M6-05: k-Means Clustering

Part of the "Towards Machine Learning" Learning Badge

Video Walkthrough: *https://discovery.cs.illinois.edu/m6-05/*

Where Linear Regression Fails Us

Every single model will have datasets that work really well and every model will have datasets that work really poorly. Consider the dataset to the right — what is the line of best fit through this data?

Instead of **predicting** a value, what if we wanted to **classify** a new point into an existing group?

---

k-Means Clustering

One of the simplest algorithms to classify data is _____.

Algorithm

The entire algorithm can be completed in five steps:

[Setup]:

1. Let **k** be the number of clusters among **n** data points
2. Choose **k** different starting _____, **non-data points** we refer to as $\mathbf{c}_0$, $\mathbf{c}_1$, …, $\mathbf{c}_{k-1}$

[Assignment]:

3. For every data point, assign it to the **centroid** closest to the data

[Update]:

4. Update the location of every **centroid** to be equal to the average value of the data assigned to that centroid
5. Repeat Steps 3-5 until the location of all centroids move by less than some error, ϵ (eg: 0.001).

Run k-means clustering with k=2 and the starting centroids:

- `(0.2, 0.2)`
- `(0.2, 0.4)`