Data Analysis (046193) HW4

Submission date: 14/6/17

* Submission must be done in pairs.

* Submit a ZIP file containing your files named with 9 digit of your ID. Submission Example: "200567989_123456789.zip".

* Computer part should be completed in the provided IPython notebook and attached inside your zip file.

* Use python version 2.7.

MLE

1. compute the maximum likelihood estimator for a Bernoulli random variable parameter.

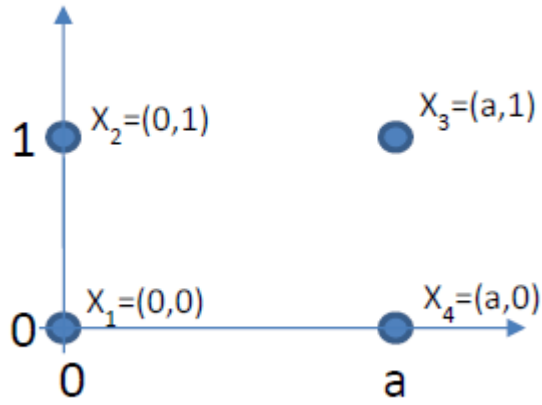
Dimensionality Reduction

2. In this question you are given infinite data points from a given probability distribution as input for the PCA algorithm.
   a. Given Gaussian probability distribution $\mathrm{N}(\mu, \Sigma)$ where $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 11 & -9 \\ -9 & 11 \end{bmatrix}$ compute the first normalized principal direction.
   b. Let us denote the first principal direction as v. Assume we projected all data points on v, compute the variance of the projected points.
   c. We are given a new probability distribution which defined as follows: with probability p we are sampling from the Gaussian distribution defined by $\mu = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 11 & -9 \\ -9 & 11 \end{bmatrix}$ and with probability $(1-p)$ we are sampling the vector $z = a\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ for some scalar a. Compute the first normalized principal direction.
3. PCA: Prove that the 2 first principal components selected by the PCA algorithm preserve maximal variability in the data. (Outline in the lecture notes).

4. At the end of Tutorial 6 we've derived the t-sne gradient where d_{ij} distribution was modeled using the t-distribution. Assume we select a different model, $Q_{ij} = \frac{\exp(-d_{ij}^2)}{Z}$, $Z = \sum_{k \neq l} \exp(-d_{ij}^2)$. Derive the new gradient.

Clustering

5. K-means: Find all possible outputs (after convergence) for the K-mean algorithm with K=2 for the following data (as function of a>0).



*You have to find all possible classification of the 4 points into 2 clusters that given as initialization for the algorithm will not change after inner iteration.

Computer part

1. Answer all questions in the Ipython notebook attached and submit within the zip file.
2. Use sklearn's dbscan clustering implementation, and compare the results with your k-means implementation.