CPTS 437: Introduction to Machine Learning (Spring 2018)

Homework #6

Due 4/26/2018

- You need to submit a report in hard-copy before lecture and your code to Blackboard.
- Hard-copy is due in class before lecture and electronic copy is due 2:50PM on Blackboard on the due date.
- Unlimited number of submissions are allowed on Blackboard and the latest one will be graded.
- LFD refers to the textbook "Learning from Data".
- Please read and follow submission instructions. No exception will be made to accommodate incorrectly submitted files.

---

1. (20 points) Let $A = U\Sigma V^T$ be the SVD of A, where $A \in \mathbb{R}^{m\times n}$, $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices, $\Sigma = \text{diag}(\sigma_1, \cdots, \sigma_r, 0, \cdots, 0)$, and $r = \text{rank}(A)$. Show that
   (a) The first r columns of U are eigenvectors of AA^T corresponding to nonzero eigenvalues.
   (b) The first r columns of V are eigenvectors of A^TA corresponding to nonzero eigenvalues.

2. (10 points) Given a symmetric matrix $A \in \mathbb{R}^{3\times 3}$, suppose its eigen-decomposition can be written as
$$A = \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \\ u_{31} & u_{32} & u_{33} \end{pmatrix} \begin{pmatrix} 3 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{21} & u_{31} \\ u_{12} & u_{22} & u_{32} \\ u_{13} & u_{23} & u_{33} \end{pmatrix}. \quad (1)$$
   What is the singular value decomposition of this matrix?

3. (20 points) Given a data matrix $X = [x_1, x_2, \cdots, x_n] \in \mathbb{R}^{p\times n}$ consisting of n data points, and each data point is p-dimensional,
   - Outline the procedure for computing the PCA of X;
   - State what is the "minimum reconstruction error" property of PCA.
   - Prove the minimum reconstruction error property of PCA by using the best low-rank approximation property of SVD.

4. (20 points) **Hierarchical clustering**: Use the similarity matrix in Table 1 to perform single (MIN) and complete (MAX) link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

5. (30 points) **Principal Component Analysis:** In this homework, you will apply the principal component analysis to a collection of handwritten digit images from the USPS dataset. The USPS dataset is in the "data" folder: USPS.mat. The starting code is in the "code" folder. The whole data has already been loaded into the matrix A. The matrix A has shape 3000×256 and contains all the images. Each row in A corresponds to a handwritten digit image (between 0 and 9) with size 16×16. You are expected to implement your solution based on the given codes. The only file you need to modify is the "solution.py" file. You can test your solution by running the "main.py" file.

Table 1: Similarity matrix.

	p1	**p2**	**p3**	**p4**	**p5**
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

(a) (10 points) Complete the ***pca()*** function. Your code will be tested on p = 10, 50, 100, 200, total four different number of the principal components.

(b) (5 points) Complete the ***reconstruction()*** function to reconstruct the data using the selected principal components from (a).

(c) (5 points) Complete the ***reconstruct_error()*** function to measuring the reconstruction error.

(d) (10 points) Run "main.py" to see the reconstruction results and summarize your observations from the results into a short report. When you run the "main.py" file, a subset (the first two) of the reconstructed images based on p = 10, 50, 100, 200 principal components will be automatically saved on the "code" folder. Please attach these images into your report also.

Deliverable: You should submit (1) a hard-copy report (along with your write-up for other questions) that summarizes your results and (2) the "solution.py" file to the Blackboard.

Note: You are NOT supposed to use existing PCA code; instead, you should write your own PCA function. Please read the "Readme.txt" file carefully before you start this assignment.