

**Question 1 (Ensemble Learning) [2 points]**

Alan recently came across the concept of ensemble methods in the 5525 class and decided to apply it to stock market prediction. To predict whether the stock market will rise or fall on a given day, he decided to flip a coin 1000 times and predict the stock market to go up if heads turned up in majority and vice-versa. He thinks that this approach could get him a better prediction of the stock market because an ensemble of independent classifiers could potentially obtain a better prediction. Do you agree with him? Give a brief justification.

**Question 2 (Class Imbalance) [5 points]**

We have test data of 1000 samples with two classes: + class (100 samples) and – class (900 samples). Consider a random classifier C0 that classifies a test instance to the + class randomly with a probability  $p$ .

- [2 points] What is the expected precision and recall for C0?
- [1 point] Write the expression for the F-measure of C0.
- [2 points] Consider another classifier C1 whose F-measure is known to be 0.15. For what range of values of  $p$  is C1 better than C0 in terms of the F-measure?

**Question 3 (Class Imbalance) [14 points]**

You are trying to evaluate four different blood tests, T1, T2, T3, and T4, that have been developed to detect a particular type of cancer. These tests have been developed by different organizations over different datasets, and their evaluations have been reported in the following confusion matrices.

Test T1:

| Dataset: (1100 patients) | Predicted by Blood Test |                     |
|--------------------------|-------------------------|---------------------|
| Actual                   | Cancer (+ class)        | No Cancer (- class) |
| Cancer (+ class)         | 40                      | 60                  |
| No Cancer (- class)      | 100                     | 900                 |

Test T2:

| Dataset: (200 patients) | Predicted by Blood Test |                     |
|-------------------------|-------------------------|---------------------|
| Actual                  | Cancer (+ class)        | No Cancer (- class) |
| Cancer (+ class)        | 20                      | 80                  |
| No Cancer (- class)     | 5                       | 95                  |

Test T3:

| Dataset: (200 patients) | Predicted by Blood Test |                     |
|-------------------------|-------------------------|---------------------|
| Actual                  | Cancer (+ class)        | No Cancer (- class) |
| Cancer (+ class)        | 50                      | 50                  |
| No Cancer (- class)     | 50                      | 50                  |

Test T4:

| Dataset: (200 patients) | Predicted by Blood Test |                     |
|-------------------------|-------------------------|---------------------|
| Actual                  | Cancer (+ class)        | No Cancer (- class) |
| Cancer (+ class)        | 50                      | 50                  |
| No Cancer (- class)     | 10                      | 90                  |

- a. [8 points] Using these confusion matrices, compute the TPR (recall), FPR, precision, and F-measure for the four tests and report these numbers in the table below.

|         | TPR | FPR | Precision | F-measure |
|---------|-----|-----|-----------|-----------|
| Test T1 |     |     |           |           |
| Test T2 |     |     |           |           |
| Test T3 |     |     |           |           |
| Test T4 |     |     |           |           |

- b. [2 points] Between T1 and T3, which test would you consider to be better? Briefly explain.
- c. [2 points] Between T1 and T4, which test would you consider to be better? Briefly explain.
- d. [2 points] Between T2 and T4, which test would you consider to be better? Briefly explain.

#### Question 4 (ROC Curve) [16 points]

You are asked to evaluate the performance of two classification models, M1 and M2, for a binary classification problem with classes '+' and '-'. For every test instance, x, each of the two models provides a posterior probability of x belonging to class '+'. The following table provides a list of 10 test instances with their true classes, and their posterior probabilities of belonging to class '+', according to M1 and M2.

| Instance | True Class | $P(+ M1)$ | $P(+ M2)$ |
|----------|------------|-----------|-----------|
| 1        | +          | 0.98      | 0.27      |
| 2        | +          | 0.31      | 0.45      |

|    |   |      |      |
|----|---|------|------|
| 3  | + | 0.92 | 0.95 |
| 4  | + | 0.31 | 0.46 |
| 5  | + | 0.93 | 0.23 |
| 6  | - | 0.33 | 0.13 |
| 7  | - | 0.47 | 0.08 |
| 8  | - | 0.46 | 0.19 |
| 9  | - | 0.24 | 0.37 |
| 10 | - | 0.45 | 0.04 |

