CS273A Final Exam

Introduction to Machine Learning: Winter 2020

Tuesday March 17th, 2020

Your name:

Yushang Lai

Your ID #(e.g., 123456789)

74082916

UCINetID (e.g.ucinetid@uci.edu)

yushanl2@uci.edu

- Due to the ongoing health emergency, this exam is **take home** and may be submitted in person or on Canvas (scanned).
- Total time is 2 hours 15 minutes for either in-person delivery (to the classroom) or submission to Canvas.
- READ THE EXAM FIRST and organize your time; don't spend too long on any one problem.
- Please put your name and ID **on every page**.
- Please **write clearly** and **show all your work**.
- If you need clarification on a problem, please post **privately** on Piazza and we will try to answer it.

Problems

Total, *(74 points.)*

This page is intentionally blank, use as you wish.

Name: Yushang Lai ID#: 74082916

Bayes Classifiers, *(12 points.)*

In this problem you will use Bayes Rule: $p(y|x) = p(x|y)p(y)/p(x)$ to perform classification. Suppose we observe some training data with two binary features x_1 , x_2 and a binary class y. After learning the model, you are also given some validation data.

Table 1: Training Data

x_1	x_2	y
~~0~~	~~0~~	~~0~~
~~0~~	~~1~~	~~0~~
0	1	1
0	1	1
1	0	1
1	0	1
~~1~~	~~1~~	~~0~~
~~1~~	~~1~~	~~0~~

Table 2: Validation Data

x_1	x_2	y
0	0	1
0	1	0
1	0	0
1	1	0

In the case of any ties, we will prefer to predict class 0.

(1) Give the predictions of a joint Bayes classifier on the validation data. What is the validation error rate? (Put final answers in boxes.) *(4 points.)*

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

Error Rate: $\boxed{3/4}$

(2) Give the required probabilities to define a **naïve** Bayes classifier. *(4 points.)*

$p(y=0) = 1/2$ $\quad$ $p(y=1) = 1/2$

$p(x_1=0|y=0) = 1/2$ $\quad$ $p(x_1=0|y=1) = 1/2$

$p(x_1=1|y=0) = 1/2$ $\quad$ $p(x_1=1|y=1) = 1/2$

$p(x_2=0|y=0) = 1/4$ $\quad$ $p(x_2=0|y=1) = 1/2$

$p(x_2=1|y=0) = 3/4$ $\quad$ $p(x_2=1|y=1) = 1/2$

(3) Give the predictions of a naïve Bayes classifier on the validation data. What is the validation error rate? (Put final answers in boxes.) *(4 points.)*

x_1	x_2	y
0	0	1
0	1	0
1	0	1
1	1	0

Error Rate: $\boxed{1/4}$

$p(x_1|y)\ p(x_2|y)\ p(y)$

$y=0$ $\quad$ $y=1$

$y=0$	$y=1$
1/8	1/4
3/8	1/4
1/8	1/4
3/8	1/4

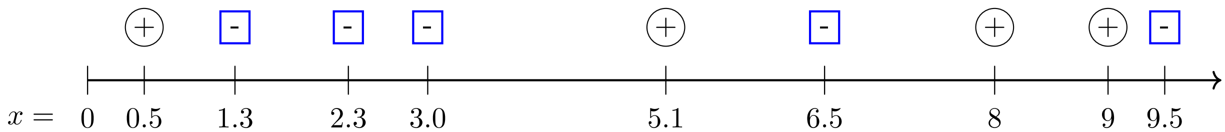
This page is intentionally blank, use as you wish.

Name: Yushang Lai ID#: 74082916

Cross-Validation, *(9 points.)*

Consider the following dataset with *nine* points shown below, for a binary classification task ($y = +, -$) with a scalar feature x. In case of ties, **prefer the negative class**. Put final answers in the box.

5−
4+



(1) Compute the **leave-one-out** cross-validation error rate of a 1-nearest neighbor classifier. *(3 points.)*

✗ ✗ ✓ ✓ ✗ ✗ ✓ ✗ ✗

$\boxed{2/3}$

(2) Compute the **leave-one-out** cross-validation error rate of a 3-nearest neighbor classifier. *(3 points.)*

✗ ✓ ✓ ✓ ✗ ✗ ✗ ✗ ✗

$\boxed{2/3}$

(3) Compute the **leave-one-out** cross-validation error rate of an 8-nearest neighbor classifier. *(3 points.)*

✗ ✓ ✓ ✓ ✗ ✓ ✗ ✗ ✓

$\boxed{4/9}$

This page is intentionally blank, use as you wish.

