CSE 575: Statistical Machine Learning Assignment #3

Instructor: Prof. Hanghang Tong
Out: Mar. 25th, 2016; Due: Apr. 15th, 2016

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 [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 exercise, we will use the following iterative procedure

- Initialize the cluster center $\mu_k, (k = 1, ..., K)$;
- Iterate until convergence
  - 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.
  - 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) **Convergence of Kmeans** [10 pts]

Prove that the above procedure will converge in finite steps.

- *hints: consider whether or not the number of possible cluster assignments is finite.*
- **Solutions:** Notice that for each cluster assignment, the corresponding cluster centers μ_k(k=1,...K) are unique. Therefore, in each iteration, we must try a new cluster assignment. On the other hand, notice that all possible cluster assignments are finite (K^N). Therefore, the algorithm must converge in finite iterations.

(2) **Kmeans and GMM** [10 pts]

Remember in GMM, $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$, where $\pi_k = p(z_k = 1)$ is the prior for the k^{th} component; and μ_k,Σ_k are the mean and covariance matrix for k^{th} component respectively. In the E-step, we will update $p(z_k = 1|x_n) = \frac{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}$

Now suppose that

(1) $\Sigma_k = \epsilon \mathbf{I}$ where ϵ is some *given* positive number;

(2) $\pi_k \neq 0$ $(k = 1, ..., K)$;

(3) $\|x_n - \mu_i\| \neq \|x_n - \mu_j\|$ for any $i \neq j$.

Under the above assumptions, prove that when $\epsilon \to 0$, $p(z_k = 1|x_n) = r_{n,k}$, where $r_{n,k}$ is the cluster assignment used in Kmeans.

- **Solutions:**

$$\begin{aligned} p(z_k = 1|x_n) &= \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)} \\ &= \frac{\pi_k \text{exp}\{-\frac{1}{2\epsilon}\|x_n - \mu_k\|^2\}}{\sum_{i=1}^{K} \pi_i \text{exp}\{-\frac{1}{2\epsilon}\|x_n - \mu_i\|^2\}} \\ &= \frac{1}{1 + \sum_{i \neq k}(\frac{\pi_i}{\pi_k})\text{exp}\{\frac{1}{2\epsilon}(\|x_n - \mu_k\|^2 - \|x_n - \mu_i\|^2)\}} \end{aligned} \tag{1}$$

Therefore, if $\|x_n - \mu_k\| = \min_i \|x_n - \mu_i\|$, for each $i \neq k$, we have $\|x_n - \mu_k\|^2 - \|x_n - \mu_i\|^2 < 0$. Thus as $\epsilon \to 0^+$, $\text{exp}\{\frac{1}{2\epsilon}(\|x_n - \mu_k\|^2 - \|x_n - \mu_i\|^2)\} \to 0$. So, $p(z_k = 1|x_n) \to 1$.

On the other hand, if $\|x_n - \mu_k\| \neq \min_i \|x_n - \mu_i\|$. Let $\|x_n - \mu_{\tilde{k}}\| \neq \min_i \|x_n - \mu_i\|$, we have $\|x_n - \mu_k\|^2 - \|x_n - \mu_{\tilde{k}}\|^2 > 0$. Thus as $\epsilon \to 0^+$, $\text{exp}\{\frac{1}{2\epsilon}(\|x_n - \mu_k\|^2 - \|x_n - \mu_{\tilde{k}}\|^2)\} \to +\infty$. So, $p(z_k = 1|x_n) \to \frac{1}{1+\infty} = 0$.

2 K-means and Matrix Factorization [10 points]

Given n data points in d dimensional space, we can represent them as an $n \times d$ data matrix X, where the rows of X are different data points and columns are different features.

1. [5 points] K-means clustering can be viewed as a special form of matrix low-rank approximation. That is the optimization objective of k-means is equivalent to

$$\text{argmin}_{F,G} \|X - F \cdot G\|_{fro}^2 \tag{2}$$

where $\|.\|_{fro}$ is the Frobenius norm, F and G are two low-rank matrices with some appropriate constraints. What is the size constraint on F and G, respectively? What are additional constraints we need to impose on F and/or G, so that Equation (2) is equivalent to the optimization objective of k-means?