- a) **[5 + 3 points]** Plot the ROC curve for both M1 and M2 (you should plot them on the same graph.) (*Hint:* to obtain the ROC curve of a model, you need to compute the TPR and FPR for every threshold on the posterior probability estimated by the model.) Using the ROC curves, compute the AUC for M1 and M2. Which model do you think is better? Explain your reasons.
- b) **[4 points]** Suppose you choose a cutoff threshold to be  $t = 0.4$  for both the models, M1 and M2. In other words, any test instance whose posterior probability is greater than  $t$  will be classified as a positive example. Compute the Precision, Recall, and F-Measure for M1 and M2 after using the cutoff threshold of  $t = 0.4$ . Which model is better using F-measure as the evaluation criterion? Are the results consistent with what you expect from the ROC curve?
- c) **[4 points]** Repeat part (b) using  $t = 0.7$ . Which model is better using F-measure as the evaluation criterion? Are the results consistent with what you expect from the ROC curve?

#### Question 5 (Types of Clustering) [12 points]

Each of the following parts describes a collection of groups, or groupings. Describe each of these groupings in terms of the characteristics we discussed for sets of clusters. More specifically, label every grouping as to whether they are:

- hierarchical or partitional
- overlapping or non-overlapping
- complete or partial


Each grouping should be labeled with three characteristics, e.g., partitional, overlapping, and incomplete. If you feel there may be some ambiguity about what characteristics a grouping has, provide a short justification of your answer.

- a) The objects are movies, and the groups are the types of movie: action, drama, kids, science-fiction, etc.

- b) Grouping of students in a university based on the organization (department, college, institute etc.) they belong to. A student may belong to multiple organizations. Also, some students don't have declared majors and hence may not belong to any organization.
- c) Grouping of all the students in the Computer Science department based on the letter grade they get in CS(STAT) 5525.
- d) You want to group all locations on Earth as to whether they belong to a tropical rainforest, a deciduous forest, or an evergreen forest. Here, each location corresponds to a region of surface area 1km X 1km, and a location can have more than one variety of forests.

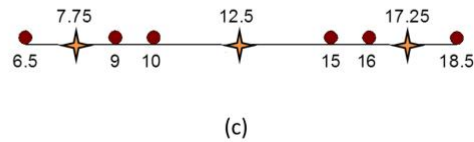
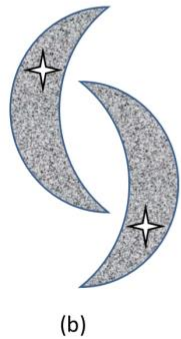
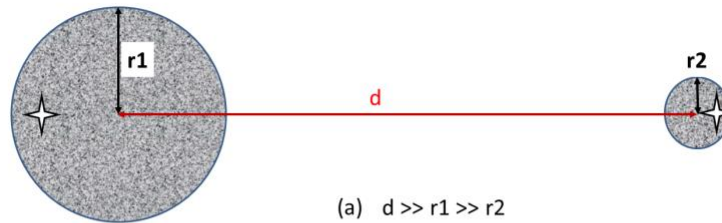
**Question 6 (k-means) [2 x 3 = 6 points]**

To answer the following true / false questions on how k-means operates, refer to figures (a), (b), and (c), below. Note that we are referring to the very basic k-means algorithm presented in class and not to any of its more sophisticated variants, such as bisecting k-means or k-means++.

Note that for all three figures, the initial centroids are given by the symbol:  Initial point

For figures (a) and (b) assume the shaded areas represent points with the same uniform density. For Figure (c), the data points are given as red dots and their values are indicated under the dots. No explanation for your answer is necessary unless you feel there is some ambiguity in the figure or the question.

