SENG 474, CSC 578D: Data Mining: Fall 2018
Assignment 3

1. (9 pt) Consider the dataset in Fig 1, with points belonging to two classes, blue squares and red circles.

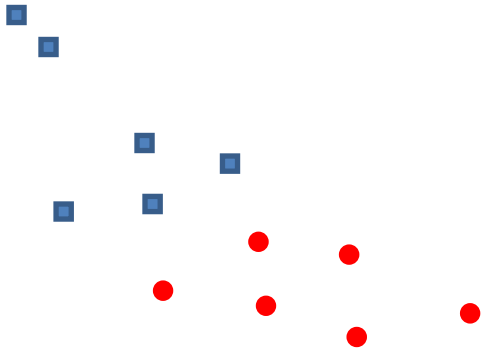

Fig. 1

(a) [1 pt] Draw (approximately) the SVM line separator.
(b) [1 pt] Suppose we find (1/2)*$\mathbf{w}^2$ to be 2 in the SVM optimization. What is the margin, i.e. the distance of closest points to the line?

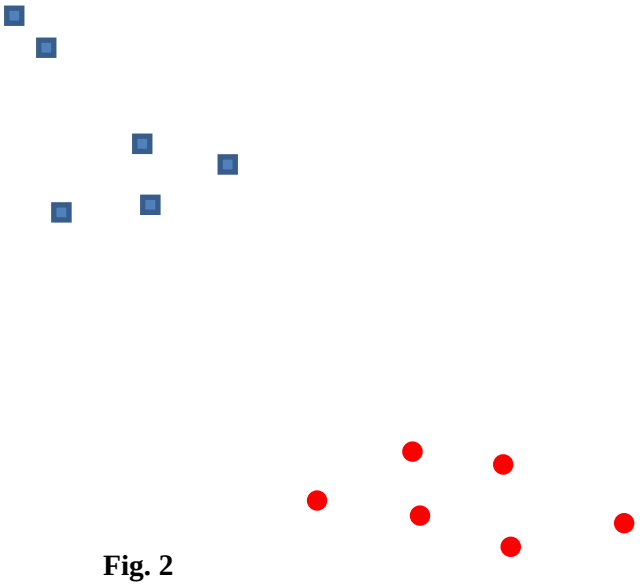

Fig. 2

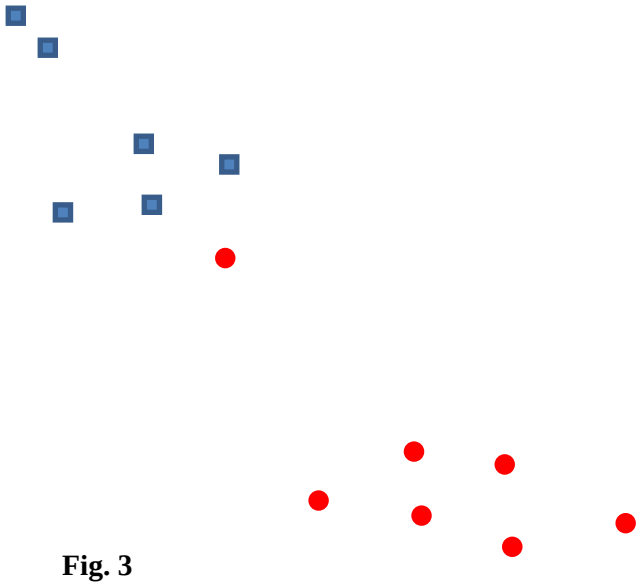

Fig. 3

(c) [1 pt] Now consider the dataset in Fig 2 (the red points are shifted below). Will (1/2)*$\mathbf{w}^2$ be smaller or greater than previously? Explain.
(d) [2 pt] Using a ruler, and the fact that (1/2)*$\mathbf{w}^2$ was 2 previously, find (approximately) the magnitude of the new line coefficient vector, **w**'.
(e) [3 pt] Consider the dataset in Fig 3 (with one additional red circle quite close to the blue squares). Assuming optimization using slack variables and C=1, draw a line that does not perfectly separate the points, but which is nonetheless better than the line that perfectly separates the points. (Draw it in the figure, and explain why).
(f) [1 pt] Why would we rather prefer the line in (e) to the line that perfectly separates the points?

2. (4 pt) Consider the task of building a classifier from random data, where the attribute values are generated randomly irrespective of the class labels. Assume the data set contains records from two classes, "+" and "−." Half of the data set is used for training while the remaining half is used for testing.

(a) (1 pt) Suppose there are an equal number of positive and negative records in the data and a classifier predicts every test record to be positive. What is the expected error rate of the classifier on the test data?

(b) (1 pt) Repeat the previous analysis assuming that the classifier predicts each test record to be positive class with probability 0.8 and negative class with probability 0.2.

(c) (1 pt) Suppose two-thirds of the data belong to the positive class and the remaining one-third belong to the negative class. What is the expected error of a classifier that predicts every test record to be positive?

(d) (1 pt) Repeat the previous analysis assuming that the classifier predicts each test record to be positive class with probability 2/3 and negative class with probability 1/3.

