

CS273A Final Exam  
Introduction to Machine Learning: Fall 2017  
Thursday December 14th, 2017

Your name:

Row/Seat Number:

Your ID #(e.g., 123456789)

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

- Please put your name and UCINetID **on every page**.
- Total time is 1 hour and 50 minutes. READ THE EXAM FIRST and organize your time; don't spend too long on any one problem.
- Please **write clearly** and **show all your work**.
- If you need clarification on a problem, please raise your hand and wait for the instructor or TA to come over.
- You may use **one** sheet of your own handwritten notes for reference, and a calculator. .
- Turn in your notes and any scratch paper with your exam.

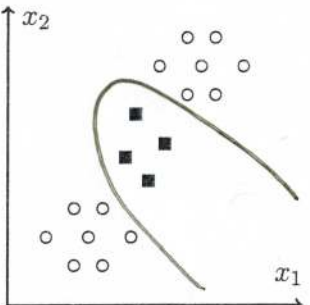
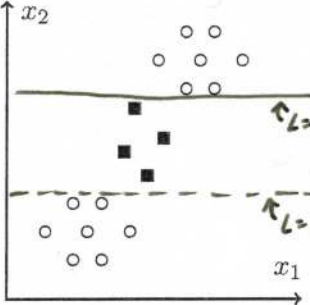
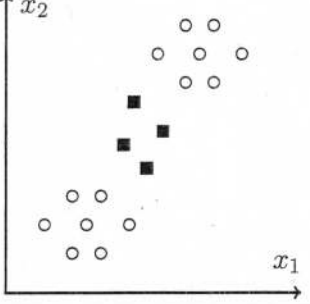
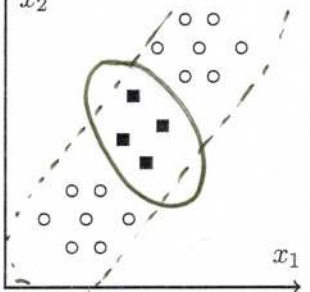
## Problems

1	Separability, <i>(8 points.)</i>	3
2	Dimensionality Reduction, <i>(8 points.)</i>	5
3	Decision Trees, <i>(8 points.)</i>	9
4	K-Means Clustering, <i>(10 points.)</i>	13
5	Multiple Choice, <i>(10 points.)</i>	17
6	Gradient Descent, <i>(10 points.)</i>	19
7	Agglomerative Clustering, <i>(12 points.)</i>	23
8	Reinforcement Learning, <i>(14 points.)</i>	27

Total, *(80 points.)*

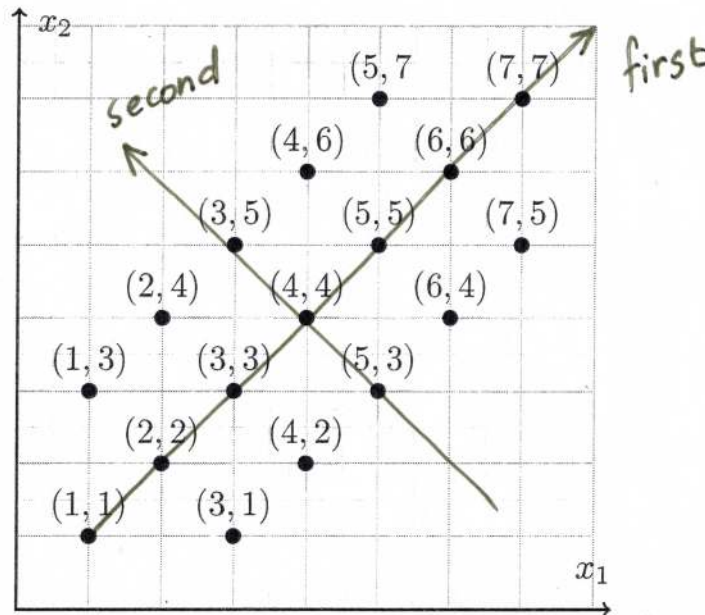
**Problem 1 Separability, (8 points.)**

For each of the following examples of training data and classifiers, state whether there exists a set of parameters that can separate the data and justify your answer briefly (~1 sentence+plotting).