- a) **True or False:** For Figure (a) and the given initial centroids, when the k-means algorithm completes, each shaded circle will have one cluster centroid at its center.
- b) **True or False:** For Figure (b) and the given initial centroids, when the k-means algorithm completes, there will be one cluster centroid in the center of each of the two shaded regions and each of the two final clusters will consist only of points from one of the shaded regions. In other words, none of the two final clusters will have points from both shaded regions.
- c) **True or False:** For Figure (c) and the given initial centroids, the final clustering for k-means contains an empty cluster.



## Practice Questions

### Question 7 (Ensemble Learning)

For a binary classification problem, you are given a collection of base classifiers where every base classifier has an error rate of  $e$ . The ensemble prediction is simply the majority vote of the predictions of the base classifiers. State whether the following statements are “true” or “false.”

- If  $e = 0.5$  and the predictions of all base classifiers are independent, the error rate of the ensemble classifier will be smaller than  $e$ .
- If  $e = 0.3$  and all base classifiers are identical, the error rate of the ensemble classifier will be smaller than  $e$ .

### Question 8 (Class Imbalance)

Consider a test data of 1000 samples with two classes: +ve class (100 samples) and –ve class (900 samples). We have two random classifiers C1 and C2. Classifier C1 classifies test data randomly to +ve class with a probability  $p$  (and to –ve class with probability  $1-p$ ) and classifier C2 classifies test data to +ve class randomly with a probability  $2p$ .

- What is the expected value of TPR and FPR for classifiers C1 and C2? Is C2 a better classifier than C1?
- What is the expected value of precision and recall for classifiers C1 and C2? Which evaluation metric pair between {TPR and FPR} and {precision and recall} do you think is correctly indicating the relative performance of classifiers C1 and C2?

### Question 9 (Class Imbalance)

- a) You are working with a doctor to evaluate how well a new, inexpensive blood test can detect a particular type of cancer. 1000 subjects are recruited from a population at high risk for the cancer and evaluated for cancer using a very expensive, but 100% accurate medical procedure. 100 subjects are found to have cancer. The 100 subjects with cancer and another 100 subjects without cancer are given the inexpensive blood test. Results are shown in the following confusion matrix.

|           | Predicted by Blood Test |           |     |
|-----------|-------------------------|-----------|-----|
| Actual    | Cancer                  | No Cancer |     |
| Cancer    | 90                      | 10        | 100 |
| No Cancer | 10                      | 90        | 100 |
|           | 100                     | 100       | 200 |

Using the above confusion matrix, compute the precision, recall, TPR, and FPR for the Cancer class.

- b) The doctor is very excited about these results, but wants to see what the results will be after all the blood tests are evaluated. The remaining 800 subjects (none of which have cancer) are given the blood test.

The confusion matrix for all 1000 subjects is given below.

|           | Predicted by Blood Test |           |      |
|-----------|-------------------------|-----------|------|
| Actual    | Cancer                  | No Cancer |      |
| Cancer    | 90                      | 10        | 100  |
| No Cancer | 90                      | 810       | 900  |
|           | 180                     | 820       | 1000 |

For this new confusion matrix, compute the precision, recall, TPR, and FPR for the Cancer class.

- c) Which of the measures (precision, recall, TPR, FPR) have changed and which have stayed the same? Comment on why some measures were affected and others were not.
- d) You were disappointed by the change you observed in one of the measures from the first confusion matrix to the second, but the doctor was not. The doctor tells you to consider sensitivity (Recall) and specificity (1 - FPR). Compute the sensitivity and specificity for both confusion matrices. Do their values change from the first confusion matrix to the second?

- e) When would we prefer (Sensitivity, specificity) as the preferred choice of evaluation measures? When would we prefer (precision, recall) as the preferred choice of evaluation measures?