Solutions: $F : n \times k$ is the cluster membership matrix (each row of F has one and only one 1; and $G : k \times d$ is cluster-description matrix (each column of G is a cluster center). k is the number of clusters.

2. [5 points] Suppose we want to solve the optimization problem in Equation (2) in an alternative way. That is, after some initialization on F and G, we alternatively update F and G iteratively. In each iteration, we (a) first fix F and update G as $\text{argmin}_G \|X - F \cdot G\|_{fro}^2$; and then we fix G and update F as $\text{argmin}_F \|X - F \cdot G\|_{fro}^2$. We repeat this process until convergence. Which step in k-means algorithm does step-(a) correspond to? Which step in k-means algorithm does step-(b) correspond to?

Solutions: step-(a): fix the cluster-membership, update the cluster centers. step-(b): fix the cluster centers, update the cluster membership.

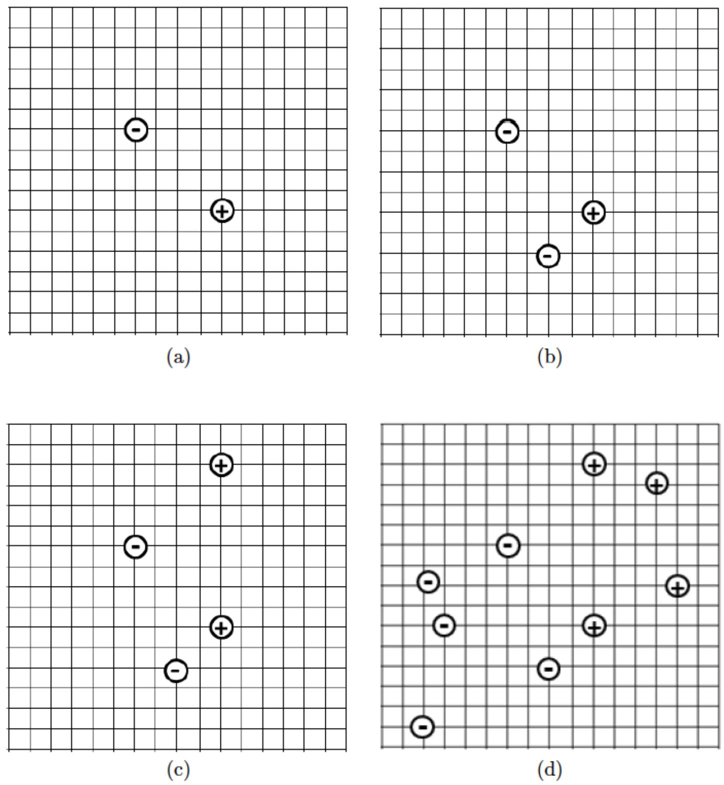


Figure 1: Training Data Set for 1NN Classifiers

3 Leave-One-Out-Cross Validation (LOOCV) for 1NN (i.e., *1*-Nearest Neighbors Classifier) [20 points]

For each of the following figures, we are given a few data points in the 2-d space, each of which is labeled as either '+' or '-'. We want to train 1NN, using L_2 distance. What is the LOOCV for each of the four figures? Justify your answers.

Solutions: 100%; 100%; 100%; $2/9$

4 Leave-One-Out-Cross Validation (LOOCV) for Support Vector Machines [15 points]

1. [5 points] Suppose we use a linear SVM (i.e., no kernel), with some large C value, and are given the following data set.

