

Summary report on Machine Learning

Rafał Górniak & Piotr Hynasiński

December 15, 2024

1 Introduction

This report provides an overview of machine learning, its methods, and applications. Machine learning is a branch of artificial intelligence that enables systems to learn and improve from experience without being explicitly programmed. It is widely used in various fields such as healthcare, finance, and technology, playing a critical role in data-driven decision-making.

The field of machine learning is mainly divided into three categories: supervised, unsupervised and reinforcement learning. Each category addresses different types of problems and leverages certain approaches and techniques to solve them. In this report, we focus on some of the most commonly used algorithms within the area of mentioned types, highlighting their features and practical usage.

The paper is divided into six sections. The first five describe, one by one, each of the approaches and functions taken into account, including classification and recommendation techniques. The last section presents results and conclusions as well as our opinions concerning the appropriateness and effectiveness of the described methods.

Every piece of data used in here comes from TMDB movie database. As an input for training, we took train.csv containing exactly 90 rows with rating for particular movie for 358 users. The goal was to assess 30 new films for users, being stored in test.csv. Thanks to server API, we were able to fetch proper fetures for each of the movies, about what there would be more later on.

2 K-Nearest neighbour

2.1 Overview

K-NN algorithm, belonging to supervised learning is primarily used for classification and regression tasks. The key idea behind its concept is that similar data points are likely to possess similar labels, which is based usually on the distance or measured with some kind of similarity feature calculations.

In description of nearest neighbour approach the information about extracted (selected) features, about their normalization and about similarity measure should be provided.

3 Decision tree

While presenting decision tree approach you should describe induction of single tree (e.g. what conditions were considered, how information gain was measured, when the induction process was stopped). This section should include an image showing how one induced tree (for one person) looks like.

4 Random forest

While presenting random forest approach you should describe working principles of random forest ensemble (e.g. how the diversity of ensemble members was assured, what was the procedure of results aggregation). This section should include images showing how sample induced trees in the forest (for one person) look like.

5 Person similarity

Describing person similarity approach report how people were compared and how the final movie evaluation was calculated.

6 Collaborative filtering

Discussing collaborative filtering approach focus on the process of training and on methods ensuring the generalization abilities of the trained model (e.g. selection of feature space dimensionality).

7 Summary and results

Presentation of results for every approach should be similar to make their comparison easier.