	<p>Perceptron with quadratic features, <math>[x_1 \ x_2 \ x_1x_2 \ x_1^2 \ x_2^2]</math>: (2 points.)</p> <p><i>Yes</i></p>
	<p>Depth two decision tree: (2 points.)</p> <p><i>Yes</i></p>
	<p>Gaussian Bayes Classifier with equal covariance: (2 points.)</p> <p><i>No, since it will be a linear classifier</i></p>
	<p>Gaussian Bayes Classifier with unequal covariance: (2 points.)</p> <p><i>Yes, by having a sharply peaked Gaussian on the squares.</i></p>

## Problem 2 Dimensionality Reduction, (8 points.)

Consider the following set of points in two dimensions.



- (1) On the above figure, draw directions of the first two principal components. (2 points.)
- (2) What is the reconstruction error (MSE) of these points when both the principal vectors are used to reconstruct each point. (2 points.)

0, perfect reconstruction

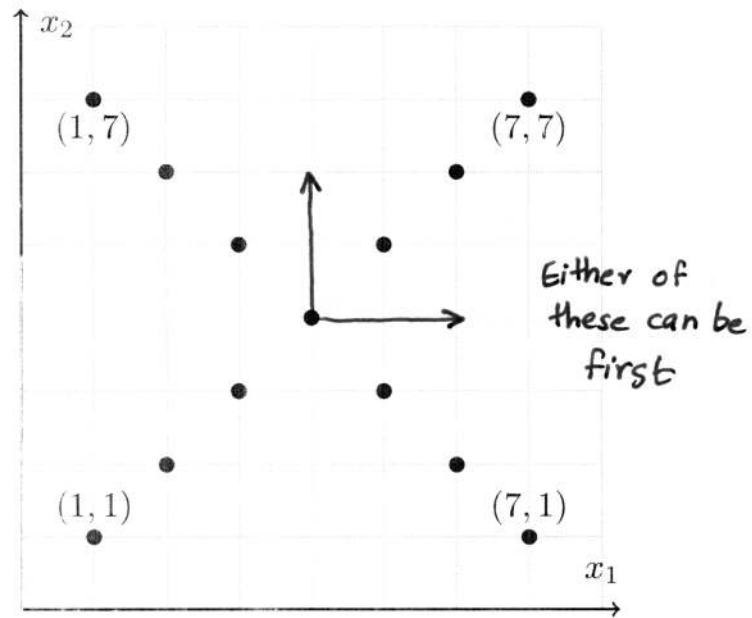
- (3) What is the reconstruction error (MSE) of these points when only the first principal vector is used to reconstruct each point. (2 points.)

$\sqrt{2}$  error for every off-diagonal points.

$$MSE = \frac{1}{17} \times 10 \times (\sqrt{2})^2 = \frac{20}{17} = \underline{1.176}$$

Continued on the next page.

- (4) For the following points, draw the directions of the first two principal vectors. (2 points.)



**Problem 3 Decision Trees, (8 points.)**

Consider the table of measured data given on the right. We will use a decision tree to predict the outcome  $y$  using three features,  $x_1, x_2, x_3$ . 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 - over class +. You may find the following values useful:

$x_1$	$x_2$	$x_3$	$y$
0	0	0	+
0	0	1	-
0	1	0	-
0	1	1	+
1	0	0	-

$$\log_2(1) = 0 \quad \log_2(2) = 1 \quad \log_2(3) = 1.59 \quad \log_2(4) = 2$$

$$\log_2(5) = 2.32 \quad \log_2(6) = 2.59 \quad \log_2(7) = 2.81 \quad \log_2(8) = 3$$

- (1) Which attribute has the highest information gain? Justify your answer. (3 points.)

