CS 188 Summer 2020 Introduction to Artificial Intelligence

Section 13

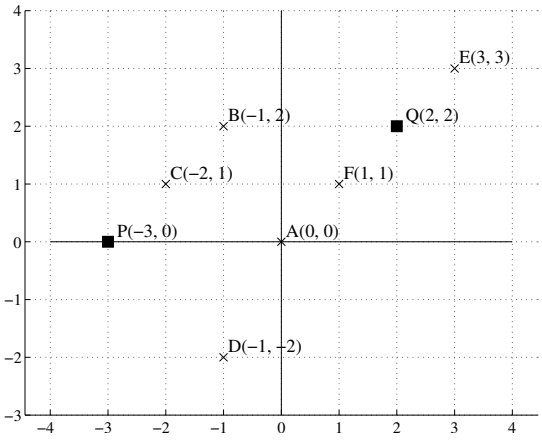
Clustering & Decision Trees Solutions

Q1. Clustering

In this question, we will do k-means clustering to cluster the points $A, B \ldots F$ (indicated by $\times$'s in the figure on the right) into 2 clusters. The current cluster centers are P and Q (indicated by the ■ in the diagram on the right).

Recall that k-means requires a distance function. Given 2 points, $A = (A_1, A_2)$ and $B = (B_1, B_2)$, we use the following distance function $d(A, B)$ that you saw from class,

$$d(A, B) = (A_1 - B_1)^2 + (A_2 - B_2)^2$$



(a) Update assignment step: Select all points that get assigned to the cluster with center at P:

☐ A ■ B ■ C ■ D ☐ E ☐ F ☐ No point gets assigned to cluster P

(b) Update cluster center step: What does cluster center P get updated to?

The cluster center gets updated to the point, P' which minimizes, $d(P', B) + d(P', C) + d(P', D)$, which in this case turns out to be the centroid of the points, hence the new cluster center is

$$\left(\frac{-1-2-1}{3}, \frac{2+1-2}{3}\right) = \left(\frac{-4}{3}, \frac{+1}{3}\right)$$

Changing the distance function: While k-means used Euclidean distance in class, we can extend it to other distance functions, where the assignment and update phases still iteratively minimize the total (non-Euclidian) distance. Here, consider the Manhattan distance:

$$d'(A, B) = |A_1 - B_1| + |A_2 - B_2|$$

We again start from the original locations for P and Q as shown in the figure, and do the update assignment step and the update cluster center step using Manhattan distance as the distance function:

(c) Update assignment step: Select all points that get assigned to the cluster with center at P, under this new distance function $d'(A, B)$.

■ A ☐ B ■ C ■ D ☐ E ☐ F ☐ No point gets assigned to cluster P

(d) Update cluster center step: What does cluster center P get updated to, under this new distance function $d'(A, B)$?

The cluster center gets updated to the point, P' which minimizes, $d'(P', A) + d'(P', C) + d'(P', D)$, which in this case turns out to be the point with X-coordinate as the median of the X-coordinate of the points in the cluster and the Y-coordinate as the median of the Y-coordinate of the points in the cluster. Hence the new cluster center is $(-1, 0)$

Q2. Decision Trees

You are given points from 2 classes, shown as +'s and ·'s. For each of the following sets of points,

1. Draw the decision tree of depth at most 2 that can separate the given data completely, by filling in binary predicates (which only involve thresholding of a *single* variable) in the boxes for the decision trees below. If the data is already separated when you hit a box, simply write the class, and leave the sub-tree hanging from that box empty.

2. Draw the corresponding decision boundaries on the scatter plot, and write the class labels for each of the resulting bins somewhere inside the resulting bins.

If the data can not be separated completely by a depth 2 decision tree, simply cross out the tree template. We solve the first part as an example.