CSE 575: Statistical Machine Learning

Mid-Term 2

Instructor: Prof. Hanghang Tong
October 26th, 2017

First Name:			
Last Name:			
Email:			
ASU ID:			
Q	Topic	Max Score	Score
1	Logistic Regression	20	
2	SVM	20	
3	KNN	20	
4	Kmeans	40	
Total:		100	

- This exam book has **8** pages, including this cover page and a blank page at the end.
- You have 75 minutes in total.
- Good luck!

1 Training Logistic Regression [20 points]

Suppose we have one positive example $x_1 = (1)$; and one negative example $x_2 = (-1)$ in one-dimensional space. We use the standard gradient ascent method (without any additional regularization terms) to train a logistic regression classifier. Recall that in logistic regression, we assume $P(y = 1|x, \mathbf{w}) = \frac{1}{1+e^{-(w_1+w_2x)}}$, where $\mathbf{w} = (w_1, w_2)$ is the weight vector we aim to learn from the training data by maximizing the log conditional data likelihood $l(\mathbf{w}) = \sum_{i=1}^{2} lnP(y_i|x_i, \mathbf{w})$.

- [5 pts.] Plot the training examples in a 1-d plot.

  Solutions: one at 1 and the other at -1.

- [5 pts.] How many independent parameters are there in this logistic regression classifier?

  Solutions: 2

- [3 pts.] Assume that the weight vector starts at the origin, i.e., $\mathbf{w} = (0, 0)'$. What is the derivative of the log conditional likelihood function $l(w)$ wrt $\mathbf{w} = (0, 0)'$? Suppose the learning rate $\eta = 1$, what is the new $\hat{\mathbf{w}}$ after the first iteration?

  Solutions: $\frac{\partial l(\mathbf{w})}{\partial \mathbf{w}}|_{\mathbf{w}=(0,0)'} = (0, 1)'$. Therefore $\hat{\mathbf{w}} = (0, 0)' + 1 \cdot (0, 1)' = (0, 1)'$. (2 pts for the derivative and 1 pt for the updated w)

- [3 pts.] What is the derivative of the log conditional likelihood function $l(w)$ wrt $\hat{\mathbf{w}}$? Suppose we use the same learning rate $\eta = 1$, what is the new $\hat{\hat{\mathbf{w}}}$ after the second iteration?

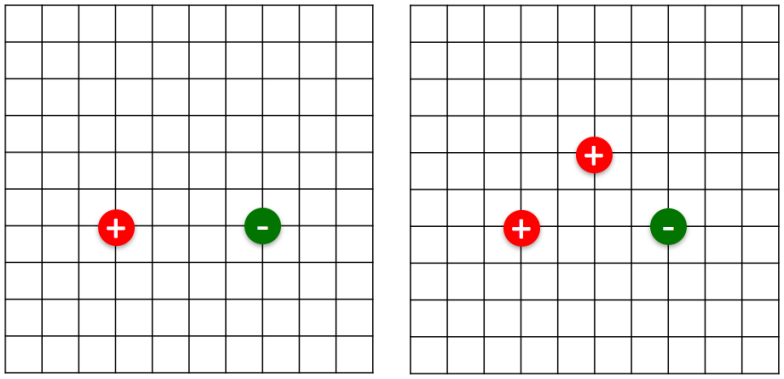
  Solutions: $\frac{\partial l(\mathbf{w})}{\partial \mathbf{w}}|_{\mathbf{w}=(0,1)'} = (0, \frac{2}{1+e})'$. Therefore $\hat{\hat{\mathbf{w}}} = (0, 1)' + 1 \cdot (0, \frac{2}{1+e})' = (0, \frac{3+e}{1+e})'$. (2 pts for the derivative and 1 pt for the updated w)

- [4 pts.] If we use the gradient ascent by an infinite number of iterations, what would be the final $\mathbf{w}$?

