Laboratory assignment

Component 1

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

October 28, 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.