CS 7641 CSE/ISYE 6740 Mid-term Exam (2014 Fall)

Le Song

10/16 Thr, 1:35 - 2:55 pm

- Name:
- GT ID:
- E-mail:

Problem	Point	Your Score
1	25	
2	15	
3	10	
4	10	
5	10	
6	30	
Total	100	

Instructions:

- Try your best to be clear as much as possible. No credit may be given to unreadable writing.
- The exam is open book and open note, but **no electronic devices** (including smart phones) are allowed.
- Good luck!

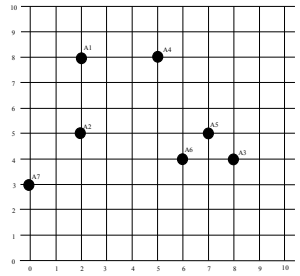
1 Clustering [25 pts]

We have the following 7 data points:

$$A_1 = (2,8), A_2 = (2,5), A_3 = (8,4), A_4 = (5,8), A_5 = (7,5), A_6 = (6,4), A_7 = (0,3)$$

The distance matrix based on the Euclidean distance is given below:

	A_1	A_2	A_3	A_4	A_5	A_6	A_7
A_1	0	$\sqrt{9}$	$\sqrt{52}$	$\sqrt{9}$	$\sqrt{34}$	$\sqrt{32}$	$\sqrt{29}$
A_2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{8}$
A_3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{4}$	$\sqrt{65}$
A_4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{50}$
A_5					0	$\sqrt{2}$	$\sqrt{53}$
A_6						0	$\sqrt{37}$
A_7							0



Suppose we run the K-means with Euclidean distance to cluster the 7 points into 3 clusters. The initial centers of each cluster are A_1, A_4 and A_7. Run the K-means algorithm for only 1 iteration.

(a) At the end of this iteration, write down the new clusters. [5 pts]

Answer: $\{A_1\}, \{A_3, A_4, A_5, A_6\}, \{A_7, A_2\}$

(b) At the end of this iteration, compute the coordinates of new centroids. [5 pts]

Answer: (2,8),(6.5,5.25),(1,4)

Now suppose we use Manhattan distance $|x_1 - x_2| + |y_1 - y_2|$ for the distance measure between two points (x_1, x_2) and (y_1, y_2). The initial centers of each cluster are still A_1, A_4 and A_7. Run the K-means algorithm for only 1 iteration.

(c) At the end of this iteration, write down the new clusters. [5 pts]

Answer: $\{A_1, A_2\}, \{A_3, A_4, A_5, A_6, A_8\}, \{A_7\}$

Consider the bottom up hierarchical clustering to realize the partition of the 7 data points into clusters. In homework 1, we've defined some of the most commonly used distance metrics between two clusters:

- Single linkage: the minimum distance between any pairs of points from the two clusters, i.e.

$$\min_{\substack{i=1,\dots,m \\ j=1,\dots,p}} \|x_i - y_j\|$$

- Complete linkage: the maximum distance between any parts of points from the two clusters, i.e.

$$\max_{\substack{i=1,\dots,m \\ j=1,\dots,p}} \|x_i - y_j\|$$

Suppose in current iteration, we have clusters: $\{A_1, A_2, A_7\}, \{A_4\}, \{A_3, A_5, A_6\}$.

(d) For the next iteration, which two clusters will be merged if we use single linkage? [5 pts]

Answer: $\{A_1, A_2, A_7\}, \{A_4\}$

(e) For the next iteration, which two clusters will be merged if we use complete linkage? [5 pts]

Answer: $\{A_4\}, \{A_3, A_5, A_6\}$

2 Principal Component Analysis [15 pts]

Suppose we have four points in 3-dimensional Euclidean space, namely $(2, 0, 2)$, $(3, -1, 3)$, $(4, -2, 4)$, and $(5, -3, 5)$.

