Assignment 4

CS6510: Applied Machine Learning
IIT-Hyderabad
Jan-Apr 2019

Max Marks: 20
Due: 25 Apr 2019 11:59 pm

This homework is intended to cover programming exercises in the following topics:

- Clustering, Dimensionality Reduction

Instructions

- Please use Google Classroom to upload your submission by the deadline mentioned above. Your submission should comprise of a single file (PDF/ZIP), named `<Your_Roll_No>_Assign4`, with all your solutions.
- For late submissions, 10% is deducted for each day (including weekend) late after an assignment is due. Note that each student begins the course with 7 grace days for late submission of assignments. Late submissions will automatically use your grace days balance, if you have any left. You can see your balance on the CS6510 Marks and Grace Days document.
- You have to use PYTHON for the programming questions.
- Please read the department plagiarism policy. Do not engage in any form of cheating - strict penalties will be imposed for both givers and takers. Please talk to instructor or TA if you have concerns.

Questions: Theory

(No programming required)

1. **Hierarchical Clustering *(4 marks)*:** Given below is the distance matrix for 6 data points

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0					
x_2	0.12	0				
x_3	0.51	0.25	0			
x_4	0.84	0.16	0.14	0		
x_5	0.28	0.77	0.70	0.45	0	
x_6	0.34	0.61	0.93	0.20	0.67	0

(a) Draw a dendrogram for the final result of hierarchical clustering with single link. [*1 mark*]

(b) Draw a dendrogram for the final result of hierarchical clustering with complete link. [*1 mark*]

(c) Change two values from the matrix so that the answer to the last two questions is same. [*2 marks*]

2. **Principal Component Analysis *(4 marks)*:** Suppose each data point $\mathbf{x}$ is an M-dimensional vector of the form $\mathbf{x} = a\delta_k = (0, ..., 0, a, 0, ...)^T$, where a is in the k^{th} slot, and k,a are random variables. k is uniformly distributed over $1, \cdots, M$ and $P(a)$ is arbitrary.

(a) Calculate the covariance matrix. [*1 mark*]

(b) Show that it has one eigenvector of form $(1, ..., 1)$ and that the other eigenvectors all have the same eigenvalue. [*2 marks*]

(c) Discuss whether PCA is a good way to select features for this problem. [*1 mark*]

(**Hint:** Use expectation to compute the covariance matrix: $C = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T]$. You should get C of the form $C_{ij} = \lambda + \mu\delta_{i,j}$ for some λ, μ.)

Questions: Programming

3. **Clustering *(7 marks)*:** DBSCAN, as we discussed in class, is a density-based clustering algorithm. In this problem, you need to implement your own DBSCAN algorithm. You can read more about it from paper that proposed this method [link].

(a) Use the Kmeans clustering algorithm from sklearn and find the number of clusters in dataset1 shared with you. Plot the data points with different colors for different clusters. [*1 mark*]

(b) Implement your own DBSCAN algorithm on the same dataset and plot the data points. [*3 marks*]

(c) What differences do you see between the DBSCAN and k-means methods, and why? [*1 mark*]

(d) Consider the dataset2 (also shared with you) with three clusters. Use (a) and (b) for dataset2, and compare the performance. List your observations clearly, and make conclusions on pros and cons of DBSCAN and k-means. [*2 marks*]

4. **Dimensionality Reduction *(*5 *marks):***

(a) For this problem, you will study Principal Component Analysis on the Iris dataset. The Iris dataset contains classifications of iris plants based on four features: sepal length, sepal width, petal length, and petal width. There are three classes of iris plants on this dataset: Iris Setosa, Iris Versicolor, and Iris Virginica. You can download the dataset from `http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data`. Use the first and second principal components to plot this dataset in 2 dimensions (you can use in-built methods for this). Use different colors for each of the three classes in your plot. Share your observations in your report. [*1.5 marks*]

(b) Now, use `sklearn`'s inbuilt `t-SNE` function to visualize the same data. Include these in your plots, and compare the plots with the PCA plots. Do you see any differences/similarities? Share them in your report. [*1.5 marks*]

(c) Now, compare PCA and t-SNE on the SwissRoll dataset (you can use the inbuilt `sklearn` function to generate this dataset using `http://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_swiss_roll.html`. Now, what do you see? Share your observations and inferences in your report. [*2 marks*]

OPTIONAL QUESTIONS

(No submission required; for your own practice and learning)

5. Show mathematically that k-means is equivalent to a GMM obtained using Expectation Maximization when the component assignment prior is uniform, and when each component has the same covariance.

6. *(For the mathematically inclined)* Suppose we approximately represent each point of the dataset as a linear combination of its k nearest neighbors. Let $(W_k)_{ij}$ be the weight of point x_j in the expansion of x_i minimizing the squared representation error.

   (a) Prove that $M_k(x_i, x_j) = ((IW_k)^T(IW_k))_{ij}$ is a positive semidefinite kernel on the domain $X = \{x_1, \cdots, x_n\}$.

   (b) Let λ be the largest eigenvalue of $(IW_k)^T(IWk)$. Prove that the LLE kernel $K_k^{LLE}(x_i, x_j) = (\lambda 1)I + W_k^T + WkW_k^TW_k)_{ij}$ is positive semidefinite on $\{x_1, \cdots, x_n\}$.

   (c) Prove that kernel PCA using the LLE kernel provides the LLE embedding coefficients for a $d$dimensional embedding as the coefficient eigenvectors $\alpha^2, \cdots, \alpha^{d+1}$. Note that if the eigenvectors are normalized, then dimension i will be scaled by $\lambda_i^{1/2}, i = 1, \cdots, d$.

   (d) Prove that the pseudo-inverse of M is positive semidefinite.

   (e) Prove that performing kernel PCA on M^+(pseudo-inverse of M)is equivalent to LLE up to scaling factors.