Name: Yushang Lai ID#: 74082916

Decision Trees, *(10 points.)*

Consider the table of measured data given at right. We will use a decision tree to predict the outcome y using the three features, $x_1, \ldots, x_3$. In the case of ties, we prefer to use the feature with the smaller index (x_1 over x_2, etc.) and prefer to predict class 1 over class 0. You may find the following values useful (although you may also leave logs unexpanded):

$\log_2(1) = 0$ $\log_2(2) = 1$ $\log_2(3) = 1.59$ $\log_2(4) = 2$
$\log_2(5) = 2.32$ $\log_2(6) = 2.59$ $\log_2(7) = 2.81$ $\log_2(8) = 3$

y=0 000 010 010 010 100
y=1 110 110 101

x_1	x_2	x_3	y
0	0	0	0
1	1	0	1
0	1	0	0
0	1	0	0
0	1	0	0
1	0	0	0
1	1	0	1
1	0	1	1

(1) What is the entropy of y? *(2 points.)*

$$H(y) = \frac{3}{8}\log\frac{8}{3} + \frac{5}{8}\log\frac{8}{5}$$
$$= \frac{3}{8}(\log 8 - \log 3) + \frac{5}{8}(\log 8 - \log 5) = 3 - \frac{3}{8}\log 3 - \frac{5}{8}\log 5$$

(2) Which variable would you split first? Justify your answer. *(2 points.)*

Split 1 first since x_1, x_3 both has entropy 0 part but 1 has smaller index

X_1

	=0	=1
y=0	000 010 010 010	100
y=1		110 110 101

X_2

y=0	000 100	010 010 010
y=1	101	110 110

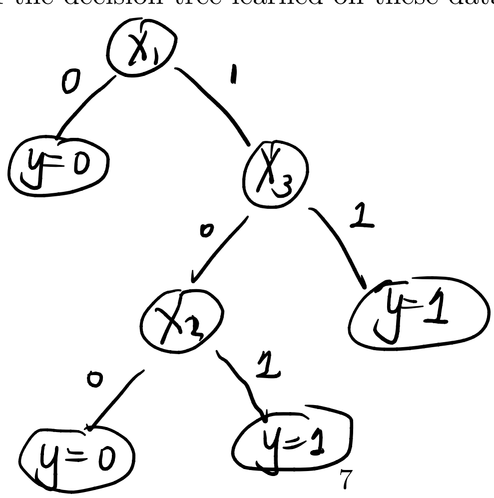
X_3

y=0	000 010 010 010 100	
y=1	110 110	101

(3) What is the information gain of the variable you selected in part (2)? *(3 points.)*

$$G(y_1) = H(y) - [\tfrac{1}{2} H(\phi) + \tfrac{1}{2} H(1/4)]$$
$$= H(y) - \tfrac{1}{2}[\tfrac{1}{4}\log 4 + \tfrac{3}{4}\log(4/3)]$$
$$= H(y) - \tfrac{1}{8}[2 + 3(\log 4 - \log 3)]$$
$$= H(y) - \tfrac{1}{8}[2 + 6 - 3\log 3] = H(y) - \tfrac{8}{8} + \tfrac{3}{8}\log 3 = 2 - \tfrac{5}{8}\log 5$$

(4) Draw the rest of the decision tree learned on these data. *(3 points.)*

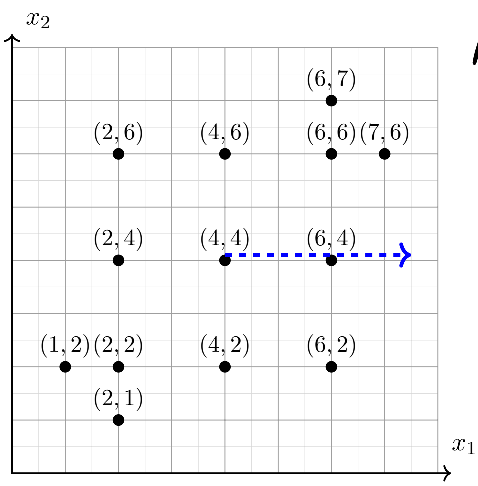


This page is intentionally blank, use as you wish.

