CSE 575: Statistical Machine Learning: Mid-Term 1

Instructor: Prof. Hanghang Tong

February 19th, 2016

First Name:			
Last Name:			
Email:			
ASU ID:			
Q	Topic	Max Score	Score
1	MLE	20	
2	Decision Boundary of 1NN	20	
3	Distance Metric and 1NN	20	
4	Naive Bayes	25	
5	Logistic Regression	15	
Total:		100	

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

1 Maximum Likelihood Estimation (20 points)

[5 points.] Suppose we flip a coin, and observe either a head or a tail. The probability of observing a head in each trial is p $(0 \leq p \leq 1)$. If we flip the coins five times, and observe *(head, head, head, head, head)*, what is the maximum likelihood estimation of p? Justify your answer.

Solution: likelihood is p^5, and $p = 1$ maximizes the likelihood.

[5 points.] Suppose we flip a coin, and observe either a head or a tail. The probability of observing a head in each trial is p $(0 \leq p \leq 1)$. If we flip the coins five times, and observe *(head, head, head, tail, head)*, what is the maximum likelihood estimation of p? Justify your answer.

Solution: likelihood is $p^4(1-p)$. take the logorithm, and calculate its derivative, and set it as zero, we have

$$4/p - 1/(1-p) = 0$$

which gives $p = 0.8$.

[5 points.] Suppose we flip a coin, and observe either a head or a tail. The probability of observing a head in the i^{th} trial is $i * p$ $(i = 1, 2, ..., 5)$. If we flip the coins five times, and observe *(head, head, head, head, head)*, what is the maximum likelihood estimation of p? Justify your answer.

Solution: likelihood is $p * 2p * 3p * 4p * 5p$. In the meanwhile, we have $0 \leq 5p \leq 1$. $p = 0.2$ gives the maximum likelihood

[5 points.] Suppose we flip a coin, and observe either a head or a tail. The probability of observing a head in the i^{th} trial is $i * p$ $(i = 1, 2, ..., 5)$. If we flip the coins five times, and observe *(head, head, head, tail, head)*, what is the maximum likelihood estimation of p? Justify your answer.

Solution: likelihood is $p * 2p * 3p * (1 - 4p) * 5p$. take the logorithm, and calculate its derivative, and set it as zero, we have

$$4/p - 4/(1-4p) = 0$$

which gives $p = 0.2$.

2 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 points for each case**).