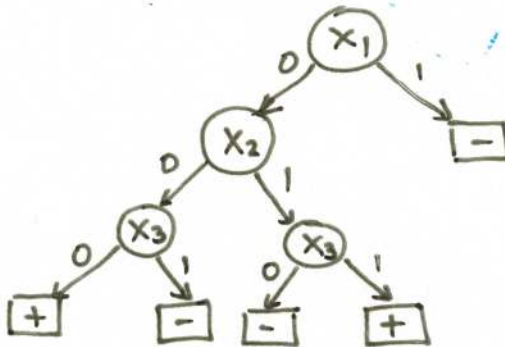
$$H(y|x_1) = \frac{4}{5} H\left(\frac{1}{2}\right) + \frac{1}{5} H(1) = \frac{4}{5} = 0.8$$

$$H(y|x_2) = \frac{3}{5} H\left(\frac{2}{3}\right) + \frac{2}{5} H\left(\frac{1}{2}\right)$$

$$= \frac{3}{5} \left( -\frac{2}{3} \log \frac{2}{3} - \frac{1}{3} \log \frac{1}{3} \right) + \frac{2}{5} = \frac{3}{5} \left( \log 3 - \frac{2}{3} \right) + \frac{2}{5}$$

$$H(y|x_3) = \frac{3}{5} H\left(\frac{2}{3}\right) + \frac{2}{5} H\left(\frac{1}{2}\right) = 0.95 \quad \therefore \boxed{x_1}$$

- (2) Based on this choice, build the complete decision tree learned on this data. (3 points.)



Continued on next page.

Name:

UCINETID:

Suppose we also have the following validation data.

$x_1$	$x_2$	$x_3$	$y$
1	1	0	+
0	1	1	-
1	1	1	-

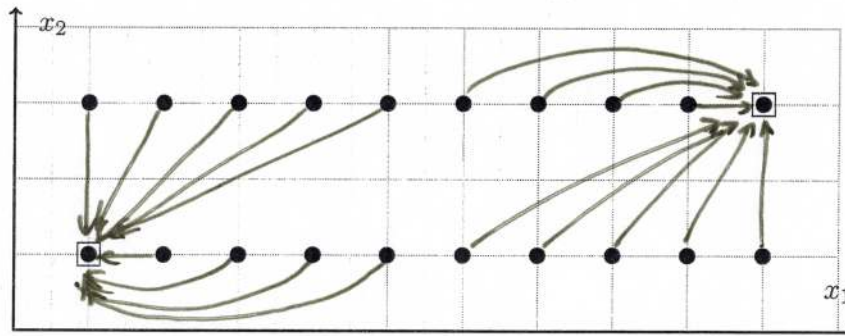
- (3) Is it possible to construct a decision tree that will have perfect accuracy both on train and validation? If yes, provide the tree. If not, explain why not. (2 points.)

No, because two instances have same features  
but different labels  $\uparrow$  011