Name: Yushang Lai ID#: 7408296

Dimensionality Reduction, ***(10 points.)***

(1) For the following points in two dimentions, consider performing linear dimensionality reduction along the given vector (dashed line). What is the reconstruction error, in MSE, when using this vector? *(4 points.)*



$$MSE = \frac{1}{13}\left((7-4)^2 + 4\times(6-4)^2 + 4\times(2-4)^2 + (1-4)^2 + (4-4)^2\times 3\right)$$

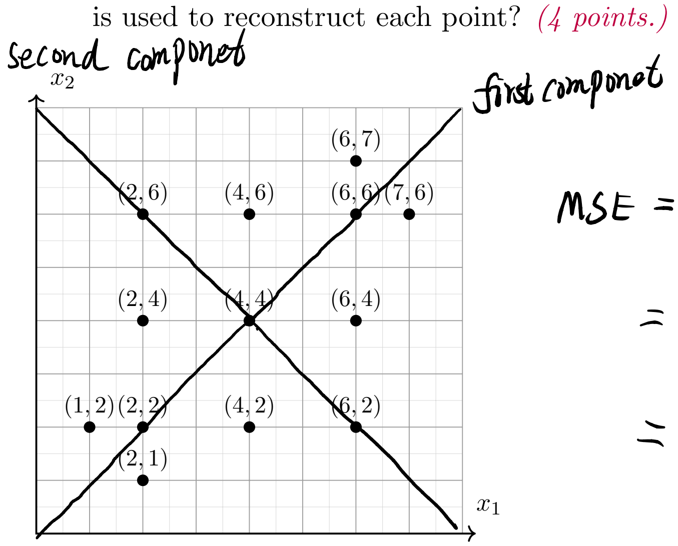
$$\frac{1}{13}\left(8\times 2^2 + 2\times 3^2 + 3\times 0^2\right)$$

$$= \frac{1}{13}(32+18)$$

$$= \frac{50}{13}$$

(2) On the figure below, draw the directions of the first two principal components. *(2 points.)*

(3) What is the reconstruction error (MSE) of these points when only the first principal component is used to reconstruct each point? *(4 points.)*



$$MSE = \left(2\times(2\sqrt{2})^2 + 4\times\sqrt{2}^2 + 4\times\left(\frac{\sqrt{2}}{2}\right)^2 + 3\times 0^2\right)\times\frac{1}{13}$$

$$= \frac{1}{13}(16+8+2)$$

$$= \frac{26}{13} = 2$$

This page is intentionally blank, use as you wish.

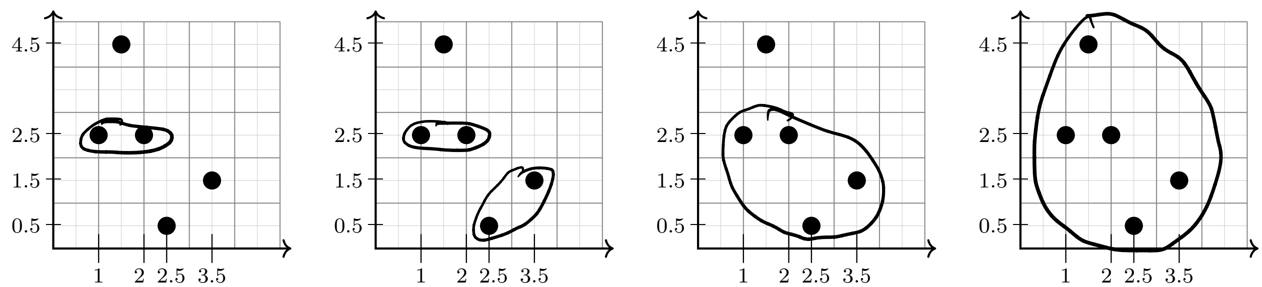
Name: Yushang Lai ID#: 74082916

Hierarchical Clustering, *(11 points.)*

Consider the two-dimensional data points plotted in each panel. In this problem, we will cluster these data.

Linkage

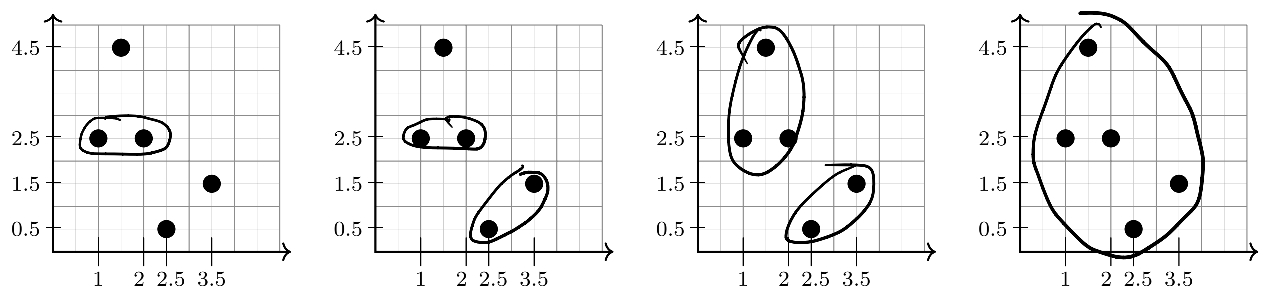
(a) Execute the hierarchical agglomerative clustering (linkage) algorithm on these data points, using "single linkage" (minimum distance) for the cluster scores. Stop when the algorithm would terminate, or after 4 steps, whichever is first. Show each step separately in a panel. *(4 points.)*



