HW4 SPR Fall 2022

Deadline 27 Azar

In order to do this homework, you have to go through feature selection and feature extractions like principal component analysis, and Fisher Linear Discriminant Analysis theories and concepts.

Feature Selection

Dataset: "vote, hepatitis, kr-vs-kp, anneal, diabetes"

- In this part, you have 5 datasets and you should implement 3 methods of feature selection on them. Then you have to classify the original datasets and new datasets (datasets after implementing the feature selection methods) with logistic regression.
- Use the feature selection methods below to choose the K best features of each dataset (You can use the python libraries for the feature selection methods):

  - Chi-Square (k= 5, k=10)
  - RFE (k= 5, k=10)
  - Univariate (k= 5, k=10)

  - Report the K best features for each dataset.
  - Consider the first 80% of each dataset data for train and 20% for test (You can use libraries for splitting your data).
  - Classify all the datasets with logistic regression and get the accuracies.(You can use python libraries for this part)
  - Report the accuracies of each dataset.
  - Compere the accuracies. Why the accuracies of new datas are different from original ones? Explain completely.
  - How these methods can select the best K features? Explain Each method separately.
  - Is feature selection a good method? When should we use it? Explain completely.
  - Which feature selection method worked better on the datasets Why?

PCA

Dataset: "olivetti_faces" in sklearn library

- In this part, you will compute a PCA from a set of images of faces. The database provides facial images of different people. Attention that eigenfaces are sets of eigenvectors that can be used to work with face recognition applications. Each eigenface, as we will see in a bit, appears as an array vector of pixel intensities. We can use PCA to determine which eigenfaces contribute the largest variance in our data and eliminate those that do not contribute much. This process lets us determine how many dimensions are necessary to recognize a face as 'familiar.'
- The Olivetti dataset consists of 400 gray images with a size of 64*64. There are 40 different people and for each person, we have 10 images.
- Use the library **sklearn** to get the Olitive dataset. Then use the method "fetch_olivetti_faces()" to work with the images.

- Visualize the dataset.
- Preprocess and zero mean the dataset.
- Implement the PCA function, then apply it to the dataset.
- Visualize the reduced dataset using 2D and 3D plots. (Using the first two principal components and the first three).
- Reconstruct the original data using K principle components (Show reconstructed images of each individual for K=1,20,50,150).
- Plot the MSE between the original and reconstructed images in terms of the number of eigenvectors.
- Visualize some of the first principal components.
- How many principal components are enough so that you have acceptable reconstruction? How do you select them?
- Plot the cumulative variance retained by the first k components, for $k \in 1 : 300$. How much variance is retained by the first component? By the first five components? How many components are needed to keep 75% of the variance? To keep 95% of the variance?

Fisher LDA

Dataset: "olivetti_faces" in sklearn library

- In this part, you will apply a Fisher LDA from a set of images of faces.
- The Olivetti dataset consists of 400 gray images with a size of 64*64. There are 40 different people and for each person, we have 10 images.
- Use the library **sklearn** to get the Olitive dataset. Then use the method "fetch_olivetti_faces()" to work with the images.

- Implement and apply the Fisher LDA function for a multi-class problem.
- What is the problem with applying Fisher LDA to the dataset?
- Reconstruct the original data by using K basis vectors obtained from LDA. (Show reconstructed images of one person for k=1, 40, 60).
- What would happen if we had a large number of outliers in the dataset?
- Plot the MSE between the original and reconstructed images in terms of the number of eigenvectors.

✓ Pay extra attention to the due date. It will not extend.

✓ Be advised that submissions after the deadline would not grade.

✓ Prepare your full report in PDF format and include the figures and results.

✓ Do not use sklearn or any similar library and write your own code.

✓ Use only the python programming languages.

✓ Submit your assignment using a zipped file with the name of:

"FirstName_LastName_ TeammateFirstName _Teammate LastName.zip"

✓ Email your zip file to 'melikaaa9896@gmail.com'.

✓ Using other students' codes or the codes available on the internet will lead to zero grades.