Cognitive Algorithms Assignment 6

Unsupervised Algorithms for Text Data

Due on Wednesday, January 22nd, 2014 10am via ISIS

In this assignment, you will detect trends in text data and implement Principal Component Analysis (PCA). The text data consists of preprocessed news feeds gathered from `http://beta.wunderfacts.com/` in October 2011, and you will be able to detect a trend related to Steve Jobs death on 5th October 2011.
The data consists of 26800 Bag-of-Words (BOW) features of news published every hour, i.e. the news are represented in a vector which contains the occurence of each word. Here we have many more dimensions (26800) than data points (645). This is why we will implement Linear Kernel PCA instead of standard PCA.
Download the python template `assignment6.py` from the ISIS web site, and download the data set `newsdata.npz`.

1. (**18 points)** Implement Linear Kernel Principal Component Analysis by completing the function stub `pca`. Given data $X \in \mathbb{R}^{D \times N}$, PCA finds a decomposition of the data in k orthogonal principal components that maximize the variance in the data,

   $$X = W \cdot H,$$

   with $W \in \mathbb{R}^{D \times k}$ and $H \in \mathbb{R}^{k \times N}$. The Pseudocode is given in Algorithm 1.

   The function `test_assignment6` helps you to debug your code. You should get a result similiar to Figure 1. (do not worry if you get slighlty complex results due to numerical instabilities).

2. (**4 points)** What happens when you forget to center the data in `pca`? Hand in the resulting plot for the 2D toydata example and explain the result.

Algorithm 1: Linear Kernel PCA

Require: data $x_1, \ldots, x_N \in \mathbb{R}^d$, $N \ll d$, number of principal components k

Center Data
$X = X - 1/N \sum_i x_i$
Compute Linear Kernel
$K = X^\top X$
Compute eigenvectors corresponding to the k largest eigenvalues
$\alpha = \text{eig}(K)$
$W = X\alpha$
Project data onto W
$H = W^\top X$
return W, H

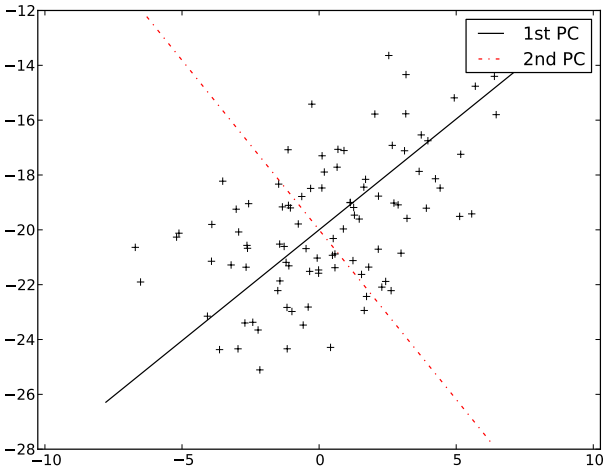

Figure 1: PCA result on 2D toy data

3. **(4 points)** Suppose we only have two data points, $\begin{bmatrix}0\\1\end{bmatrix}$ and $\begin{bmatrix}0\\2\end{bmatrix}$, $X = \begin{bmatrix}0 & 0\\1 & 2\end{bmatrix}$. What would be the principal directions $W = [\mathbf{w}_1, \mathbf{w}_2]$? What will be the variance of the projected data onto each of the principal components $\text{Var}(\mathbf{w}_1^T X)$, $\text{Var}(\mathbf{w}_2^T X)$? What is H?

   Hint: You can obtain the result simply by visualizing the two data points and remembering PCA's objective. Or you can calculate the result using *standard* PCA. With Linear Kernel PCA, you will not be able to compute $\mathbf{w}_2$, because the corresponding eigenvalue is 0.

4. **(4 points)** Detect trends in the text data by calling the provided function `plot_trends` once for PCA and once for Non-Negative Matrix Factorization (NMF) (the code for NMF is provided as well). Which differences do you notice between the algorithms? Hand in the plot of the most prominent trend related to Steve Jobs death for each algorithm.

Please hand in your completed `assignment6.py` via ISIS. Please write your name and your Matrikel Number as the first line of the code. Also hand in a pdf file that contains your name, the answers to the questions and the plots for Question 2 and Question 4. Please also copy your code of your function `pca` in the pdf file.