(a) Find the first principal component. [5 pts]

Answer: $\frac{1}{\sqrt{3}}[1, -1, 1]^\top$

(b) When we reduce the dimensionality from 3 to 1 based on the principal component you found in (a), what is the reconstruction error in terms of variance? [5 pts]

Answer: 0

(c) Suppose $X \in \mathbb{R}^{d \times n}$ is centered data, with n data points. Let $w \in \mathbb{R}^d$ be the principal component. Prove that there exists $\alpha \in \mathbb{R}^n$, such that $w = X\alpha$. [5 pts]

Answer: Since, $1/nXX^\top w = \lambda w$, w lies in the column space of X, therefore, $w = X\alpha$.

3 Expectation Maximization [10 pts]

In the table below, there are 5 observations, out of which in example 4 X_2 value is missing. Variable X_2 is dependent on X_1 and hence the distribution can be specified using three parameters $\hat{P}(X_1 = 1)$, $\hat{P}(X_2 = 1|X_1 = 1)$ *and* $\hat{P}(X_2 = 1|X_1 = 0)$. To approximate the missing value, we will use EM.

Example	X_1	X_2
1	0	1
2	1	0
3	1	0
4	1	?
5	0	1

The EM process has run for several iterations. At this point the parameter estimates are

$$\hat{\theta}_{X_1=1} = \hat{P}(X_1 = 1) = 0.6$$

$$\hat{\theta}_{X_2=1|X_1=1} = \hat{P}(X_2 = 1|X_1 = 1) = 0.6$$

$$\hat{\theta}_{X_2=1|X_1=0} = \hat{P}(X_2 = 1|X_1 = 0) = 1$$

(a) Proceed one more E-step. List all the updated variables and their values. [5 pts]

Answer: The expected value of X2 for example 4 $\hat{P}(X_2 = 1|X_1 = 1)$

(b) Proceed one more M-step after (a). List all the updated variables and their values. [5 pts]

Answer: New estimates for all probabilities. $\hat{\theta}_{X_1=1}$ and $\hat{\theta}_{X_1=1|X_1=1}$ do not change. $\hat{\theta}_{X_1=1|X_1=1}$ becomes 0.2. Parametrizing the distribution in the following manner -

$$P(X_1 = 1) = \Theta$$

$$P(X_2 = 1|X_1 = 1) = \alpha$$

$$P(X_2 = 1|X_1 = 0) = \beta$$

the log-likelihood can be written as (based on the solutions to Missing value estimation problem pdf shared in T-square)

$$log(1-\Theta)\beta + 2log(1-\alpha)\Theta + log\sum_{X_2} P(X_1^4, X) + log(1-\Theta)\beta \quad (1)$$

Now using Jensen's inequality,

$$log\sum_{X_2} P(X_1^4, X) \geq \sum_{X_2} P(X_2|X_1^4)^t \; logP(X_1^4) \geq$$

Where α^t is the previous estimate of α from the E-step. Therefore for α we need to maximize -

$$2log(1-\alpha)\Theta + \alpha^t logP(X_1^4, 1) + (1-\alpha^t)logP(X_1^4, 0) \quad (2)$$

$$= 2log(1-\alpha)\Theta + \alpha^t log\alpha + (1-\alpha^t)log(1-\alpha) \quad (3)$$

Taking derivatives w.r.t α, we find $\alpha = \frac{\alpha^t}{3}$. Thus the new value will be 0.2.

4 Logistic Regression [10 pts]

Logistic regression is named after the log-odds of success (the logit of the probability) defined as below:

$$\ln\left(\frac{P[Y=1|X=x]}{P[Y=0|X=x]}\right)$$

where

$$P[Y=1|X=x] = \frac{\exp(w_0 + w^T x)}{1+\exp(w_0 + w^T x)}$$

(a) Show that log-odds of success is a linear function of X. [6 pts]

Answer: $ln(\frac{P[Y=1|X=x]}{P[Y=0|X=x]}) = ln(\frac{\frac{exp(w_0+w^T x)}{1+exp(w_0+w^T x)}}{\frac{1}{1+exp(w_0+w^T x)}}) = w_0 + w^T x$

(b) Consider a point that is correctly classified and distant from the decision boundary. Why would SVMs decision boundary be unaffected by this point, but the one learned by logistic regression be affected? [4 pts]

Answer: The hinge loss used by SVMs gives zero weight to these points while the log-loss used by logistic regression gives a little bit of weight to these points.

