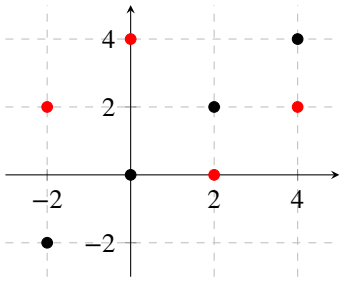
CS 188 Summer 2020 Introduction to Artificial Intelligence

Final Review

Clustering and Decision Trees Solutions

Q1. More Decision Trees

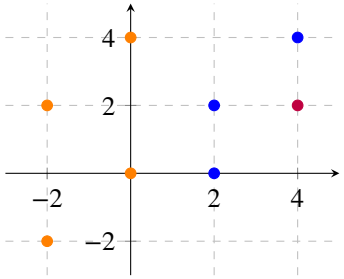
Throughout this problem, we'll be working with the following classification dataset. Values for two features are plotted along the x and y axes, and the two classes are shown in red vs. black.



(a) What is the depth of the shallowest decision tree that can correctly classify this training data? Assume that each decision involves comparing a single feature value against a chosen threshold.

Depth: 3

(b) Now suppose that we cluster the data points into three clusters (shown in orange, blue, and purple in the plot below):



(i) Could this clustering have been produced by running k-means (with the Euclidean distance function)?

○ Yes ● No

The point (4, 4) is closer to (4,2) than the centroid of the blue cluster, so this cluster assignment can't be the result of running k-means to covergence.

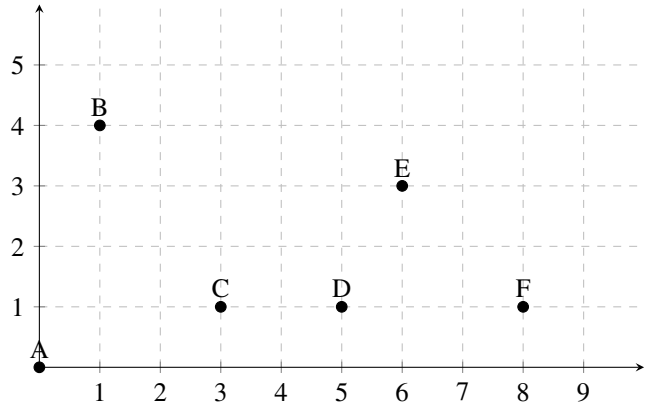
(ii) Now suppose that we incorporate cluster assignments as an additional feature value into our data. For example, the point $(0, 4)$, which originally had just two features ($f_1 = 0$, $f_2 = 4$), now has three features ($f_1 = 0$, $f_2 = 4$, and $f_3 = \text{Orange}$).

Using these augmented features, what is the depth of the shallowest decision tree that can correctly classify the data?

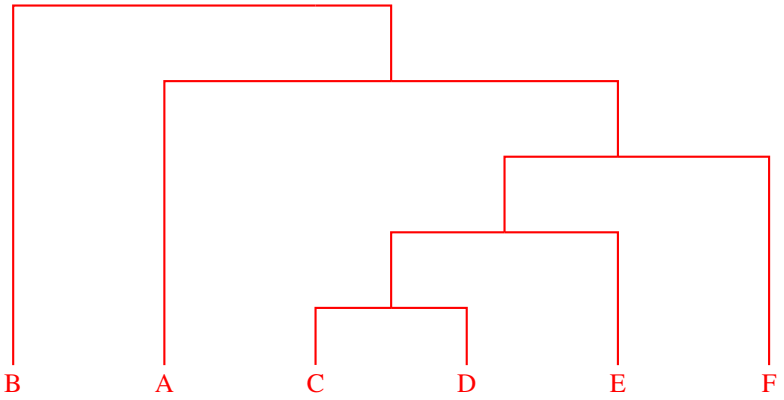
Depth: 2

Q2. K-Means Clustering

Consider the following set of data points. For all clustering algorithms in this problem, assume that Euclidean distance is used.



(a) Draw a dendrogram produced by agglomerative clustering for this collection of data points. Use *single-link* clustering, meaning that the distance between two clusters is measured using the closest pair of points between the two clusters.



(b) You now wish to run k-means clustering on this data, with $k = 2$. Rather than initializing the cluster centers randomly, you decide to use the output of agglomerative clustering for initialization. After splitting on the root node of the tree from part (a), what are the initial clusters used for initializing k-means?

Cluster 1: $\{B\}$. Cluster 2: $\{A, C, D, E, F\}$

(c) Based on the initial cluster assignments in part (b), what are the initial locations of the cluster centers?

Cluster 1: $(1, 4)$. Cluster 2: $(22/5, 6/5) = (4.4, 1.2)$

(d) Now re-assign each data point to its closest mean. What are the new cluster assignments?

Cluster 1: $\{A, B\}$. Cluster 2: $\{C, D, E, F\}$

(e) Using the cluster assignments in part (c), update the locations of the cluster centers.

Cluster 1: $(0.5, 2)$. Cluster 2: $(5.5, 1.5)$

(f) In this part, $\phi = \sum_i \text{dist}(x_i, c_{a_i})$ refers to the sum of distances from each data point to the center of its assigned cluster.

k-means always finds a local minimum of ϕ ● True ○ False
k-means always finds a global minimum of ϕ ○ True ● False
agglomerative single-link clustering always finds a local minimum of ϕ ○ True ● False
agglomerative single-link clustering always finds a global minimum of ϕ ○ True ● False

k-means is an optimization procedure for ϕ, and is guaranteed to converge to a local (but not global) optimum. On the other hand, agglomerative clustering is a greedy algorithm that provides no guarantees about the optimality of its output.