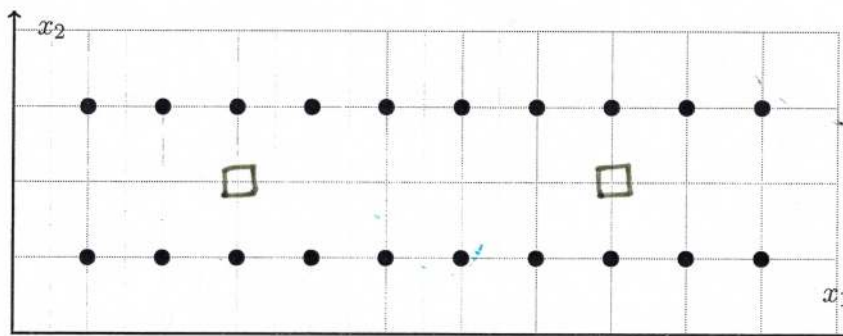


**Problem 4 K-Means Clustering, (10 points.)**

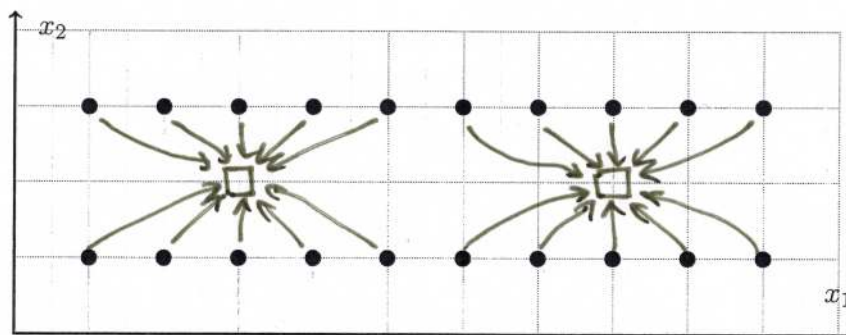
- (1) For the following set of 20 points in two dimensions, and the two initial cluster centers (as squares), indicate on the figure what the cluster assignments in the first step of K-Means will look like ( $K = 2$ ). (2 points.)



- (2) Do another step of K-Means, to update the cluster centers based on the previous assignment. Plot the new cluster centers. (2 points.)



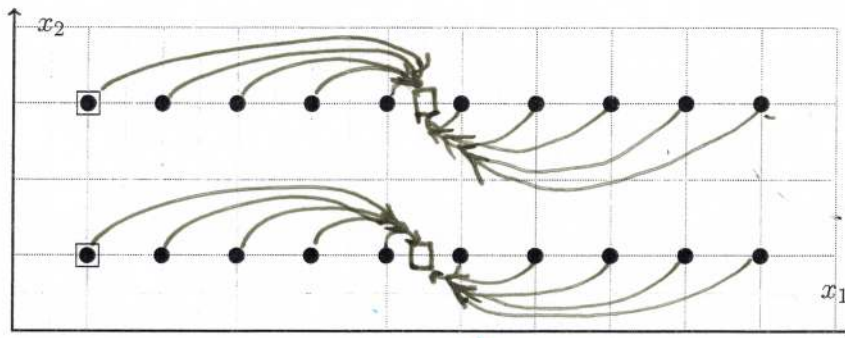
- (3) Compute the assignment of the points to clusters, using these new cluster centers. (2 points.)



- (4) Do you think the algorithm has converged? Why do you think so? (2 points.)

Yes, since the assignments did not change

- (5) Draw the final cluster assignment and the cluster centers when K-Means is initialized with the following initial clusters (squares). (2 points.)





**Problem 5 Multiple Choice, (10 points.)**

Here, assume that we have  $m$  data points  $y^{(i)}, x^{(i)}, i = 1 \dots m$ , each with  $n$  features,  $x^{(i)} = [x_1^{(i)} \dots x_n^{(i)}]$ . For each of the choices below, will it likely increase, decrease, or have no effect on overfitting (circle your choice)? If you think it is equally likely to go either way, pick *No Effect*.

- |   |   |               |                 |                  |
|---|---|---------------|-----------------|------------------|
| 1 | Creating an ensemble of classifiers by Bagging      | <u>Reduce</u> | Increase        | No Effect        |
| 2 | Creating an ensemble of classifiers by Boosting     | Reduce        | <u>Increase</u> | No Effect        |
| 3 | Reducing the number of features using PCA           | <u>Reduce</u> | Increase        | No Effect        |
| 4 | Evaluating AUC instead of accuracy                  | Reduce        | Increase        | <u>No Effect</u> |
| 5 | Training on a single bootstrap of the training data | Reduce        | <u>Increase</u> | No Effect        |

Suppose you are working with a dataset for which the number of features is much higher than the number of data points, i.e.  $n \gg m$ . For each of the questions below, circle the algorithm that you think is better in this scenario in terms of time complexity. If both are equally fast, pick *Either*.