5 Maximum Likelihood [10 pts]

(a) You are in a casino in Las Vegas, playing slot machine games. You can win \$20 with machine A with probability of θ. Machine B has 4 times higher probability of winning, with just one forth of dividend (fair enough!). Suppose you played 10 times with either of machine A or B, and the result was as follows. What is the maximum likelihood estimation for θ? [5 pts]

Machine	Result	Machine	Result
A	Win	B	Win
A	Win	B	Lose
B	Lose	B	Lose
B	Lose	B	Win
B	Win	B	Lose

Answer: 1/8

(b) Uniform distribution [5 pts]

A uniform distribution in the range of $[\theta, \theta+1]$ is given by

$$p(x|\theta) = \begin{cases} 1 & \theta \le x \le \theta+1 \\ 0 & \text{otherwise} \end{cases}.$$

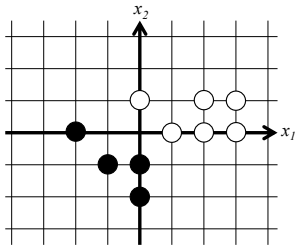
What is the maximum likelihood estimator of θ?

Answer: Any θ satisfying $\theta \le x^1, x^2, ..., x^n \le \theta + 1$

6 Classification [30 pts]

Suppose we have 10 training data points with a binary label (Black and White) on two-dimensional Euclidean space, where x_1 and x_2 are integers with $|x_1| \leq 3$ and $|x_2| \leq 3$, as given below. We will build several types of classifiers on this dataset to classify three unseen test points, namely $(0, 0)$, $(1, -1)$, and $(-1, 0)$.

x_1	x_2	y
0	1	White
0	-1	Black
0	-2	Black
1	0	White
2	0	White
2	1	White
3	0	White
3	1	White
-1	-1	Black
-2	0	Black



(a) Suppose we build a Bayes classifier under the following assumption. First, prior probability is determined by training data points. Second, $p(\mathbf{x}|\mathbf{y} = \text{White})$ is proportional to $\mathbf{x}_1+\mathbf{x}_2+3$, while $p(\mathbf{x}|\mathbf{y} = \text{Black})$ is proportional to $-\mathbf{x}_1 - \mathbf{x}_2 + 3$, where their (omitted) normalization constants are same. Classify the test points below with this Bayes classifier, and briefly explain why. If you are unable to classify a datum with the method, please indicate "N/C" (not classifiable). [6 pts]

- $(0, 0)$: White: 3 * 6/10 > 3 * 4/10
- $(1, -1)$: White: 3 * 6/10 > 3 * 4/10
- $(-1, 0)$: Black: 2 * 6/10 < 4 * 4/10

(b) Use Naive Bayes classifier to classify test points below, and briefly explain why. If you are unable to classify a datum with the method, please indicate "N/C" (not classifiable). [6 pts]

- $(0, 0)$: N/C: 1/6 * 1/2 * 6/10 = 1/2 * 1/4 * 4/10
- $(1, -1)$: N/C: 1/6 * 0 * 6/10 = 0 * 1/2 * 4/10
- $(-1, 0)$: Black: 0 * 1/2 * 6/10 < 1/4 * 1/4 * 4/10

(c) k-nearest neighbors classifier decides class of test points by majority vote of k nearest training points to the queried test point. When there are more than one set of k nearest neighbors, try all of them to output a label from each, and take one by majority vote among them. If it ties, mark it as N/C. Classify test points below with k = 1, 5, 7, and 10 respectively. (That is, answer four class assignments for each.) [6 pts]

- $(0, 0)$: White, Black, Black, White
- $(1, -1)$: N/C, Black, White, White
- $(-1, 0)$: Black, Black, Black, White

(d) Use the (hard margin) support vector machine to classify test points below, and briefly explain why. If you are unable to classify a datum with the method, please indicate "N/C" (not classifiable). [6 pts]

- $(0, 0)$: N/C (on decision boundary)
- $(1, -1)$: N/C (on decision boundary)
- $(-1, 0)$: Black

(e) How many support vectors do you have with the SVM in (d)? [6 pts]

Answer: 3