Laboratory assignment

Component 1

Authors: Ichim Stefan, Mirt Leonard
Group: 246/1

October 29, 2024

1 Task 1 - Unsupervised Learning

1.1 Problem Definition

- We want to create an algorithm that could assist us in observing similarities between different housings in California, observations which could have been missed by market analysts.

1.2 Problem Specification

Inputs

Tabluar data from the "California Housing Prices" dataset. Features which will be used for the pattern learning consist of: longitude, latitude, housing median age, total rooms, total bedrooms, population, households, median income and ocean proximity. The median house value will act as the target value and will be used to evaluate if patterns from the same clusters have similar values.

Preconditions

Pattern features will need to be scaled for the algorithm to work with numbers from a smaller interval.

$$x = \frac{x - mean(x)}{std(x)}$$

Outputs

The unsupervised algorithm will indicate which cluster a pattern will be part of.

Postconditions

A natural number, representing the number of the cluster.

1.3 Learning Task Specification

Task

Create appropriate clusters for various patterns in the dataset.

Performance

The algorithm will be evaluated using both external and internal performance measures. According to these results, hyperparameters will suffer changes so that the algorithm reaches higher results.

For internal measures, silhouette score and Calinski-Harabasz Index can be used. The first one measures how similar an object is to its own cluster compared to other clusters, and the latter measures the ratio of between-cluster dispersion to within-cluster dispersion.

An external measure could be represented by a "feature hold-out" technique, where we choose to exclude a feature from the pattern when building our clusters, and then later on using that feature to examine how well the created clusters allign with it.

Experience

The experience consists of patterns represented by the entries of the dataset.

2 Task 2 - Supervizezd Regression

Problem Definition

The problem we want to solve is the prediction of the price of housing in California based on the input data. The goal is to develop a model that accurately estimates the selling price for any given house by learning patterns from historical housing data.

Problem Specification

Input data: median income, housing median age, average rooms, average bedrooms, house holds, average occupation, latitude, longitude, ocean proximity. Median house value will be used to train the decision tree regression. The input data is taken from California Housing Dataset available in scikit-learn.
Precondition:

- Data Completeness: All input features (e.g., median income, housing median age, average rooms, etc.) should have no missing values or should be imputed if missing.
- Data Quality: Numerical values should be within realistic ranges (e.g., median income should be a positive number). Latitude and longitude values should accurately represent California's geographic boundaries.
- Data Consistency: Features should be in consistent units and scales (e.g., income in thousands of dollars, age in years). All records should represent housing data from California to match the geographical scope of the dataset.

Output data: median house price in thousands of dollars
Postcondition: the output is a continuous numerical value

Specification of the learning Task

The model will be trained by recursively splitting the data based on feature values to create "branches" and "leaves" of a tree. Each split is chosen to minimize prediction error in the resulting subsets.
The objective is to train a model that minimizes the prediction error for median house prices. The model learns patterns in the historical data to make accurate predictions on new, unseen data.

Evaluation Metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE) to measure the accuracy of predictions.