    a. We set the total number of instances to be 300
    b. The number of centers is 4 with the standard deviation 0.6

2. Plot the generated data with labels by using matplotlib.

3. Implement the K-means function return the labels and centers.

4. Fit the model on the dataset and plot the figure with default seed

5. Fit the model on the dataset and plot the figure with seed=2

6. Compare the results from 4 and 5. Is there any differences? If yes, why?

7. Implement the K-means++ function return the labels and centers.

8. Fit the model on the dataset and plot the figure with default seed

9. Fit the model on the dataset and plot the figure with seed=2

10. Compare the results from 8 and 9. Is there any differences? If yes,

why? 11. Compare the results from 4,5,8 and 9. State your observations.

Part II. Principal Component Analysis (PCA)

Introduction

Principal Components Analysis (PCA) is a dimensionality reduction algorithm that can be used to significantly speed up your unsupervised feature learning algorithm. For example, when you train your model on a dataset such as images, some of our data points may be meaningless in explaining our desired target variable. Therefore, we refer to drop them from our training.

Task1: Users to Movies

In this task, you will work with a Users-to-Movies example (as shown below) from http://web.stanford.edu/class/cs246/slides/06-dim_red.pdf. Each row contains the scores provided by a user while each column has the scores given by different users on the same movie. The first 4 users prefer Science Fiction and the others prefer Romance. You will need to implement the classic PCA algorithm and calculate the features after PCA.

$$\begin{array}{c|ccccc|}
 & \text{Matrix} & \text{Alien} & \text{Serenity} & \text{Casablanca} & \text{Amelie} \\
\hline
\uparrow & 1 & 1 & 1 & 0 & 0 \\
 & 3 & 3 & 3 & 0 & 0 \\
\text{SciFi} & 4 & 4 & 4 & 0 & 0 \\
\downarrow & 5 & 5 & 5 & 0 & 0 \\
\uparrow & 0 & 2 & 0 & 4 & 4 \\
\text{Romance} & 0 & 0 & 0 & 5 & 5 \\
\downarrow & 0 & 1 & 0 & 2 & 2 \\
\hline
\end{array}$$

Requirements

1. Understand the data by plotting in 3D space

2. Preprocessing step: centering the dataset

3. Implement PCA using Singular Value Decomposition (SVD) – Solution1

4. Plot the eigenvalues (note: the singular values are the square-root of the eigenvalues) and select the right number *K* of principal components. What number do you choose for *K* ? Explain your choice.

5. Calculate the compressed data with the *K* you choose from step 4

6. Implement PCA by directly computing the eigenvectors (V) and eigenvalues (D) from covariance matrix – Solution2

7. Print the V and D from step 6

Task2: Human Faces

In this assignment, you will need **The Labeled Faces in the Wild** dataset which is designed for the face recognition task. The dataset containing images of faces. Each image is a 62x47 pixel array. The images are read into a matrix. The rows of the matrix are the images (examples). The features (columns) are the pixels. Each example is represented by a vector of real numbers of length 2914, listing the pixels from left to right, row by row, from top to bottom.

https://scikit-learn.org/stable/datasets/index.html#labeled-faces-in-the-wild-dataset

Requirements

1. Load dataset and display the fourth face in the dataset

2. Compute the mean of all the examples in the dataset fea. (That is, compute an image such that each pixel i of the image is the mean of pixel i in all the images in fea.)

3. Display the mean image calculated in step 2

4. Do dimensionality reduction with either of the pca algorithms you implemented in Task

5. Compute the 5 top principal components of the data matrix fea, and print them

6. What are the values of the associated 5 attributes of the fourth image in the dataset?

7. Project the fourth face in the dataset onto the first 5 principal components.

8. Project the fourth face with first 5 principal components back into the original image space, and then display it.

9. Repeat step 5-8 with the first 50 principle components (instead of 5).

Deliverables

A .ipynb file containing the following:

1. Source code
2. Detailed description of the project if needed

Before submitting your project, please make sure to test your program on the given dataset.

Notes

You may discuss the general concepts in this project with other students, but you must implement the program on your own. ***No sharing of code or report is allowed.*** *Violation of this policy can result in a grade penalty.*

Late submission is acceptable with the following penalty policy:
10 points deduction for every day after the deadline