CSCE 489: Machine Learning (Spring 2019)

Homework #5

Due 4/17/2019 before class

1. You need to submit a report in hard-copy before lecture and your code to eCampus. Your hard-copy report should include (1) answers to the non-programming part, and (2) results of the programming part. Your submission to eCampus should be your code files ONLY. Please put all your code files into a compressed file named "HW#_FirstName_LastName.zip"
2. Hard-copy is due in class before lecture, and code files are due 9:10AM to eCampus on the due date.
3. Unlimited number of submissions are allowed on eCampus and the latest one will be graded. If you make a resubmission after the due date, it will be considered late.
4. LFD refers to the textbook "Learning from Data".
5. Please read and follow submission instructions. No exception will be made to accommodate incorrectly submitted files/reports.
6. All students are highly encouraged to typeset their reports using Word or LaTeX. In case you decide to hand-write, please make sure your answers are clearly readable.

---

1. (20 points) This question is related to the loss functions we discussed in class.
   (a) Describe what are hinge loss, logistic regression loss, and 0-1 loss mathematically. Describe their similarities and differences using the unified picture we developed in class.
   (b) By relying on the result in the above question, consider a point that is correctly classified and distant from the decision boundary. Why would SVM's decision boundary be unaffected by this point, but the one learned by logistic regression be affected?
2. (10 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.
3. (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?
4. (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 p features,

- outline the procedure for computing the PCA of X;
- state what is the "minimum reconstruction error" property of PCA.

5. (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.

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

6. (20 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.

   (a) (5 points) Complete the ***_do_pca()*** method. 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()*** method to reconstruct the reduced data.

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

   (d) (5 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.

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