  Solutions: $\mathbf{w} = (0, \infty)'$. (get 2 pts if saying w is infinity)

2 Support Vectors and Margins [20 points]

In each of the following figures, we are given some data points in 2-d space and we aim to train a hard margin linear SVM classifier. The length of each grid is 1. For each case, circle the support vectors of your SVM classifier (3 pts for each case). What is the size of the margin of your SVM classifier in each case (2 pts for each case)?



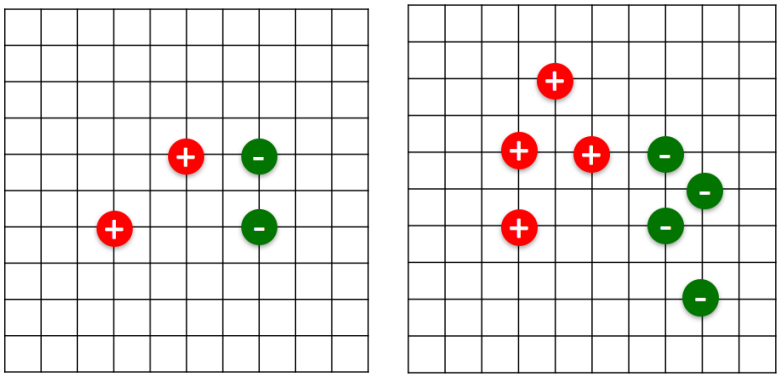


Figure 1: Support Vectors and Margins

Solution:

a both points are support vectors. margin is 4

b all three points are support vectors. margin is $2\sqrt{2}$

c the right three points are support vectors. margin is 2

d the rightmost positive point and the two leftmost negative points are support vectors. margin is 2

3 The decision boundary 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 '-'. Draw the decision boundary for 1NN, assuming we use L_2 distance (5 pts for each case).