#### Question 10 (Class Imbalance)

You are given a classification algorithm that predicts whether it will rain tomorrow (+) or not (-). The confusion matrix of the algorithm on 1000 test days is given below:

|  | Predicted |    |     |
|--|-----------|----|-----|
|  |           | +  | -   |
|  | Actual    |    |     |
|  | +         | 20 | 50  |
|  | -         | 80 | 850 |

- Compute accuracy, precision, recall, and F-measure with respect to '+' class.
- Which of these metrics is a poor indicator of the overall performance of your algorithm? Which of these metrics is a good indicator of the overall performance? Give a one sentence reason why this is the case?
- Construct a trivial classifier that achieves better accuracy by classifying all test instances to the same class, irrespective of their attributes.

#### Question 11 (Types of Clustering)

Each of the following parts describes a collection of groups, or groupings. Describe each of these groupings in terms of the characteristics we discussed for sets of clusters. More specifically, label every grouping as to whether they are:

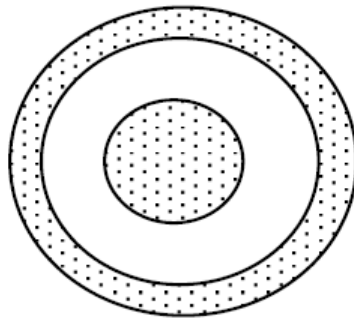
- hierarchical or partitional
- overlapping or non-overlapping
- complete or partial

- The groups are all possible sets of 10 items drawn from a set of 100 items.
- Each group is the set of all people in the military who have the same rank.

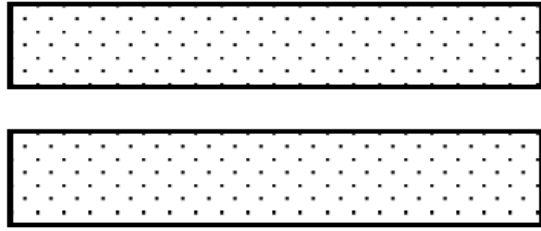
#### Question 12 (k-means)

For the following sets of two-dimensional points, (1) provide a sketch of how they would be split into clusters by K-means for the given number of clusters and (2) indicate approximately where the resulting centroids would be. Assume that we are using the squared error objective function. If you believe that there is more than one possible solution, then please indicate whether each solution is a global or local minimum. Darker areas indicate higher density. Assume a uniform density within each shaded area.

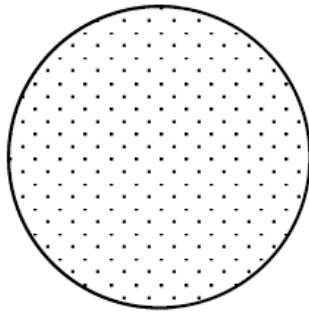
- (a)  $k = 3$ .



(b)  $k = 2$

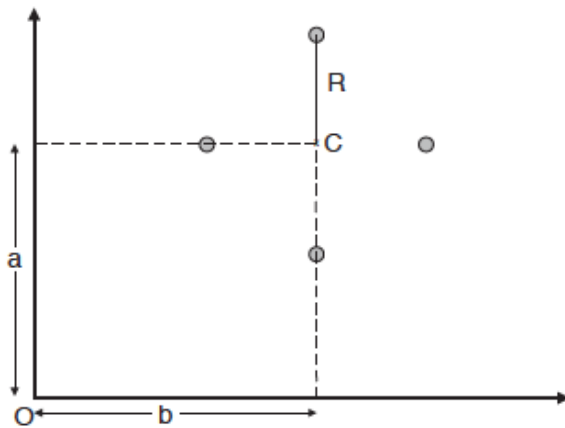


(c)  $k = 2$



### Question 13 (k-means)

Consider the 4 data points shown in the following figure. The distance between each data point to the center  $C$  is  $R$ .



- Compute the total SSE of the data points to the centroid,  $C$ .
- Compute the total SSE of the data points to the origin,  $O$ .



c. Using the above calculations, compute the SSE for the 8 data points shown below with respect to the centroid, D. Note that there are two groups of 4 points each, and the distance of every point from the center of its group is  $R/2$ .