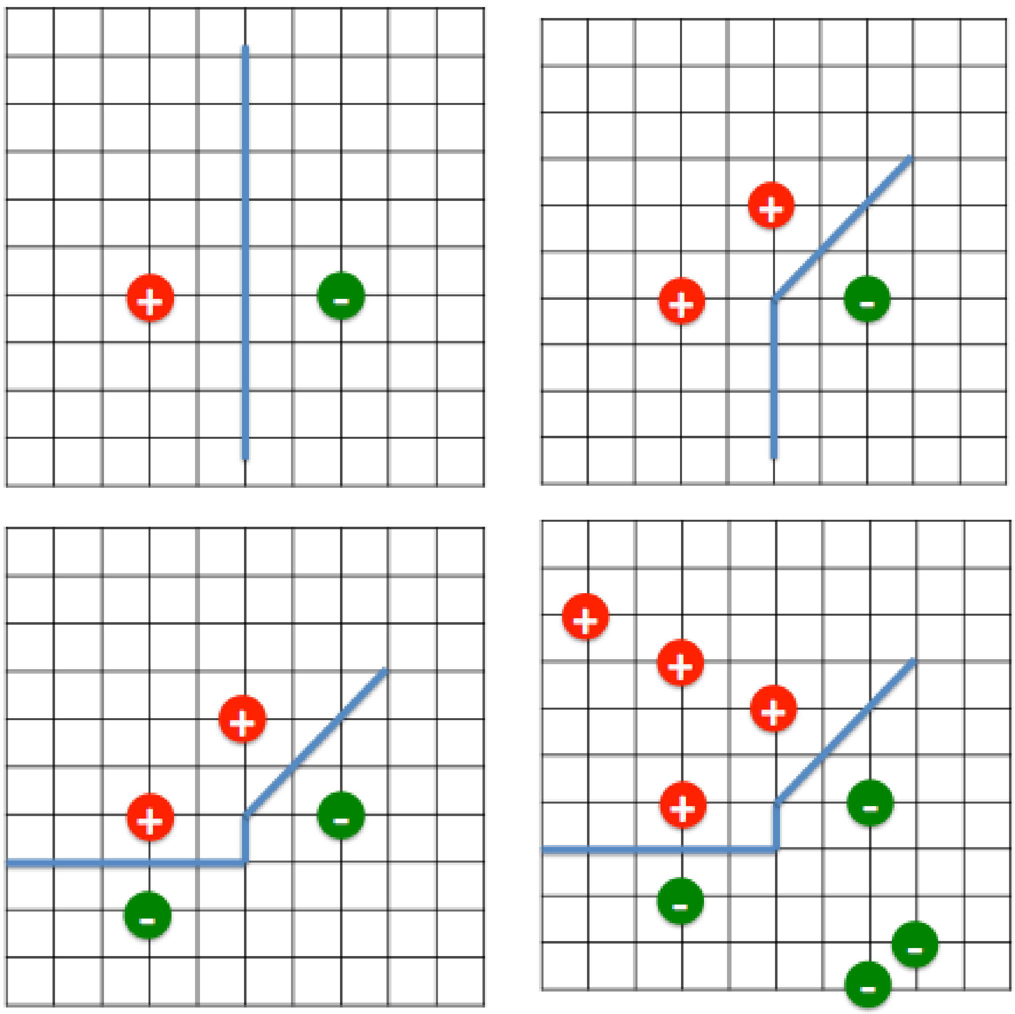

Figure 1: Decision Boundary for 1NN Classifiers (the lines and the transition points must be on the right locations in order to get the points)

3 1NN Classifier and Distance Metric (20 points)

[6 points]. Given two training data points: (1) a positive example at $(a, 0)$ and (2) a negative example at $(-a, 0)$, where $a > 0$, what is the decision boundary of 1NN classifier if we use L_2 distance? Justify your answer.

Solution: the y-axis.

[6 points]. Given two training data points: (1) a positive example at $(a, 0)$ and (2) a negative example at $(-a, 0)$, what is the decision boundary of 1NN classifier if we use L_2 distance **when** $a = 0$? Justify your answer.

Solution: the entire space.

[6 points]. Given two training data points: (1) a positive example at $(1, 0)$ and (2) a negative example at $(-1, 0)$, what is the decision boundary of 1NN classifier if we use L_∞ distance? Recall that the L_∞ distance between two d dimensional data points $X = (x_1, x_2, ..., x_d)$ and $Y = (y_1, y_2, ..., y_d)$ is defined as $\|X - Y\|_\infty = \max_{i\ (i=1,2,...d)} |x_i - y_i|$.

Solutions: shown in the following figure.

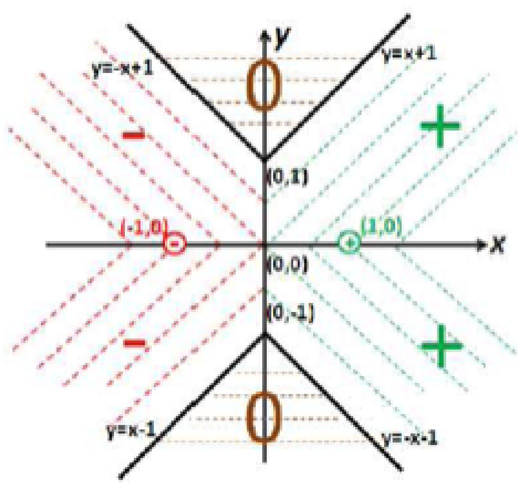

Figure 2: Decision Boundary of 1NN Classifiers with L_∞ Distance.

[2 points]. Given two training data points: (1) a positive example at $(a, 0)$ and (2) a negative example at $(-a, 0)$, what is the decision boundary of 1NN classifier if we use L_∞ distance **when** a goes to infinite?

