CSE 575: Statistical Machine Learning Assignment #3

Instructor: Prof. Hanghang Tong
Out: Oct. 16th, 2018; Due: Nov. 15th, 2018

Submit electronically, using the submission link on Blackboard for Assignment #3, a file named `yourFirstName-yourLastName.pdf` *containing your solution to this assignment (a .doc or .docx file is also acceptable, but .pdf is preferred).*

1 Kmeans [15 points]

Given N data points $x_i, (i = 1, ..., N)$, Kmeans will group them into K clusters by minimizing the loss/cost/distortion function $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{n,k} \|x_n - \mu_k\|^2$, where μ_k is the center of the k^{th} cluster; and $r_{n,k} = 1$ if x_n belongs to the k^{th} cluster and $r_{n,k} = 0$ otherwise. In this exercise, we will use the following iterative procedure

- Initialize the cluster center $\mu_k, (k = 1, ..., K)$;
- Iterate until convergence
  - Step-a: Update the cluster assignments for every data point x_n: $r_{n,k} = 1$ if $k = \text{argmin}_j \|x_n - \mu_j\|^2$; $r_{n,k} = 0$ otherwise.
  - Step-b: Update the center for each cluster k: $\mu_k = \frac{\sum_{n=1}^{N} r_{n,k} x_n}{\sum_{n=1}^{N} r_{n,k}}$

(1) [5 pts] Prove that the above procedure will converge in finite steps.

- *hints: consider whether or not the number of possible cluster assignments is finite.*

(2) [10 pts] Suppose we run Kmeans on the following dataset with six data points (i.e., the six black dots) to find two clusters. If we set the initial cluster centers are $\mu_1 = (0, 1)$ and $\mu_2 = (3, 3)$. How many iterations does the algorithm take until convergence (3 points)? If we only run Kmeans for one iteration, what is the cluster assignment for each data point after Step-a (4 points)? What are the updated cluster centers after Step-b (3 points)?

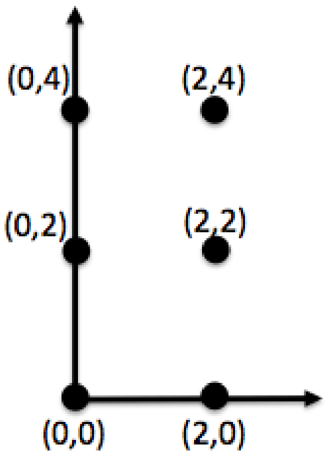
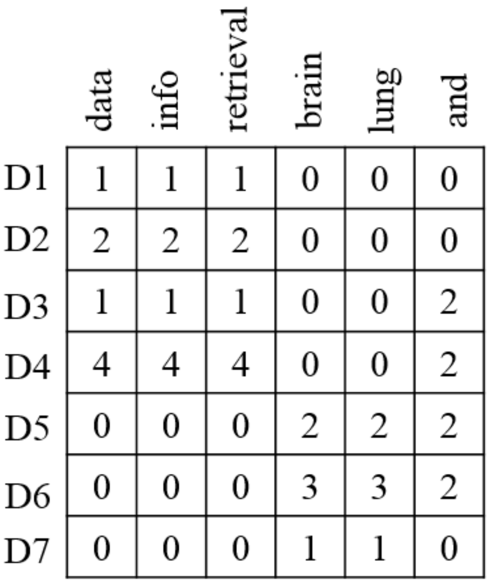


Figure 1: input data points

2 Mixture Model for Document Clustering [15 points]

Given a collection of 7 documents ($D1 - D7$), assume our vocabulary consists of 6 words $\{$'data', 'information', 'retrieval', 'brain', 'lung', 'and'$\}$. The entry in the table indicates the frequency of the corresponding word in the corresponding document. We want to run the mixture model that we introduced in class to find two clusters of the input documents.



	data	info	retrieval	brain	lung	and
D1	1	1	1	0	0	0
D2	2	2	2	0	0	0
D3	1	1	1	0	0	2
D4	4	4	4	0	0	2
D5	0	0	0	2	2	2
D6	0	0	0	3	3	2
D7	0	0	0	1	1	0

Figure 2: Input Data Set for Clustering

Suppose that at the previous iteration, we have the following estimation for the two language models: $P(\theta_1) = P(\theta_2) = 0.5$, $\theta_1 = [0.2,\ 0.3,\ 0.3,\ 0,\ 0,\ 0.2]$ and $\theta_2 = [0.1,\ 0,\ 0,\ 0.3,\ 0.3,\ 0.3]$.

We also have the following estimation for the cluster membership $E[z_{ij}]$ $(i = 1, ..., 7;\ j = 1, 2$ (i.e., $E[z_{ij}]$ is the probability that the $i - th$ document belongs to $j - th$ cluster):

0.9	0.1
1	0
0.8	0.2
0.8	0.2
0.2	0.8
0.1	0.9
0	1

Table 1: $E[z_{ij}]$

- [3 pts] What is $P(d = d_1)$? (You can use the following fact: $0^0 = 1$, $1^0 = 1$ and $0^1 = 0$).
- [4 pts] We run EM algorithm one iteration. What is the updated $E[z_{ij}]$ $(i = 1, ..., 7;\ j = 1, 2)$ after the E-step?
- [4 pts] What is the updated $P(\theta_1)$, $P(\theta_2)$, θ_1, and θ_2, respectively, after the M-step?
- [4 pts] If we run EM algorithm for another iteration, what is the updated $E[z_{ij}]$ $(i = 1, ..., 7;\ j = 1, 2)$ after the E-step?