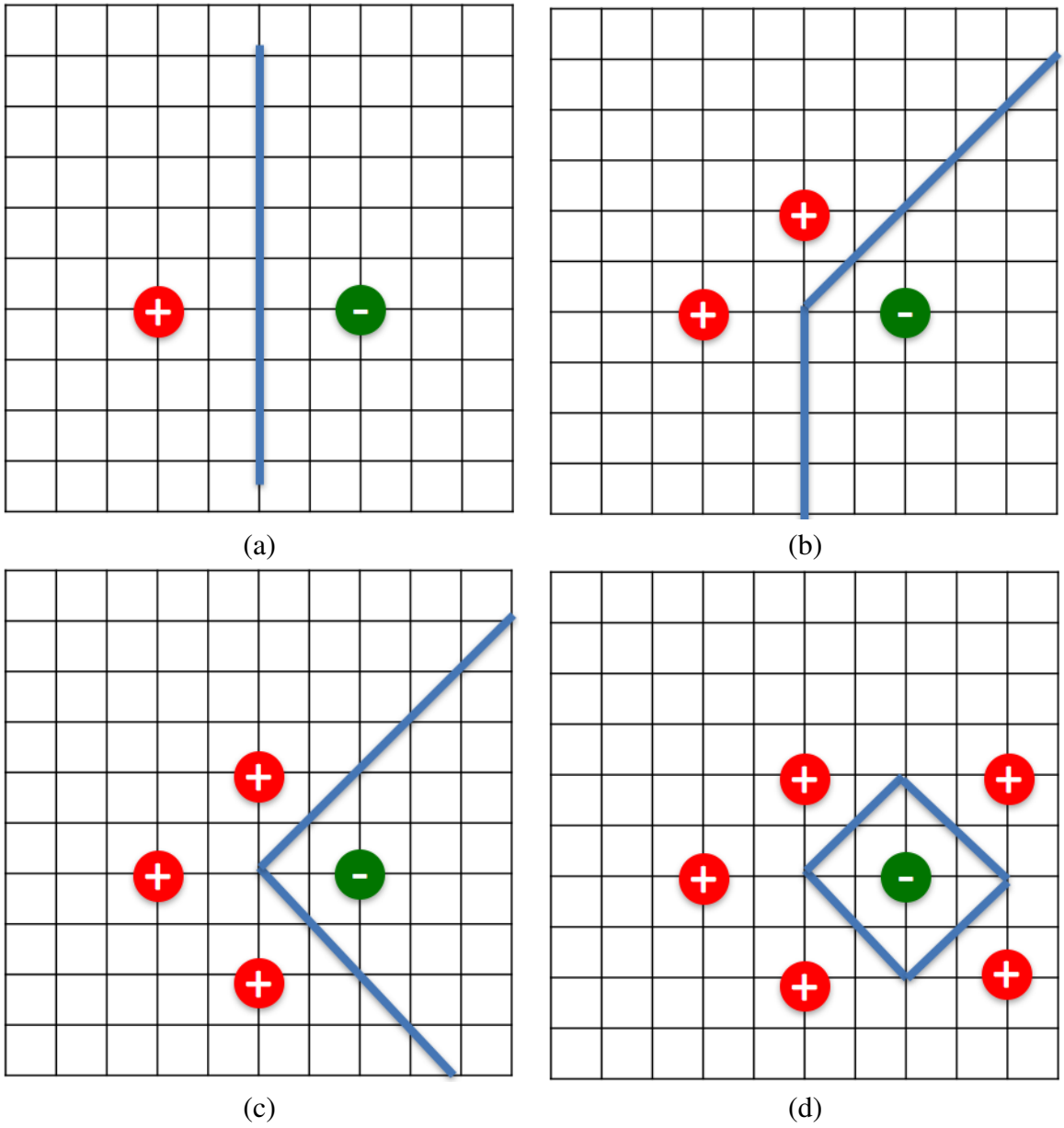


Figure 2: Training Data Set for 1NN Classifiers

4 Kmeans [40 points]

Given N data points x_i $(i = 1, ..., N)$, Kmeans will group them into K clusters by minimizing the loss/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.
- [5 pts.] Given 8 data points in 1-d space: $x_1 = -3$, $x_2 = -1$, $x_3 = 0.5$, $x_4 = 2$, $x_5 = 3$, $x_6 = 4$, $x_7 = 7$ and $x_8 = -5$. Plot these eight data points in 1-d space.

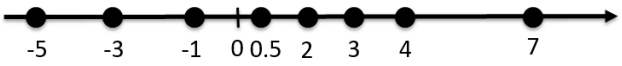


Figure 3: input data points

- [10 pts.] Suppose the initial cluster centers are $\mu_1 = -1$ and $\mu_2 = 3$ (i.e., $K = 2$). If we only run Kmeans for one iteration on the above data set, what is the cluster assignment for each data point after Step 1 (5 pts)? What are the updated cluster centers after Step 2 (5 pts)?

  sol: the left four belong to the first cluster; and the right four to the second cluster. $\mu_1 = -2.125$ and $\mu_2 = 4$.

- [15 pts.] Suppose the initial cluster centers are $\mu_1 = -4$, $\mu_2 = 0$ and $\mu_3 = 3$ on the above data set (i.e., $K = 3$). If we only run Kmeans for one iteration, what is the cluster assignment for each data point after Step 1 (5 pts)? What are the updated cluster centers after Step 2 (5 pts)? What is the loss function J after this iteration (5 pts)?

  sol: $\{-5, -3\}$, $\{-1, 0.5\}$ and $\{2, 3, 4, 7\}$ $\mu_1 = -4$, $\mu_2 = -0.25$, $\mu_3 = 4$. $J = 1 + 1 + (\frac{3}{4})^2 + (\frac{3}{4})^2 + 4 + 1 + 9 = 17.125$

- [10 pts.] If we run Kmeans on the above dataset to find eight clusters (i.e., $K = 8$). What is the **optimal** cluster assignment for each data point? (5 pts)? What is the corresponding loss function J for the optimal clustering assignment (5 pts)?

  sol: each data point forms its own clusters. $J = 0$.

Blank Page