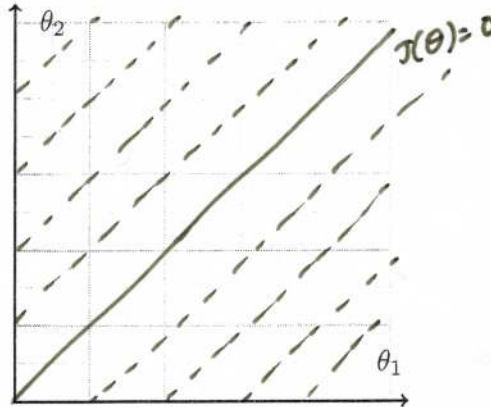
- |    |                           |                             |               |
|----|---------------------------|-----------------------------|---------------|
| 6  | <u>Naive Bayes</u>        | Joint Bayes                 | Either        |
| 7  | Single-linkage Clustering | Complete-linkage Clustering | <u>Either</u> |
| 8  | Primal-form SVM           | <u>Dual-form Kernel-SVM</u> | Either        |
| 9  | Quadratic regression      | <u>Decision Trees</u>       | Either        |
| 10 | <u>K-Means Clustering</u> | Mixture of Gaussians        | Either        |

**Problem 6 Gradient Descent, (10 points.)**

Here's a new loss function for you to consider for a problem with two parameters:

$$J(\theta) = (\theta_1 - \theta_2)^2$$

- (1) Draw the contour plot of the cost function  $J$  in terms of  $\theta_1$  and  $\theta_2$ . Contour plots consists of lines/curves for different values of the cost function, i.e. try  $J(\theta) = c, c = 0, 1, 4$ . (3 points.)



- (2) Derive the gradient equations for the cost function, i.e.  $\frac{\partial}{\partial \theta_1} J(\theta)$  and  $\frac{\partial}{\partial \theta_2} J(\theta)$ . (2 points.)

$$\frac{\partial}{\partial \theta_1} (\theta_1 - \theta_2)^2 = 2(\theta_1 - \theta_2)$$

$$\frac{\partial}{\partial \theta_2} (\theta_1 - \theta_2)^2 = 2(\theta_1 - \theta_2)(-1) = 2(\theta_2 - \theta_1)$$

Continued on next page

- (3) Starting with  $\theta_1 = 0$  and  $\theta_2 = 1$ , compute the gradient, and for learning rate  $\alpha = 0.5$ , perform two steps of gradient descent. (3 points.)

$$t=1 \quad \begin{aligned} \theta_1 &= 0 - \frac{1}{2} \cdot 2 \cdot (0-1) = 1 \\ \theta_2 &= 1 - \frac{1}{2} \cdot 2 \cdot (1-0) = 0 \end{aligned}$$

$$t=2 \quad \begin{aligned} \theta_1 &= 1 - \frac{1}{2} \cdot 2 \cdot (1-0) = 0 \\ \theta_2 &= 0 - \frac{1}{2} \cdot 2 \cdot (0-1) = 1 \end{aligned}$$

- (4) Do you think the above algorithm will converge? If not, what do you think you need to do so that it converges? (2 points.)

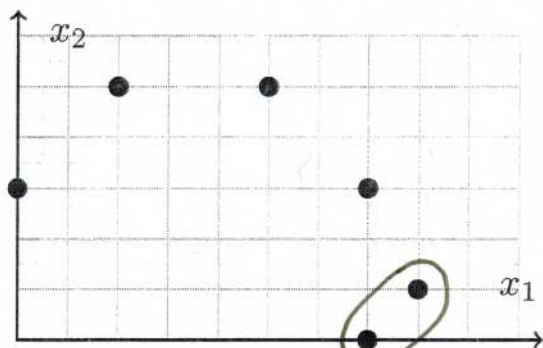
No, reduce learning rate

**Problem 7 Agglomerative Clustering, (12 points.)**

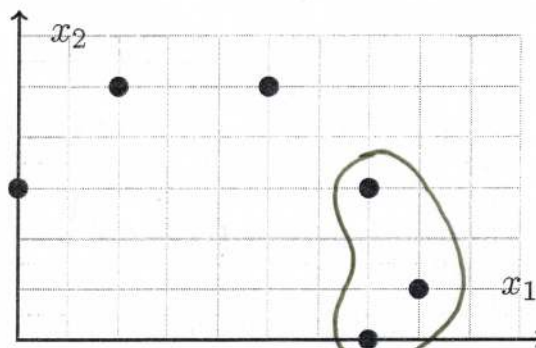
You are given *six* points on a two-dimensional space. For this configuration, you will be showing the steps of both *single-linkage* and *complete-linkage* clustering, starting with all singletons.

