Maskinlæring Task1 Report

Hallvard Enger Bjørgen, 506446

September 2023

1 Linear Regression

1.1 Technical Understanding

How does the algorithm work on a technical level and what kind of machine learning problems is it suited for? Linear regression is a supervised machine learning algorithm for regression problems where a linear relationship exists between the dependent variable and independent variables. It aims to find optimal coefficients to minimize a cost function. Gradient Descent optimizes these coefficients. It's suited for problems with linear relationships but may not perform well with highly nonlinear data, where other models like polynomial regression or decision trees may be better.

Mathematically, the linear model can be represented by "target = bias + coefficients times each their independent variable".

1.2 Inductive Bias

What is its inductive bias? The main inductive bias is the assumption that the relationship between dependent variable and predictors is linear. The fact that the predictors are independent also bears the meaning that their effect should be additive, i.e. an increase in one should mean the same increase in the dependent variable independent of what values are present for the other predictors.

1.3 2nd Dataset

What happens in the second dataset that makes it harder than the first and how does this problem relate to the algorithm's inductive bias? In the second dataset there are two classes, A and B (binary classification, and the decision boundary between these classes is a close to circular region of radius x centered at a specific point.

The classification of which class a point belongs to is no longer dependent on the two independent variables, but rather the points distance to a center which is dependent on both the antecedents. This means that the given features (x0 and x1) no longer are indeependent of each other.

1.4 Workaround

What modifications did you do to get around this problem? To deal with this issue, a third feature, a points distance to the center of all datapoints, was calculated and added to the point. This value was then the main factor in training the model to predict the correct class. The x0 and x1 variables were still including in the training of the model as the circular decision boundary was rather oval in it's shape, so the x0 and x1 variables still had some purpose outside from determining the distance to centre.

1.5 Achieved Results

See 1 for results. A greater accuracy was achieved by feature engineering. Cross-entropy is still high, which may result from the model not being confident in each predictions.

	Accuracy	Cross Entropy
Dataset 1	0.940	0.277
Dataset 2 Train	0.786	0.526
Dataset 2 Test	0.800	0.517

Table 1: Results for linear regression

2 K-Means Clustering

2.1 Technical Understanding

How does the algorithm work on a technical level and what kind of machine learning problems is it suited for? The algorithm aims to group data into K cohesive clusters. It is unsupervised, and works as follows:

1. Choose random cluster centroids (out of the data provided is one popular way of doing it).
2. Assign each data point to their closest cluster
3. Move each cluster centroid to the actual centre of the assigned cluster points (median may be used as well)
4. Repeat step 2 and 3 until max iterations is reached, the cluster centroids are the same as the previous iteration, or some performance measurement is reached.

Works well for (amongst others) cluster analysis, anomaly detection, customer segmentation and image compression.

2.2 Inductive Bias

What is its inductive bias? The inductive bias of K-Means is that it assumes clusters are spherical, equally sized, and have similar densities. It also assumes that data points within a cluster are closer to each other than to points in other clusters. Additionally, K-Means is sensitive to the initial placement of centroids (which can be done different ways), and, in most implementations, the number of clusters (K) needs to be specified beforehand, which can influence the results.

2.3 2nd Dataset

What happens in the second dataset that makes it harder than the first and how does this problem relate to the algorithm's inductive bias? The unbalanced data magnitudes makes the algorithm work poorly if the data is not normalized (relating to the bias of the data being equally sized). Additionally, the number of clusters is above 6 (obvious from the plot), which makes the user need to increase the K. In higher dimensions where plots is not as easy, this may be harder to solve.

2.4 Workaround

What modifications did you do to get around this problem? I increased the K to 8, I changed the way of updating the centroids have the option (set upon initialization) to use the cluster median instead of mean, and I normalized the data before fitting the model.

2.5 Achieved Results

Silhouette score usually ends at around 0.59, and distortion at 4. Sometimes, the algorithm doesnt find the optimal solution. This could be possibly be improved by selecting initial centroids more intelligently.

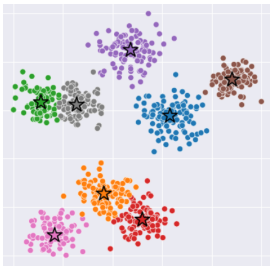

Figure 1: Nice result

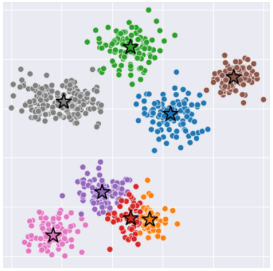

Figure 2: Worse result