INF436 Machine Learning / Homework 3-2

Principal Component Analysis (PCA)

Jae Yun JUN KIM*

February 19, 2019

Due: Before the next lab session.

Evaluation: Interrogation during the next lab session about:

- code (in group of up to 3 people)
- (theoretical, practical) questions (individual)

Remark:

- Only groups of one/two/three people accepted. Forbidden groups of larger number of people.
- No late homework will be accepted.
- No plagiarism. If plagiarism happens, both the "lender" and the "borrower" will have a zero.
- Code yourself from scratch. No homework will be considered if you solve the problem using any ML library.
- Do thoroughly all the demanded tasks.
- Study the theory for the interrogation.

1 Tasks

A) **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.

B) **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.

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