3. (5 pt) You are asked to evaluate the performance of two classifiers, A and B. The following table shows the ranking obtained by applying the classifiers to a test set of 10 instances.

Instance	True Class	Classifier A	Classifier B
1	+	0.73	0.61
2	+	0.69	0.03
3	-	0.44	0.68
4	-	0.55	0.31
5	+	0.67	0.45
6	+	0.47	0.09
7	-	0.08	0.38
8	-	0.15	0.05
9	+	0.45	0.01
10	-	0.35	0.04

a. (2 pt) Plot the ROC graphs for both A and B. (You should plot them on the same space.) Which classifier do you think is better? Explain your reasons.

b. (1 pt) For classifier A, suppose you choose the cutoff threshold to be t = 0.5. In other words, any test instances whose ranking is greater than t will be classified as a positive example. Compute the precision, recall, and F-measure for the classifier at this threshold value.

c. (1 pt) Repeat the analysis for part (b) using the same cutoff threshold on classifier B. Compare the F-measure results for both classifiers. Which classifier is better? Are the results consistent with what you expect from the ROC curve?

d. (1 pt) Show, in the ROC graph for A, the points corresponding to thresholds t = 0.5 and t = 0.1. Tell one application [from real life] for which you would set t = 0.5 and another application for which you would set t = 0.1.

4. (4 points) The Apriori algorithm uses a generate-and-count strategy for deriving frequent itemsets. Candidate itemsets of size k + 1 are created by joining a pair of frequent itemsets of size k (this is known as the candidate generation step). A candidate is discarded if any one of its subsets is found to be infrequent during the candidate pruning step. Suppose the Apriori algorithm is applied to the data set shown in the table with minsup=30%, i.e., any itemset occurring in less than 3 transactions is considered to be infrequent.

Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

(a) Draw an itemset lattice representing the data set. Label each node in the lattice with the following letter(s):
N: If the itemset is not considered to be a candidate itemset by the Apriori algorithm. There are two reasons for an itemset not to be considered as a candidate itemset: (1) it is not generated at all during the candidate generation step, or (2) it is generated during the candidate generation step but is subsequently removed during the candidate pruning step because one of its subsets is found to be infrequent.
F: If the candidate itemset is found to be frequent by the Apriori algorithm.
I: If the candidate itemset is found to be infrequent after support counting.

(b) What is the percentage of frequent itemsets (with respect to all itemsets in the lattice)?

(c) What is the pruning ratio of the Apriori algorithm on this data set?
(Pruning ratio is defined as the percentage of itemsets not considered to be a candidate because (1) they are not generated during candidate generation or (2) they are pruned during the candidate pruning step.)

(d) What is the false alarm rate (i.e, percentage of candidate itemsets that are found to be infrequent after performing support counting)?

5. (4 points) Using the data in the above table, build an FP-Tree, and then mine the frequent itemsets using FP-Growth with minsup=30%.

6. (6 points) Use the following similarity matrix to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.
("Single link clustering" means clustering using MIN, while "Complete link clustering" means clustering using MAX)

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

7. (8 points) Consider the following set of users and movies they have rated.

'Lisa Rose': {'Lady in the Water': 2.5, 'Snakes on a Plane': 3.5, 'Just my Luck': 3.0, 'Superman Returns': 3.5, 'You, Me and Dupree': 2.5, 'The Night Listener': 3.0}	'Gene Seymour':{'Lady in the Water': 3.0, 'Snakes on a Plane': 3.5, 'Just my Luck': 1.5, 'Superman Returns': 5.0, 'The Night Listener': 3.0, 'You, Me and Dupree': 3.5}
'Michael Phillips':{'Lady in the Water': 2.5, 'Snakes on a Plane': 3.0, 'Superman Returns': 3.5, 'The Night Listener': 4.0}	'Claudia Puig': {'Snakes on a Plane': 3.5, 'Just my Luck': 3.0, 'The Night Listener': 4.5, 'Superman Returns': 4.0, 'You, Me and Dupree': 2.5}
'Mick LaSalle':{'Lady in the Water': 3.0, 'Snakes on a Plane': 4.0, 'Just my Luck': 2.0, 'Superman Returns': 3.0, 'The Night Listener': 3.0, 'You, Me and Dupree': 2.0}	'Jack Matthews':{'Lady in the Water': 3.0, 'Snakes on a Plane': 4.0, 'Superman Returns': 5.0, 'The Night Listener': 3.0, 'You, Me and Dupree': 3.5}
'Toby': {'Snakes on a Plane': 4.5, 'Superman Returns': 4.0, 'You, Me and Dupree': 1.0}	

(a) **(4pt)** Suppose we build a recommender system following the user-user similarities approach with Pearson correlation as a similarity measure. What will be the rating prediction for user Michael Phillips, for movie "You, Me and Dupree"? Give the details of your computation.

(b) **(4pt)** If we use the user-bias, item-bias approach to recommendation (Netflix competition), what will b_r (short for $b_{\text{lisa rose}}$) be after the first pass over the data? Set $\lambda_1=\lambda_2=\gamma=0.1$, and start with zero bias values.