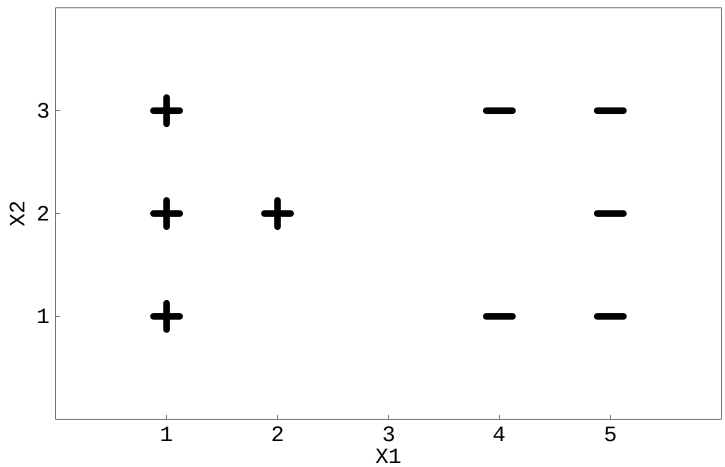


Figure 2

What is LOOCV for your SVM? Justify your answer.

Solution: 0

2. [10 points] In general, Suppose we are use a linear SVM (i.e., no kernel), with some large C value on a training data set with n examples, and there are k support vectors in the trained SVM classifier. What is the (tight) upper-bound of LOOCV of your SVM classifier? Justify your answer.

   Solution: k/n (this is because none of the non-support vectors can be mis-classified in the LOOCV process).

5 PCA [15 points]

Suppose we have the following data points in 2-d space $(0, 0)$, $(-1, 2)$, $(-3, 6)$, $(1, -2)$, $(3, -6)$.

1 [5pts] Draw them on a 2-d plot, each data point being a dot.

2 [5pts] What is the first principle component (2 pts)? Give 1-2 sentences justification (3 pts). (*Hints:* You do not need to run matlab to get the answer.)

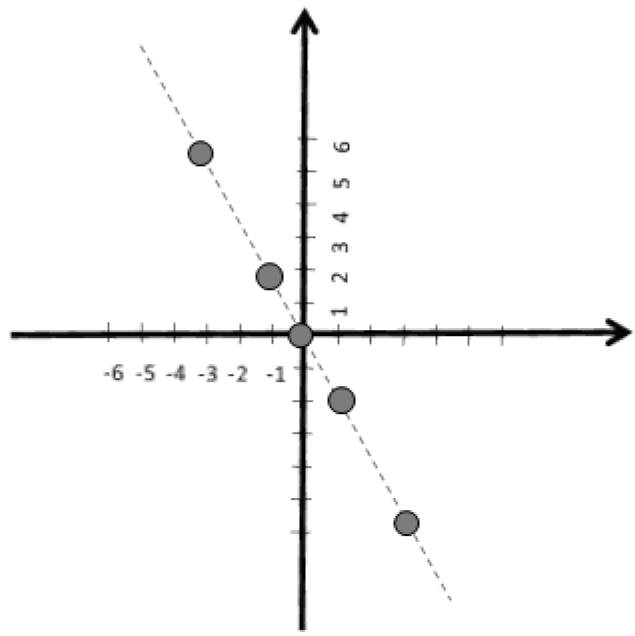
sol: $\frac{1}{\sqrt{5}}(-1, 2)$

2 [5pts] What is the second principle component (2 pts)? Give 1-2 sentences justification (3 pts). (*Hints:* You do not need to run matlab to get the answer.)

sol: $\frac{1}{\sqrt{5}}(2, 1)$

6 HMM [20 points]

Suppose that we use four distinct words to write a paragraph with 100 segments, and we treat each word in the paragraph as a segment. We want to infer three posible class labels of all the segments in this paragraph, including (a) location (b) person name and (c) background by HMM (Hidden Markov Models).



1. [5pts] What is the size of the state transition probability matrix in our HMM model?

sol: 3×3

2. [5pts] What is the size of the state-observation probability matrix?

sol: 3×4

3. [5 pts] In a particular trial, how many observations do you see [2pts]? What is the length of the path of states [3pts]?

sol: 100 and 100

4. [5pts] Suppose that the first state is about 'background', how many different possible state paths are there in total?

sol: 3^{99}