You *might* need to use the fact that  $\sqrt{2} = 1.41$ ,  $\sqrt{3} = 1.73$ ,  $\sqrt{5} = 2.24$ , and  $\sqrt{8} = 2.83$ .

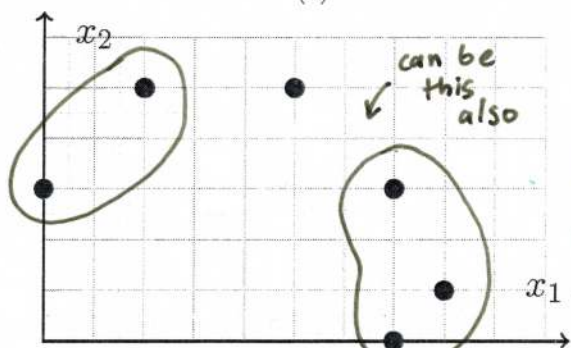
- (1) **Single-linkage Clustering:** Show the first 6 steps (or till termination, whichever sooner) of single-linkage agglomerative clustering. In case of ties, pick arbitrarily. (5 points.)



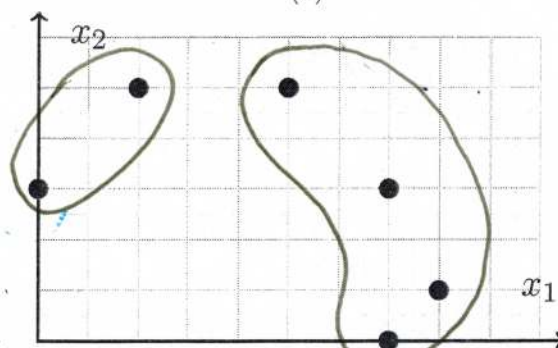
(a)



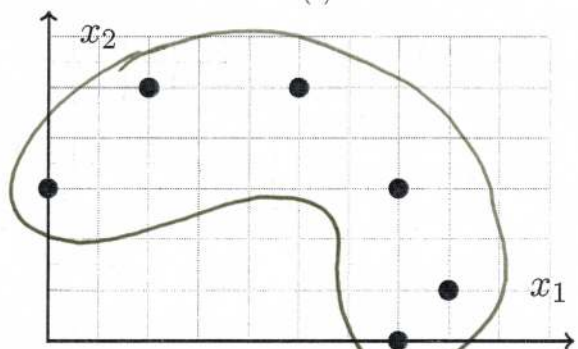
(b)



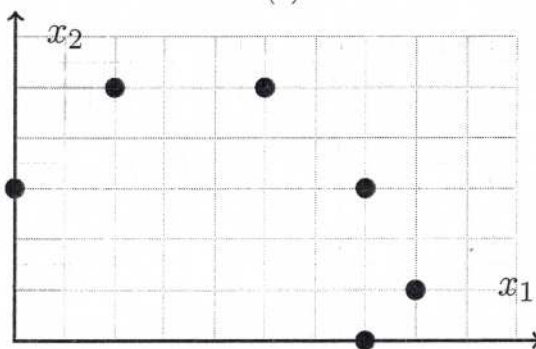
(c)



(d)



(e)

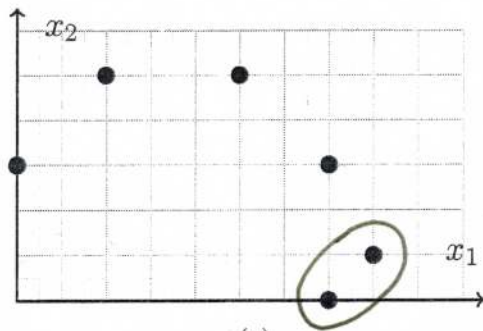


(f)

