TDT4173 Maskinlæring

Individual assignment

Name: Taheera Ahmed
Submission date: 16. september 2023

NTNU
Norwegian University of
Science and Technology

K-means

How does the algorithm work on a technical level and what kind of machine learning problems is it suited for?

Basic K-means utilizes euclidean distance and centroids to cluster different data points in space. It is used for classifying similar data in the feature space, in the same type of clusters. Meaning if you see different features, K-means tries to predict which cluster it fits the best. Which means that it can be used for tasks such as document categorization and is typically used for with data that has a smaller number of dimensions, is numeric, and is continuous [1].

What is its inductive bias, i.e., what assumptions does it make about the data in order to generalize?

K-means utilizes different techniques in order to generelize data, make assumptions and place them in the correct cluster.

1. Assumption of spherical clusters - Meaning it works better on spehircal clusters of data which is equally dense and roughly the same size.
2. Fixed number of clusters - It requires you to know how many clusters are needed.
3. Initialization sensitivity - It's sensitive to where the centroids are place initially. Meaning if they are placed wonky in the first place, it might ruin the clustering ability of the algorithm.
4. Variance of data - The data points should have rougly the same variance.

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 scale of the data made it difficult to learn from the data and to generalize, with just the basic implementation of K-means. Meaning the clusters looked absolutley horrifying with the basic implementation on the second dataset.

By looking at the mean and standard deviation of the two datasets the second dataset has different scales where x0 had a mean around 46, and x1 had a mean around 5. This made it harder for the algorithm to work effectivley.

What modifications did you do to get around this problem?

Normalizing the data the K-means algorithm worked a lot better. I also implemented K-means plus plus for better centroid initialization but this proved to not be neccessary. I also looked into tolerance of convergence, but that wasn't necessary either.

Logistic regression

How does the algorithm work on a technical level and what kind of machine learning problems is it suited for?

Logisitic regression is a machine learning algorithm which wokrs best for binary classification problems. It models the relationship between a target and one or more features by utilizing the logistic function.

It models the probability of an input belonging to a given class using the logistic function. The function transforms a linear of combination of the input features and outputs a probability score between 0 and 1 which is used for deciding how it should be classified.

The model looks like a neural network, just with one hidden layer with one node. Meaning it has as many weights as input features and outputs one value between 0 and 1. These weights will be adjusted during training with respect to minimize a loss function.

When the training is completed the logistic regression can predict a class label of new data points by using the trained network with the weights.

It can be useful for spam detection and credit scoring and low-dimensional datasets.

What is its inductive bias, i.e., what assumptions does it make about the data in order to generalize?

1. Linearity - Assumes a linear relationship between features and targets.
2. Sigmoid activation - Maps a linear combination of the input features between 0 and 1 by using sigmoid activation function.
3. No assumption about feature distributions - This means unlike K-means, Logistic Regression doesn't need data with approx. the same distribution.
4. Interpretability

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 class imbalance where varuable y is 0.61, which indicated that there is a class imbalance of class 1. The dataset also has a narrower spread of feature values compared to the first dataset.

What modifications did you do to get around this problem?

I added degree of polynomial features (which can help depict more complex relationships between features), add a learning rate, number of iterations, regularization, convergence threshold and batch size.

Referanser

[1] Ritvik Ranjan. *K-means Clustering and its applications*. Hentet 16.09.2023. URL: https://www.linkedin.com/pulse/k-means-clustering-its-applications-ritvik-ranjan/.