2^2+3^2 1^2+4^2

13 17

(b) Now repeat your agglomerative clustering algorithm, this time using "complete linkage" (maximum distance) for the cluster scores. Stop when the algorithm would terminate, or after 4 steps, whichever is first. Show each step separately in a panel. *(4 points.)*



3^2+4^2 1^2+4^2

(c) What is the (big-O) computational complexity of the hierarchical agglomerative clustering algorithm? Justify your answer **briefly** in 1-2 sentences. *(3 points.)*

suppose has m points

calculate all distance $O(m^2)$

sorted $O(m^2 \log m^2) = O(m^2 \log m)$

Each iteration (m times) i

merge closet cluster pair $O(1)$

update (m-i) cluster distance in sorted $O((m-i)\log m)$

$\Rightarrow$ total $O(m^2 \log m)$

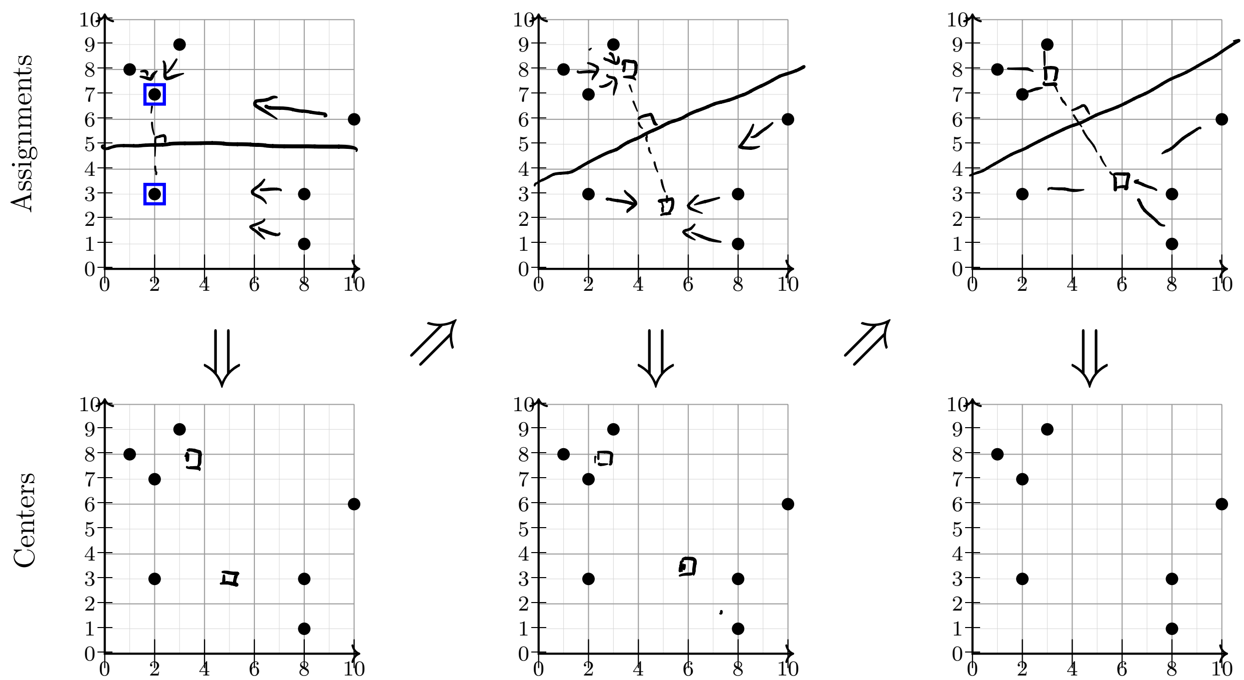
This page is intentionally blank, use as you wish.

Name: Yushang Lai ID#: 74082916

K-Means Clustering, *(10 points.)*

Consider the 2-D data points plotted in each panel. In this problem, we will cluster these data using the k-means algorithm, where each panel is used to show a single step of the algorithm.