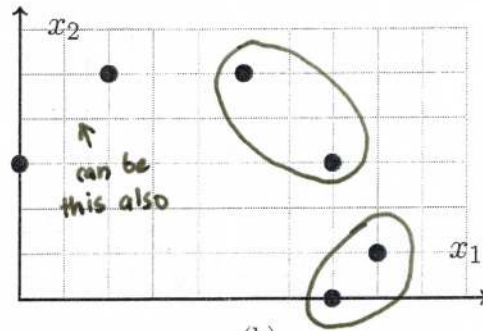
Continued on next page.



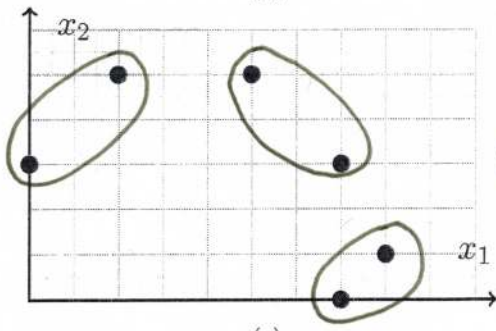
- (2) **Complete-linkage Clustering:** Show the first 6 steps (or till termination, whichever sooner) of complete-linkage clustering. In case of ties, pick arbitrarily. (5 points.)



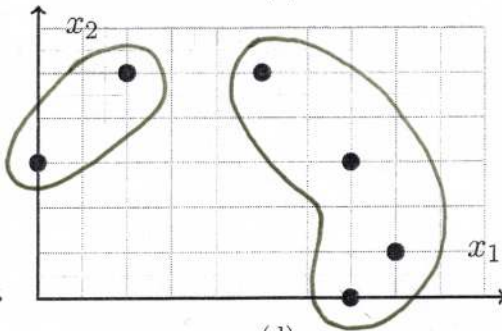
(a)



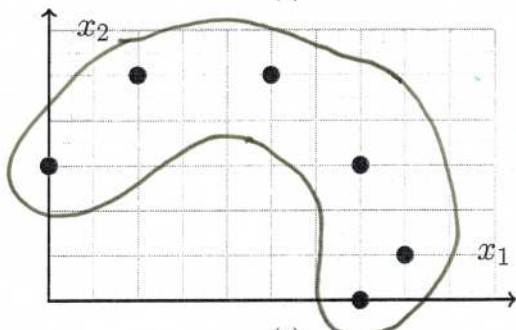
(b)



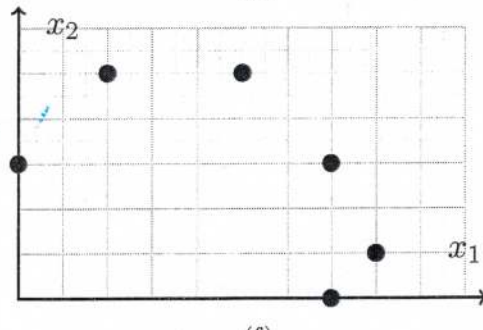
(c)



(d)



(e)



(f)

- (3) Suppose you have a collection of exam submissions that you want to cluster hierarchically to identify groups of students that have submitted very similar exams to each other. Given that you're using the number of words that are common between submissions as the similarity between them, which of single- or complete-linkage would you use, and why? (2 points.)

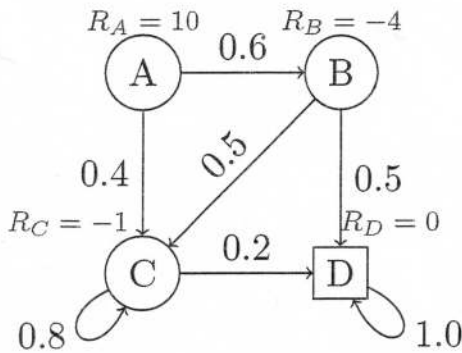
Complete.

