1 Clustering Types

There are 6 different datasets noted as A,B,C,D,E,F. Each dataset is clustered using two different methods, and one of them is K-means. All results are shown in the following figure (1). You are required to determine which result is more likely to be generated by K-means method. (Hint: check the state when K-means converges; Centers for each cluster have been noted as X; Since x and y axis are scaled proportionally, you can determine the distance to centers geometrically). The distance measure used here is the Euclidean distance.

Figure 1: Clusterings with cluster centroids marked.
For each dataset, give which clustering (1 or 2) you think more likely to have been the result of k-means. For the one you don’t think a likely result of k-means, state an algorithm you think could give that clustering and explain why.

1. Dataset A
2. Dataset B
3. Dataset C
4. Dataset D
5. Dataset E
6. Dataset F

2 Clustering Evaluation

For datasets A, B, C, and F in Figure 1, it seems visually clear that clustering 1 is better than clustering 2. For each, describe an unsupervised measure that would suggest that clustering 1 is better (you may give the same measure for multiple datasets, if appropriate.) Describe briefly why it would say the left (1) clustering is better than the right (2) clustering.

1. Dataset A
2. Dataset B
3. Dataset C
4. Dataset F

3 Hierarchical Clustering

The following figure shows two clusters A (red) and B (blue), each has four members. The coordinates of each member are labeled in the figure. Compute the distance between two clusters using Euclidean distance.
Use Hierarchical clustering to cluster this data and plot the Dendrogram for each of the following algorithms:

1. Complete link
2. Single link
3. Average link

For purposes of the clustering, you are starting with all eight points and not using the information on clusters A and B. You might want to use A and B as "known categories" to do supervised evaluation of the clustering - this could help with understanding both how well hierarchical clustering works, and evaluation. This may help you to answer the question below:

Which algorithm is more robust to noise and why?

4 Margin-based classifiers

The following figure shows two classes (+ and −) with two features (X and Y). The lines A and B represent two different classifiers.
1. Is this data linearly separable? Explain why or why not.

2. Could either of the lines represent a Support Vector Machine classifier? State which one, both, or neither. Explain your answer.

3. Could either of the lines represent a Single-Layer Perceptron classifier? State which one, both, or neither. Explain your answer.

4. Could either of the lines represent a K-Nearest Neighbor classifier? State which one, both, or neither. Explain your answer.

5 SVM

In Figure 1, C1 shows a dataset nicely split into three clusters (red, green, blue). Assume the colors are classes, and we want to use this as training data to train a support vector machine to separate the three classes. We’ve only discussed binary classification. One way to make a multi-class SVM is to first divide into two classes, then divide one class into a third. For example, with C1 we could first learn to separate the red (lower) class from the blue (upper left) and green (upper right), then use a second SVM that would learn to separate the blue and green.

1. For C1, would you be better off with a hard margin support vector machine, or soft margin? Explain your reasoning.

2. A1 only has two classes (the inner ball and outer ring.) But an SVM would have a hard time with this. Explain why this would not be easily separated with a linear classifier.

3. We can often use SVM to solve non-linear classification problems using kernel functions. Give the formula for a kernel function that would allow us to use SVM to separate the two classes in A1. Hint: Assume the origin (0,0) is at the center of the figure.
6 Bias/Variance

Suppose in Assignment 5, you did k-fold cross-validation to check the accuracy of your model. With some folds, you were very accurate. With others, your model did little better than a random assignment.

1. Is this an example of high bias, or high variance? Explain.

2. Describe two different ways you could go about improving on this problem. Note that your suggestions may not be things you could have done with assignment 4, they are generally things you could try if you experience the problem of a wide range of accuracy across different tests.

7 Ensemble

Suppose, you are working on a binary classification problem. You have three models. The table below shows a test run on the three models, along with the true value of the output.

1. What is the accuracy of each of Models M1, M2, and M3?

2. What is the accuracy of the ensemble formed by taking the majority vote?

<table>
<thead>
<tr>
<th>Test instance</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>True Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

8 Ensemble

Boosting is an ensemble approach used to reduce bias. Assume we have trained a Naïve Bayes classifier that predicts how people will do in CS grad. school, and it has high overall accuracy (say, 90%). But on some students (say, those who went to Rose-Hulman Institute of Technology), it has much lower accuracy.
1. Explain briefly how boosting would be used to give us a better classifier (i.e., describe boosting.)

2. Would you need to know that the students on who this worked poorly went to Rose-Hulman, or know anything else about them? Explain.

3. Would your answers change if we’d started with a Support Vector Machine classifier?

9 E-M Clustering

We have said that K-Means is a special form of E-M clustering. Give an example dataset (you can just draw a 2-dimensional dataset if you wish) where E-M clustering would do better than K-Means, and briefly explain why it does better.

10 Association Rules

You have run the association rules algorithm on your data set, and the two rules \{banana, apple\} \rightarrow \{grape\} and \{apple, orange\} \rightarrow \{grape\} have been found to be relevant. Which of the following must also be true? Why?

1. \{grape, apple, orange\} must be a frequent itemset.

2. \{banana, apple, grape, orange\} must be a frequent itemset.

3. \{grape\} \Rightarrow \{banana, apple\} must be a relevant rule.

4. \{banana, apple\} \Rightarrow \{orange\} must be a relevant rule.