Homework Assignment 6

Lecturer: Kyunghyun Cho

April 15, 2018

1. (a) Why is it more efficient to process data points if they are lower-dimensional vectors? State one reason.

(b) What is a potential trouble of reducing the dimensionality of input vectors before training a classifier? State one reason.

2. (a) Given a training set $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, show that the reconstruction error of principal component analysis (PCA) could be written down as

$$\frac{1}{N}\sum_{n=1}^{N} \|\mathbf{x}_n - \hat{\mathbf{x}_n}\|_2^2 = \sum_{j=q+1}^{d} \mathbf{w}_i^\top \mathbf{C}\mathbf{w}_i,$$

where $\mathbf{w}_i$ is the i-th principal component or the eigenvector of the input covariance matrix $\mathbf{C}$.

(b) Show that

$$\begin{aligned} &\Sigma = \mathbf{W}^\top \mathbf{C}\mathbf{W} \\ \iff &\sigma_j^2 = \mathbf{w}_j^\top \mathbf{C}\mathbf{w}_j, \text{ for all } j = 1, \ldots, d, \end{aligned}$$

where $\mathbf{W}$ is the weight matrix of PCA, $\mathbf{C}$ is the input covariance matrix, and

$$\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_q^2) = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma 2^2 & \cdots & 0 \\ \vdots & 0 & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma_q^2 \end{bmatrix}$$

is the covariance matrix of the code vectors.

3. (Programming Assignment) Complete the implementation of PCA and NMF using Python and scikit-learn. The completed notebooks must be submitted together with the answers to the questions above.

When submitting Jupyter notebooks, make sure to save printed outputs as well.

PCA `https://github.com/nyu-dl/Intro_to_ML_Lecture_Note/blob/master/homeworks/hw5_pca.ipynb`

NMF `https://github.com/nyu-dl/Intro_to_ML_Lecture_Note/blob/master/homeworks/hw5_nmf.ipynb`