3 Overfitting [15 points]

(1) Bias Variance Trade-off

[5 pts] Assume we are modeling a number of features using a machine learning model. Recall the relationship among bias, variance, and the complexity of hypothesis class, choose the correct answers in the following table.

	High Model Complexity		**Low Model Complexity**	
Bias	High	Low	High	Low
Variance	High	Low	High	Low

(2) Leave-One-Out-Cross Validation (LOOCV)

We are given the following training data set. Each side of the grid represents a unit lenghth.

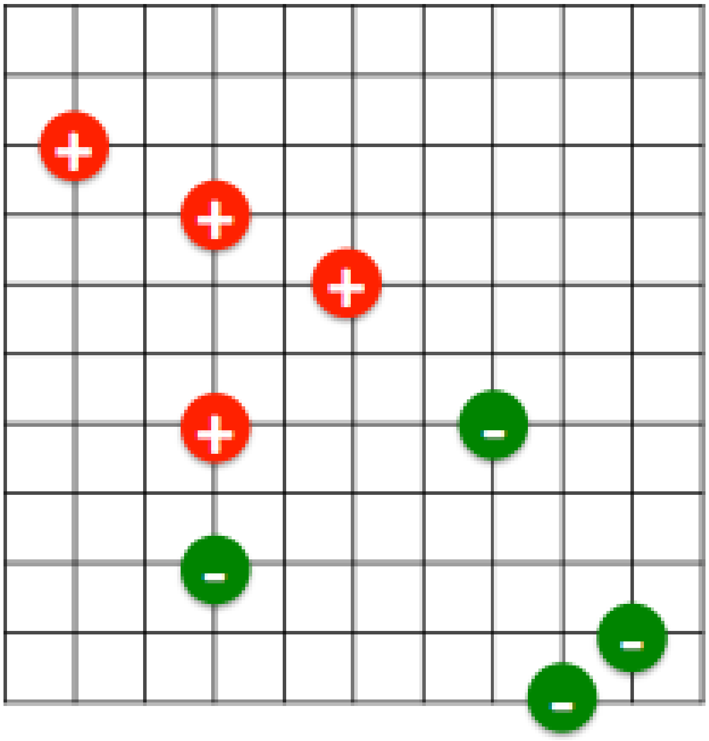


Figure 3: Data points for LOOCV

1. [5 pts] Suppose we use a linear SVM (i.e., no kernel), with some large C value. What is LOOCV for your SVM? Justify your answer.
2. [5 pts] Suppose We want to train 1NN, using L_2 distance. What is the LOOCV for your 1NN classfier? Justify your answers.

4 PCA [15 points]

Suppose we are given 20 points in 2-d space, and the features of each of these 20 data points are finite. Suppose that its first principle component is $u = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})'$.

- [5 pts] Suppose we add one more data point at $(4, 4)$, how would that affect the first princinple component?
- [5 pts] Suppose we add one more data point at $(2, 0)$, how would that affect the first princinple component?
- [5 pts] Suppose we add an infinite number of data points at $(0, 5)$, how would that affect the first princinple component?

5 Implementing Kmeans and Gaussian Mixture Model (GMM) (40 points)

Download the data file from Blackboard. The data contains 128 instances and 13 features. If you encounter missing values in the dataset, use the mean value of that feature instead. DO NOT use scikit-learn package in this problem and please implement from scratch.

(1) [20 pts] Implement Kmeans algorithm and Gaussian Mixture Model **from scratch** in either MATLAB or Python. Initialize the cluster centers by randomly picking them among all the instances.

   - **Kmeans.** Let the number of clusters k ranging from 2 to 10, for each number of clusters, upon convergence of your Kmeans algorithm, compute the objective function value $\sum_{j=1}^{k} \sum_{i=1}^{N} m_{i,j}(\mathbf{x}_i - \mathbf{C}_j)^2$ where N is the number of clusters, $m_{i,j}$ represents the cluster membership and $\mathbf{C}_j$ represents the j-th cluster center. Then plot the objective function value vs. the number of clusters k.
   - **GMM.** We use full covariance matrix here. Let the number of cluster $k = 2$, upon convergence of your GMM, draw a scatter plot to visualize your clustering results. In this problem, only use the first two features as input to draw the scatter plot. Different clusters must be in different colors.

(2) [20 pts] Now implement Principal Component Analysis (PCA) **from scratch** and apply it to reduce the dimensionality of original data from 13 to 2. Run your own Kmeans algorithm and GMM on new data with lower dimensionality.

   - **Kmeans.** Let the number of clusters k ranging from 2 to 10, upon convergence of your Kmeans algorithm, plot the objective function value vs. the number of clusters k.
   - **GMM.** Similarly, we use full covariance matrix here. Let the number of cluster $k = 2$, upon convergence of your GMM, draw a scatter plot to visualize your clustering results. In this problem, only use the first two features as input to draw the scatter plot. Different clusters must be in different colors.

Attach all plotted figures in your assignment report. And your code **SHOULD** be also submitted.