[Lab/Homework] K-means and PCA

Jae Yun JUN KIM*

Lab due: Today
Homework due: Before the next lab session, submit your .ipynb files on campus.ece.fr.

- Group 1: Before Wednesday March 24th, 2021, 15h30
- Group 2: Before Tuesday March 23rd, 2021, 14h
- Group 3: Before Wednesday March 24th, 2021, 9h30
- Group 4: Before Tuesday March 23rd, 2021, 9h30

Evaluation: Code and explanation about the code (in groups of 2 or 3 people (preferably 3))
Remark:

- Only groups of two or three people accepted (preferably three).
- Before you leave today lab session, you need to show the lab task results to the professor.
- No late lab/homework will be accepted.
- No plagiarism. If plagiarism happens, both the "lender" and the "borrower" will have a zero.
- Code yourself from scratch. No lab/homework will be considered if any ML library is used.
- Do thoroughly all the demanded tasks.
- Study the theory for the questions.

---

1 Lab tasks

1.1 K-means: Clustering some synthetic data

1. Download from the course site the 2D data stored in `data_kmeans.txt` file.
2. Cluster them using the K-means algorithm using the formulas seen in class.
3. Test your model with some new data.
4. Plot both training and test results in a 2D graph.

*ECE Paris Graduate School of Engineering, 37 quai de Grenelle 75015 Paris, France; jae-yun.jun-kim@ece.fr

2 Homework tasks

2.1 K-means:Clustering some real data

Download from the course site the 6D data stored in `grade_students.csv` file. The source of this dataset is the **The Student/Teacher Achievement Ratio (STAR) Project** organized by the Tennessee State Department of Education in the USA. The reference is the following:
`https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/10766`
This dataset contains 6 features of 5500 students from 79 schools in the state of Tennessee: students' free or reduced-price lunch status, number of absence days, the standarized Stanford Achievement Test Scores for reading, Math, listening and word study.

1. Using the given dataset, cluster the students in 3 clusters (weak, average and gifted clusters) using the K-means algorithm.

2. Interpret your results. That is, relate the input-feature values to the output values and comment your observations.

2.2 PCA: Reducing the dimension of some synthetic data

1. Download from the course site the 2D data stored in `data_pca.txt` file.

2. Implement the PCA algorithm from the formulas seen in class.

3. Indicate the principal axes of the data.

4. Test your model with some new data.

5. Plot both training and test results in a 2D graph.

2.3 PCA: Reducing the dimension of some real data

Download from the course site the 8D data stored in `diabetes.txt` file. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases in the USA. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.
The dataset contains 9 features of 769 patients, and these features are: number of pregnancies, glucose level, blood pressure [mm Hg], skin thickness [mm], insulin level [muU/ml], BMI (body mass index) [weight in kg/m$\hat{2}$], diabetes pedigree function, age, diabetes status.
For further information, you can visit the following Kaggle site:
`https://www.kaggle.com/uciml/pima-indians-diabetes-database`

1. Reduce the 8 dimensional data to a meaningful reduced dimensional space for both diabetic and non-diabetic groups separately.

2. Interpret your results. That is, relate the input-feature values to the output values and comment your observations.