Single-linkage can put exams in a cluster that do not share ANY words.

**Problem 8 Reinforcement Learning, (14 points.)**

In this question, we will consider a simple environment to compute various properties such as the value function and the optimal policy.

- (1) Given the following Markov random process consisting of four states, with the transition probability and the reward achieved when **leaving** the state shown in the figure, compute the value function with  $\gamma = 0$ . (2 points.)



$$V_A = 10$$

$$V_B = -4$$

$$V_C = -1$$

$$V_D = 0$$

- (2) For the same MRP, compute the value function for all the states with  $\gamma = 0.5$ . Show that the computed value function is correct by verifying it for any two states. (6 points.)

**Hint:** Compute this in the reverse order, i.e. D, C, B, and A.

$$V_A = 8.34$$

$$V_B = -4.417$$

$$V_C = -1.667$$

$$V_D = 0$$

$$V(s) = R_s + \gamma \sum_{s'} P_{ss'} V(s')$$

$$V_D = 0 + \frac{1}{2}(1)V_D = \frac{1}{2}V_D \quad \therefore \underline{V_D = 0}$$

$$V_C = -1 + \frac{1}{2}\left(\frac{1}{5}(0) + \frac{4}{5}V_C\right) = -1 + \frac{2}{5}V_C$$

$$\Rightarrow \frac{3}{5}V_C = -1 \Rightarrow \underline{V_C = -\frac{5}{3}}$$

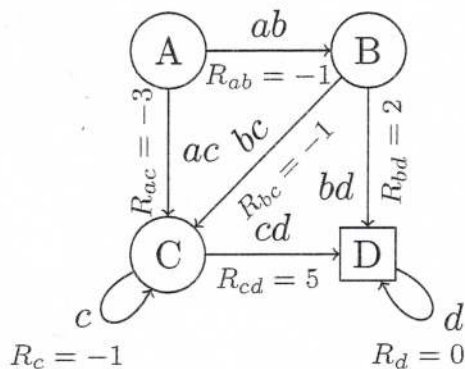
$$V_B = -4 + \frac{1}{2}\left(\frac{1}{2}(0) + \frac{1}{2}\left(-\frac{5}{3}\right)\right) = -4 - \frac{5}{12}$$

$$V_A = 10 + \frac{1}{2}\left(\frac{2}{5}\left(-\frac{5}{3}\right) + \frac{3}{5}\left(-4 - \frac{5}{12}\right)\right) = 10 - \frac{1}{3} - \frac{12}{10} - \frac{1}{8} = 10 - 0.33 - 1.2 - 0.125$$

$$\approx \underline{8.34}$$

Continued on next page.

- (3) Now consider the Markov Decision Process (MDP) below, i.e. the MRP with actions. We have provided what we think is the optimal value function for  $\gamma = 0.8$ . Show that it is indeed the case by verifying it for all the four states, or pointing out any errors. (4 points.)



$$V_A = 1.4$$

$$V_B = 3$$

$$V_C = 5$$

$$V_D = 0$$

$$V_D \Rightarrow 0 = \max(0 + 0.8(0)) = 0 \quad \checkmark$$

$$V_C \Rightarrow 5 = \max(-1 + 0.8(5), 5 + 0.8(0)) = 5 \quad \checkmark$$

$$V_B \Rightarrow 3 = \max(-1 + 0.8(5), 2 + 0.8(0)) = 3 \quad \checkmark$$

$$V_A \Rightarrow 1.4 = \max(-1 + 0.8(3), -3 + 0.8(5)) = \max\left(\frac{7}{5}, 1\right) = \frac{7}{5} \quad \checkmark$$

- (4) What is the optimal policy for this MDP? Describe it as the action for each state. (2 points.)

$$\pi(D) = d$$

$$\pi(C) = cd$$

$$\pi(B) = bc$$

$$\pi(A) = ab$$