(a) Starting from the two cluster centers indicated by squares, perform k-means clustering on the data. In the top panels, indicate the assignment of the data, and then in the panel below show the new cluster centers, so each **pair** of panels shows an iteration of k-means. Stop when converged, or after 6 steps (3 iterations), whichever is first. It may be helpful to recall from our nearest neighbor classifier that the set of points nearer to A than B is separated by a line. *(6 points.)*



converges

stop

(b) Write down the cost function optimized by the k-means algorithm, in terms of the data locations $x^{(i)}$, cluster centers μ_c, and cluster assignments $z^{(i)}$. *(2 points.)*

$$J = \sum_{i=1}^{m} \| X^{(i)} - \mu_{z^{(i)}} \|^2$$

$\| \cdot \|^2$ is the Euclidean distance

(c) What is the (big-O) computational complexity of each iteration of k-means (naïve computation), in terms of the data size m and number of clusters k? *(2 points.)*

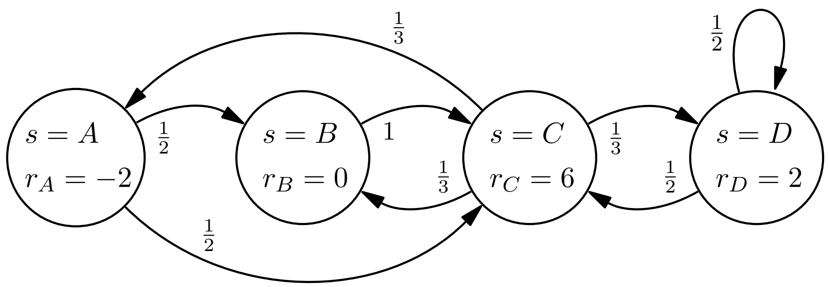
$O(km)$

This page is intentionally blank, use as you wish.

Name: Yushang Lai ID#: 74082916

Markov Processes, *(12 points.)*

Consider the Markov reward process model shown here:



$\Pr[A \to B] = 0.5$
$\Pr[A \to C] = 0.5$
$\Pr[B \to C] = 1.0$
$\Pr[C \to A] = 0.33$
$\Pr[C \to B] = 0.33$
$\Pr[C \to D] = 0.33$
$\Pr[D \to C] = 0.5$
$\Pr[D \to D] = 0.5$

where the transition probabilities are shown next to each arc and at right, and the rewards r_s associated with each state s are shown inside the circles. We will use dynamic programming to (start) computing the expected discounted sum of rewards. Assume a future discounting factor of $\gamma = \frac{1}{2}$.

(1) Compute $J^1(s)$, the expected discounted sum of rewards for state sequences of length 1 (e.g., $[A]$) starting in each state s. *(4 points.)*

$J^1(A) =$	$J^1(B) =$	$J^1(C) =$	$J^1(D) =$
-2	0	6	2

(2) Compute $J^2(s)$, the expected discounted sum of rewards for state sequences of length 2 (e.g., $[C \to B]$) starting in each state s. *(4 points.)*

$J^2(A) =$	$J^2(B) =$	$J^2(C) =$	$J^2(D) =$
$-1/2$	3	6	4

$-2 + \frac{1}{2} \times \frac{1}{2} \times 0 + \frac{1}{2} \times \frac{1}{2} \times 6 = -2 + \frac{3}{2}$

$0 + \frac{1}{2} \times 1 \times 6$

$6 + \frac{1}{2} \times \frac{1}{3} \times (-2)$ $\frac{1}{2} \times \frac{1}{3} \times 0$ $\frac{1}{2} \times \frac{1}{3} \times 2$

$2 + \frac{1}{2} \times \frac{1}{2} \times 2 + \frac{1}{2} \times \frac{1}{2} \times 6 =$

(3) Compute $J^3(s)$, the expected discounted sum of rewards for state sequences of length 3 (e.g., $[D \to C \to A]$) starting in each state s. *(4 points.)*

$J^3(A) =$	$J^3(B) =$	$J^3(C) =$	$J^3(D) =$
$5/2$	3	$6\frac{13}{12}$	$9/2$

$-2 + \frac{1}{2} \times \frac{1}{2} \times 3 + \frac{1}{2} \times \frac{1}{2} \times 6$

$-2 + \frac{9}{2}$

$0 + \frac{1}{2} \times 1 \times 6$

$6 + \frac{1}{2} \times \frac{1}{3} \times (-\frac{1}{2} + 3 + 4)$

$6 + \frac{1}{6} \times \frac{13}{2}$

$6 + \frac{13}{12}$

$2 + \frac{1}{2} \times \frac{1}{2} \times 4 + \frac{1}{2} \times \frac{1}{2} \times 6$

$2 + \frac{10}{4}$

This page is intentionally blank, use as you wish.

This page is intentionally blank, use as you wish.

This page is intentionally blank, use as you wish.