Solution: the entire space. for any test example (x, y), where x and y are finite numbers, $d_1 = max(|x-a|, |y|) = |x-a| = \infty = d_2 = max(|x+a|, |y|) = |x+a|$.

4 Naive Bayes Classifier (25 points)

Given the training data set in Figure 3, we want to train a binary classifier. In the figure, (1) the last column is the binary class label; (2) each of the first four columns is a binary feature, and (3) each row is a training example.

Input Feature $X = (x_1, x_2, x_3, x_4)$ Class Label Y

x_1	x_2	x_3	x_4	Y
1	1	0	0	1
0	1	1	0	1
0	0	1	1	0
1	1	0	0	1
1	0	0	1	0

Figure 3: Training Data Set

[5 points.] If we want to train a **Bayes Classifier**, how many *independent* parameters are there in your classifier? Justifiy your answer.

Solutions: $P(y = 1)$, for each class label, we have $2^4 - 1 = 15$ independent parameters. Total $2 * (2^4 - 1) + 1 = 31$ independent parameters.

[5 points.] If we want to train a **Naive Bayes Classifier**, how many *independent* parameters are there in your classifier? Justifiy your answer.

Solutions: (1) $P(y = 1)$, (2) for each class lable and each dimension of the feature, one independent parameter $P(x_i = 1|y = i)$ $(i = 1, 0)$. Total $2 * 4 + 1 = 9$ independent parameters.

[4 points.] Using the standard MLE (maximum likelihood estimation) to train a **Naive Bayes Classifier**, what is your estimation for $P(Y = 0)$?

Solutions: $P(Y = 0) = 0.4$.

[3 points.] Using the standard MLE (maximum likelihood estimation) to train a **Naive Bayes Classifier**, what is your estimation for $P(x_3 = 1|Y = 0)$?

Solutions: $P(x_3 = 1|Y = 0) = 0.5$.

[3 points.] Using the standard MLE (maximum likelihood estimation) to train a **Naive Bayes Classifier**, what is your estimation for $P(x_2 = 1|Y = 1)$?

Solutions: $P(x_2 = 0|Y = 1) = 0$.

[5 points.] Suppose we use the standard MLE (maximum likelihood estimation) to train a **Naive Bayes Classifier**. Now given a test example $X = (1, 0, 1, 1)$, what is the class label your classifier will predict? Justify your answer.

Solutions: $P(Y = 1|X) \propto P(Y = 1)P(x_1 = 1|Y = 1)P(x_2 = 0|Y = 1)P(x_3 = 1|Y = 1)P(x_4 = 1|Y = 1) = 0$ (due to the above answer). Therefore, the classifier will predict as negative example (i.e., $Y = 0$).

5 Logistic Regression (15 points)

Suppose we have two positive examples $x_1 = (1,1)$ and $x_2 = (1,-1)$; and two negative examples $x_3 = (-1,1)$ and $x_4 = (-1,-1)$. We use the standard gradient ascent method (without any additional regularization terms) to train a logistic regression classifier.

[4 points.] In your logistic regression classifier, how many *independent* parameters are there?

Solutions: 3. (get two points if the answer is 2)

[4 points.] Assume that the weight vector starts at the origin, i.e., $w_0 = (0,0,0)'$. What is $\prod_{i=1}^{4} P(y_i|x_i, w_0)$, where $y_i (i = 1,2,3,4)$ is the corresponding class label?

Solutions: 1/16. $P(y_i = 1|x_i, w_0) = P(y_i = 0|x_i, w_0) = 1/2$

[4 points.] Assume we use the standard gradient ascent method (without any additional regularization terms) to train a logistic regression classifier, what is the final weight vector w?

Solutions: The final weight vector $w = (0, \infty, 0)'$. (simply saying that w goes to infinite is also fine).

[3 points.] [**True or False**]. Now, suppose we add more training examples whose class labels are either positive or negative. We might end up with a non-linear binary logistic regression classifer, depending on how these new training examples are distributed in the 2-d space.

Solutions: F (binary LR is always linear).