Author

Matwej Novgorodtsev; Student ID: 244580IV

1 Exercise 1

In this exercise, four distance metrics are defined: Euclidean distance calculates the straight-line distance between two points; Manhattan distance sums the absolute differences of their coordinates; Chebyshev distance finds the maximum absolute difference along any coordinate; and Minkowski distance generalizes these, allowing for different values of 'P' to compute various distances.

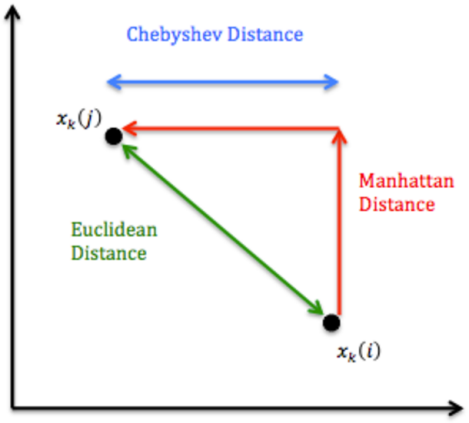

Figure 1: Distances in Machine Learning [1]

2 Exercise 2

Entropy measures the disorder in a dataset. Higher entropy means more variety in the data, while lower entropy indicates that the data is more similar. Additionally, in this exercise, we learned about the Fisher score, which helps measure how well a feature distinguishes between different classes in a dataset. It identifies the most important features for improving classification performance.

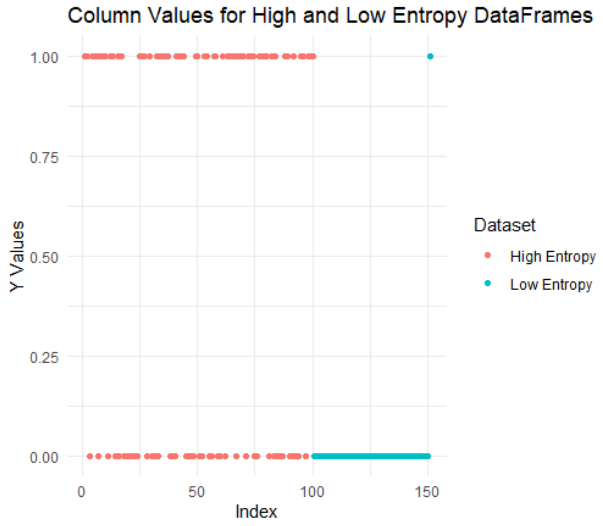

Figure 2: Data distribution for low and high entropy datasets.

3 Exercise 3

In the third task, our goal was to create a decision tree. We used the entropy function from the previous task to determine which data split provides the most information. The data used for training consisted of coordinates in three-dimensional space (X, Y, Z) with an assigned class (from 1 to 3). Cross-validation was used to improve the model's quality. The decision tree was built, and the presentation of data used and metric results are presented below:

Figure 3: Dataset with 3 classes.

Table 1: Metrics Results

Class	Precision	Recall	F1
1	0.9857	0.6252	0.7548
2	0.7820	0.7907	0.7685
3	0.4142	0.8833	0.5341

4 Exercise 4

In Exercise 4, we implemented a linear regression model using forward selection. Starting with six features, the model was reduced to three significant predictors: x3, x6, and x2. This stepwise process added variables based on their statistical significance, improving the model's accuracy.

References

[1] Distances in Machine Learning