CSE 575: Statistical Machine Learning: Mid-Term 2

Instructor: Prof. Hanghang Tong

April 1st, 2016

First Name:			
Last Name:			
Email:			
ASU ID:			
Q	Topic	Max Score	Score
1	SVM and LOOCV	40	
2	Spectral Clustering	20	
3	Kmeans	20	
4	K-fold Cross Validation	20	
Total:		100	

- This exam book has **10** pages, including this cover page.
- You have 150 minutes in total.
- Good luck!

1 Support Vector Machines and Leave-One-Out-Cross-Validation [40 points]

Given the following dataset in 1-d space, which consists of 2 positive data points at the following coordinates $\{-1, -4\}$ and 2 negative data points at the following coordinates $\{4, 1\}$. Suppose we use a linear SVM (i.e., no kernel), with some large C value, where C is the regularization parameter to control the mis-classification error on the training data set.

1. [4 points] Draw the dataset in 1-d space.

   Sol:

   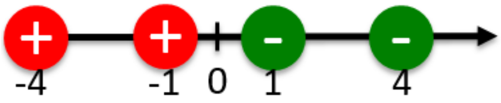

   Figure 1

2. [5 points] Draw the decision boundary of linear SVM trained on this dataset.

   Sol: at the origin.

3. [5 points] In your linear SVM, how many support vectors are there? What are they?

   Sol: 2. -1 and 1.

4. [5 points] What is the leave-one-out-cross-validation (LOOCV) error in your linear SVM? Justify your answer.

   Sol: 0.5

5. [8 points] Now, given another new dataset in 1-d space, which consists of 3 positive data points at the following coordinates $\{-0.8, -1, -4\}$ and 2 negative data points at the following coordinates $\{4, 1\}$. Suppose we use a linear SVM (i.e., no kernel), with some large C value. How many support vectors are there in your linear SVM (4 points)? What is the leave-one-out-cross-validation (LOOCV) error in your linear SVM (4 points) ? Justify your answer.

   Sol: 2 support vectors. LOOCV: 0.2.

6. [8 points] Now, given another new dataset in 1-d space, which consists of 1,000 positive data points at the following coordinates $\{2 - 3 \times i\}$ $(i =$

$1, 2, ..., 1, 000)$ and 1,000 negative data points at the following coordinates $\{3 \times i - 2\}$ $(i = 1, 2, ..., 1, 000)$. Suppose we use a linear SVM (i.e., no kernel), with some large C value. How many support vectors are there in your linear SVM (4 points)? What is the leave-one-out-cross-validation (LOOCV) error in your linear SVM (4 points)? Justify your answer.

Sol: 2 support vectors. LOOCV: 0.001.

7. [5 points] Now, given another new dataset in d-dimensional space ($d > 1$), which consists of 1,000 positive data points whose coordinates are **unknown** and 1,000 negative data points whose coordinates are **unknown**. Suppose we use a kernel SVM, with some large C value. There are 20 support vectors in the resulting SVM. What is a tight upper-bound of the leave-one-out-cross-validation (LOOCV) error in your SVM? Justify your answer.

   Sol: LOOCV: $\leq 20/2000 = 0.01$.

2 Spectral Clustering [20 points]

Given a graph with 6 nodes (i.e., data points) in the following figure, we want to run the spectral clustering for *MinCut* to find two clusters.

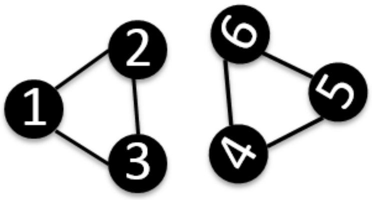

Figure 2: The Input Graph

1. [4 points] Write down the adjacency matrix W of this graph.

   Sol:

$$W = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

Figure 3: W

2. [4 points] Write down the graph Laplacian matrix L of this graph.

   Sol:

$$W = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

Figure 4: L

3. [4 points] Now, if we run the spectral clustering for *MinCut* on this graph to find two clusters, what is the resulting clustering membership vector q?

   Sol: $q = [1,\ 1,\ 1,\ -1,\ -1,\ -1]$ (scaling the numbers is fine)

4. [4 points] Based on the clustering membership vector q, what is the clustering result?

   Sol: $\{1, 2, 3\}$ vs. $\{4, 5, 6\}$

5. [4 points] What is the cut size?

   Sol: 0

3 Kmeans [20 points]

Given N data points x_i $(i = 1, ..., N)$, Kmeans will group them into K clusters by minimizing the 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 question, we will use the following iterative procedure.

- Initialize the cluster center μ_k $(k = 1, ..., K)$;
- Iterate until convergence
  - Step 1: Update the cluster assignments $r_{n,k}$ for each data point x_n.
  - Step 2: Update the center μ_k for each cluster k.

Suppose we run Kmeans on the following dataset with six data points (i.e., the six black dots) to find two clusters.

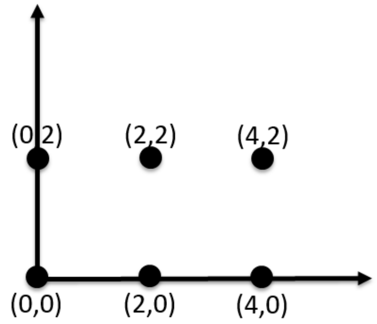

Figure 5: input data points

1 [10 points] Suppose the initial cluster centers are $\mu_1 = (0, 0)$ and $\mu_2 = (5, 0)$. 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 1 (4 points)? What are the updated cluster centers after Step 2 (3 points)?

sol: 1. left fours belong to the first cluster; and the right two belong to the second cluster. $\mu_1 = (1, 1)$ and $\mu_2 = (4, 1)$

2 [10 points] Suppose the initial cluster centers are $\mu_1 = (2, 0)$ and $\mu_2 = (2, 2)$. 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 1 (4 points)? What are the updated cluster centers after Step 2 (3 points)?

sol: 1. bottom three belong to the first cluster; and the upper three belong to the second cluster. $\mu_1 = (2, 0)$ and $\mu_2 = (2, 2)$

4 [20 points] K-fold Cross-Validation

1 [5 points] Given a dataset with $10,000$ data points, we perform a 5-fold cross-validation, how many data points do we set aside for evaluation (i.e., to calculate the test error) at each iteration?

Sol: $10,000/5 = 2,000$

2 [5 points] Given a dataset with $2,000$ data points, we perform a 100-fold cross-validation, how many data points do we set aside for evaluation (i.e., to calculate the test error) at each iteration?

Sol: $2,000/100 = 20$

3 [5 points] Given a dataset with $1,000$ data points, we perform a *k*-fold cross-validation, and at each iteration, we set aside 1 single data point for evaluation (i.e., to calculate the test error). What is the k value? Justify your answer.

Sol: $k = 1,000$

4 [5 points] **True or False**. Given a dataset with N data points, where N is an even number. If we perform a *k*-fold cross-validation, then at each iteration, we will set aside **at most** $N/2$ data points for evaluation (i.e., to calculate the test error). Justify your answer.

Sol: true.