NTNU
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for IKT og realfag

Dato 24.03.2025

Referanse
V. Singh,
Alla Anohina-Naumeca

Exercise – Part E: Machine Learning

Deadline: Refer blackboard

Reading:

- Material on Blackboard

Hand in: Single pdf file

Introduction

For this assignment, you will select a business/research problem that can be solved using machine learning, select appropriate dataset and will apply methods of supervised and unsupervised learning. You are allowed to use any platform/language for machine learning.

Before starting to work on your assignment, you must find and choose a dataset on the web to solve the problem. Some of the well-known repositories are the following:

- UC Irvine Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php
- R Datasets on Github: https://vincentarelbundock.github.io/Rdatasets/
- Kaggle Datasets: https://www.kaggle.com/datasets
- Awesome Lists: Public Datasets: https://github.com/caesar0301/awesome-public-datasets
- https://www.linkedin.com/pulse/open-source-gis-data-abhinav-bhaskar-dwafc/
- Google Dataset Search: https://datasetsearch.research.google.com/
- https://paperswithcode.com/datasets
- https://www.nasa.gov/intelligent-systems-division/discovery-and-systems-health/pcoe/pcoe-data-set-repository/
- https://www.yelp.com/dataset
- You can also use your own dataset.

Selection of dataset

- avoid using the well-known dataset.
- the dataset should be of reasonable size (at least 200 data objects)
- includes 5 to 15 well defined features
- the dataset should be labelled

Part 1 – Pre-processing/Exploring the data:

1. Select a dataset suitable for classification task, and describe the dataset based on the information given in the repository/database where the dataset was located.

2. If the dataset you have acquired from the repository is not in a format that is easy to work with (like a comma-separated-values, or .csv, file), convert it into the needed format. Your dataset file should consist of an $n \times d$ table, where d is the number of dimensions of the data and n is the number of data objects.
3. Explore, understand and analyse the dataset (EDA) using commands/plots.
4. Identify and deal qualitative data/categorical data: If the values of any feature are textual values (e.g. yes/no, positive/neutral/negative, etc.), they must be transformed into numerical values.
5. Handle missing, duplicate and outlier values: Identify and find a way to obtain them in suitable way.
6. Feature selection and correlation: Remove some features if they are strongly related to the target variables. For example, in a dataset about cars, it appears that the city_mpg (Milage in city) and highway_mpg (Milage on highway) variables convey similar information. Also variables such as length, width, and height could potentially be combined into a single variable, like volume, if needed.

Include the following information in the report:

1. description of the dataset (providing references to the sources of information used):
    a. title, source, author and/or owner of the dataset.
    b. description of the problem domain of the dataset;
    c. licensing regarding the dataset (if any);
    d. a way of how the dataset was collected;
2. description of the content of the dataset:
    a. the number of data objects in the dataset;
    b. the number of classes in the dataset, the meaning of each class and the way of representing classes (explanation of the labels assigned to classes); if the data set provides several possible data classifications, then the report must clearly identify which classification is considered in the assignment;
    c. the number of data objects belonging to each class;
    d. the number and meaning of features in the dataset, as well as their value types and ranges (this information should be presented in a table consisting of the feature representation, its meaning, value type and range of values available in the dataset);

e. a snippet of the structure of your datafile in which the columns of your datafile and class labels are shown together with some data objects;

3. conclusions coming from the analysis of boxplot, scatter plots, histograms and distributions about the separability of your classes (remember to include your graphs in the report). Address the followings:
    a. Check whether classes are balanced, or imbalanced and how that may affect model performance.
    b. relationships amongst the data elements, e.g. inter-feature relationship, if there are many features, you can summarize them.
    c. calculate statistics (mean, mode, median, standard variance, Q1 and Q3) on your data and your observation on statistics values.
    d. What are the potential issues/challenges in your data such as missing values, duplicate values, qualitative data/categorical data, outliers. How did you handle them and why?
    e. How many data groupings can be identified by studying the visual representation of the data? It is a question of whether there are any separable groupings of data if the data objects of different classes merge.
    f. Are the identified data groupings close to each other or far from each other?

Part 2 – Supervised learning

For this part of the assignment, you will be running at least 2 supervised classification algorithms on the data you collected. You can also use an artificial neural network (ANN).

- To complete this part of the assignment, you will need to take the following steps:
    1. Choose 2 supervised learning methods that are suitable for classification task.
       For each supervised learning method selected:
        a. Divide your dataset in training and test sets.
        b. Perform at least 3 experiments using the training dataset, changing the values of the algorithm hyperparameters and analysing the algorithm performance metrics.
        c. Apply the trained model to the test dataset.
        d. Evaluate and compare the performance of the trained models.
        e. Check if there are signs of overfitting/underfitting?

Norges teknisk-naturvitenskapelige universitet

Dato
24.03.2025

Referanse
V. Singh
Alla Anohina-Naumeca

Include the following information in the report:

1. Short description about supervised learning algorithms you have used and motivation for choosing it.
2. Description of the hyperparameter/s, and its meaning to the algorithm.
3. Information on test and training datasets:
    a. the total number of data objects added to the test and training datasets (by number and %).
    b. information on how many data objects from each class are included in your training and test sets (by number and %).
    c. Conclusions on the performance analysis of the models
    d. Test results of trained model.

Part 3 – Unsupervised learning

This part of the assignment aims to look at the data in an unsupervised fashion to see if the assumptions about class structure hold.

1. Apply any 2 methods of unsupervised learning: Hierarchical clustering, K-Means and/or DBSCAN.
2. Depending on the algorithm selected:
    a. Perform at least 3 experiments with Hierarchical clustering, freely changing the values of hyperparameters, and analysing the operation of the algorithm.
    b. Perform experiments with the K-means algorithm using at least five different k values, calculate the Silhouette Score, and analyse the performance of the algorithm.
    c. Perform at least 3 experiments with the DBSCAN algorithm, freely changing the values of hyperparameters, and analysing the operation of the algorithm.

Include the following information in the report:

1. Describe the applied algorithms and their hyperparameters, explaining the meaning of each.
2. Description of the experiments performed, clearly indicating the hyperparameter values used, and conclusions about the operation of the algorithm.

Norges teknisk-naturvitenskapelige universitet

Dato
24.03.2025

Referanse
V. Singh
Alla Anohina-Naumeca

3. Based on the analysis of the operation of the algorithms, conclusions made about whether the classes in the dataset are well or poorly separable. Which algorithm perform better.

ABOUT THE REPORT:

a) Hand in only **one** file in **.pdf** format.
b) On the title page, provide a link to the created project or attach the code as appendix and link to dataset on a public website (e.g. Datalore, Google, GitHub, etc.).
c) Add comments to make the code readable.
d) The figures and tables added to the report must be numbered, explained, and referenced in the body text.
e) Include enough information (screen-dumps) to show that you have solved the assignment.
f) If anything is unclear, ask instructor to clarify and update this document if required.

Grading:

This submission on Blackboard will be your final submission for portfolio. You are not allowed to update this assignment.

You will be given a score as follows:

- 4 - very good (A or B)
- 3 - good (C)
- 2 - satisfactory (D or E)
- 1 - unsatisfactory (F)