TDT4173 Machine Learning

September 13, 2023

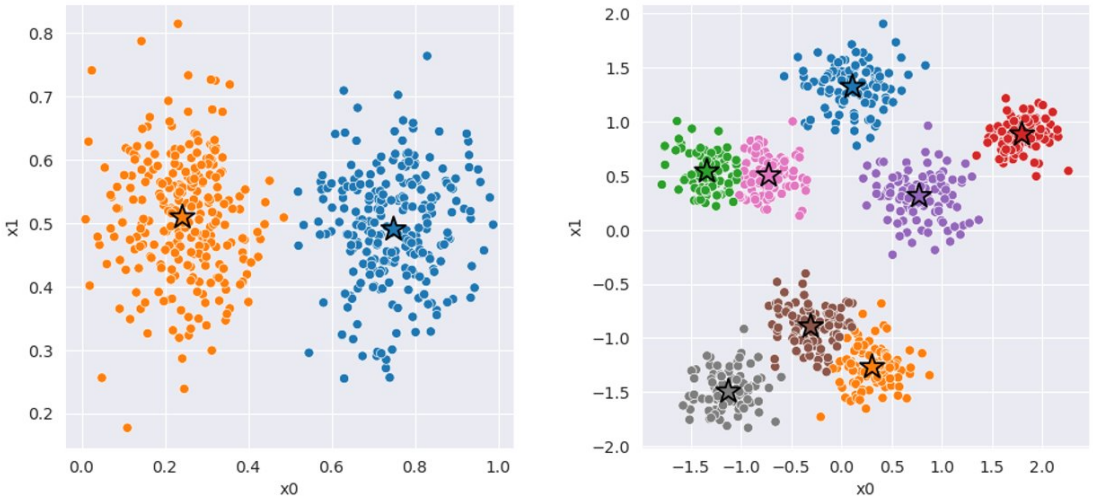
Problem 1 - K-Means Clustering

The K-Means clustering algorithm is an unsupervised machine learning algorithm. It's used to partition a data set into K different clusters based on how similar the data points are. The algorithm chooses how many clusters it want to partition based on user input. Then, each data point is assigned to its closest cluster, now centroid, until all data points are assigned. The distances is usually computed with Eucledian distance. The centroids position is updated at every step to minimize the squared distances between data points and their assogned centroid until convergence. That is, when the centroids no longer change its position between iterations.

The algorithm is used in problems where you want to group the data. This can be in customer segmentation, image compression and anomaly detection.

The inductive bias is based on assumptions about the data set in order to generalize and perform clustering:

- Clusters are spherical
- Clusters are of equal size
- Clusters have the same density
- Variance of data within each cluster roughly the same
- The number of clusters is given and fixed



In the figure above you can see the result from data set 1 and 2 respectively. In the second data set, there are more clusters and the axes don't have the same scale. The different scale can make it difficult for the algorithm to appoint points to each cluster. This can violate the inductive biases above. To counter this, I normalized the data points with the method *normalize* and added a boolean hyperparameter **normalize**. This gave the result seen in the figure to the right.

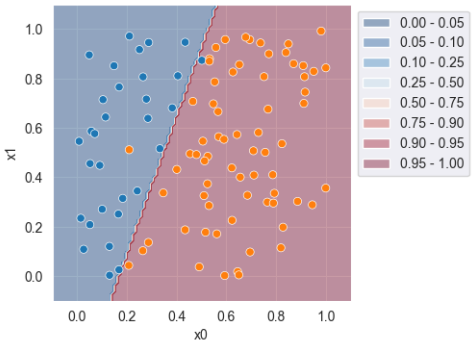
With data set 1, I got a silhouette score of 0.672 and distortion of 8.837. With data set 2, after the fixes, I got a silhouette score of 0.594 and distortion of 59.779.

Problem 2 - Logistic Regression

Logistic regression is a statistical model used for classification and predictive analytics. It estimates the probability of an event happening, based on a data set of independent variables. The dependent variable is always between 0 and 1 and it uses a logit transformation on the odds. The algorithm is suited for problems where you have to predict one of two outcomes, for example if someone voted or not.

The inductive bias is the assumption of a linear relationship between features and the logit transformation. This makes it sensitive to feature scaling.

The result from data set 1 is plotted below. Here I got an accuracy of 98% and cross entropy of 0.691 with a learning rate of 1.



The data in the second set is not sorted as nicely in data set 1. Therefore, the algorithm got an accuracy of around 60 %. To increase this, I chose to manipulate the data by mirroring the points around a horizontal axis and calling this column "X_2" in the data set. I then ran the algorithm again for an accuracy score of 91.6% and cross entropy of 2.901 on the training set. On the test set I got an accuracy of 90.8 % and cross entropy 3.178. The scatter plots of data set 2 and the manipulated data set 2 is shown below.