CDS503: Machine Learning

Group Project Guidelines

PROJECT DESCRIPTION

Task 1: Group Formation

For the final project, you must work in a group of 4 members. You are free to form your own group consisting of 4 members.

Task 2: Data Set Selection

You must choose your own data set. It can be one found from an online source or one of your own. Select a structured data set and avoid from picking unstructured data such as text. Here are a few online sources where you can explore and find free data sets:

UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php)

Kaggle: https://www.kaggle.com/datasets

KDnuggets: https://www.kdnuggets.com/datasets/index.html

Try to avoid using a data set we have already worked on in class or for an earlier project. It should not be a small or made-up data set. Describe your data:

1) source of the data set

2) how is data collected

3) identify each attribute and its type

Task 3: Problem Framing

Define a problem on the data set and describe it in terms of its real-world business application. You can choose to work on supervised learning (classification, regression) or unsupervised learning (clustering, association rules).

Task 4: Data Preparation

Depending on the data chosen, preparation of the data may be one significant tasks of the project. Perform exploratory analysis on the data. You may have to perform some conversion to ensure the attribute is in the desired type. For example, if you are doing a classification task and the number of examples is very skewed among the classes, you may consider creating a subsample with reduced examples of the larger classes to make the dataset more balanced.

Task 5: Experiment Setup

In the investigation of the solution to the problem you have defined, four sets of experiments must be included. Guidelines on the experiment sets are described below. The experiment sets are different based on your choice of supervised learning or unsupervised learning.

Supervised Learning

Make sure you prepare a training set, a development set and a test set. The performance results for comparison from all the models across the four experiments should be based on the test set.

Experiment Set 1: Comparing machine learning algorithms

- Goal: Identify the most suitable machine learning algorithm.
- Explore at least 3 supervised machine learning algorithms with parameter tuning to obtain the best performing model for the data you have selected.

Experiment Set 2: Selecting features

- Goal: Find the best set of features.
- Explore at least 3 feature selection methods to identify the most relevant features for the problem.

Experiment Set 3: Ensemble learning

- Goal: Examine if ensemble learning can help improve model performance.
- Explore at least 3 ensemble learning methods to examine if model performance can be improved.

Experiment Set 4: Varying training sample size

- Goal: Examine if more training data will lead to a better performing model.
- Select the 3 best performing models from the three other experiment sets. Use the same training set for each model. You will train each model with increments of 10% following the steps below.
    i. Randomly subsample 10% of the training data and train each model (train-10P). Then, evaluate the performance of the models on the test set.
    ii. Randomly subsample another 10% from the remaining 90% of the training data. Add this second 10% sample to the first 10% sample in (i). Train another round of the models with the sample containing 20% of the training data (train-20P). Then, evaluate the performance of the models on the test set.
    iii. Repeat (i) and (ii) in 10% increments until you reach 100% of the training data. You will have a set of 10 results (10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100%) for each model.

Unsupervised Learning

Experiment Set 1: Partitional clustering (k-means clustering)

- Goal: Find the appropriate number of clusters using k-means clustering.
- Apply k-means clustering on the full dataset. Adjust the parameters to find the best clusters to group the data.

Experiment Set 2: Hierarchical clustering

- Goal: Find the clusters using hierarchical clustering.
- Apply the hierarchical clustering methods on the full dataset. Adjust the parameters to find the best clusters to group the data.

Note: Members responsible for Experiment Set 1 and Experiment Set 2 should compare the clusters produced by k-mean clustering and hierarchical clustering.

Experiment Set 3: Generate association rules from the full data

- Goal: Identify interesting association rules from the full data.
- Run association rule mining on the full dataset and pick out interesting and useful rules.

Experiment Set 4: Generate association rules from selected attributes

- Goal: Identify interesting association rules from selected attributes.
- Run association rule mining on at least 2 subsets of attributes. Define the two subsets of attributes based on attributes that are likely to relate to one another.

For a group of 4 people, plan on doing 4 experiment sets with a comparison of results and an analysis of the experiments. Each member should be responsible for one experiment set. Each experiment set should include a discussion of the results and report the best performing model within the set. In addition, there should be a discussion of results across the four sets of experiments and a conclusion summarizing the most significant findings from your overall machine learning experiments.

If you would like to combine supervised learning and unsupervised learning in your project, please consult the instructor first.

DELIVERABLES

Group Formation (Week 4: 6 November 2024 - Wednesday)

Create a post on padlet wall (https://padlet.com/jasy_yan/cds503-group-project-sem-1-2024-2025-tqp4i05lt9jgyu32), which will include the name of each group member and a one-line description of your business application (e.g., email spam filters, movie recommendation, credit card purchase fraud detection, etc.). Check the padlet first before you post your business application to make sure your selected business application does not clash with other groups. Please first make sure you can gain access to relevant data sets. Appoint a team leader, who should sign up for a padlet account so you can edit your post at the end of the semester to submit your interactive poster (Poster Presentation).

Project Proposal Submission (Week 5: 15 November 2024 - Friday)

Submit a short plan describing:

1) Task 1: Each member in the group
2) Task 2: Data set you plan to use (also submit the full or a subset of the data set)
3) Task 3: Problem definition and business application (describe the machine learning problem your group would like to address in detail)
4) Preliminary Planning: Tentatively what data preparation and machine learning experiment set each member would like to investigate

Final Report Submission (Week 15: 24 January 2025 - Friday)

Write and submit a report describing Task 1 to Task 5.

Task 1: Group Formation

Include all member names and describe the role of each group member.

Task 2: Data Set Selection

Describe your data: 1) source of the data set, 2) how is data collected, and 3) identify each attribute and its type.

Task 3: Problem Framing

Introduce the application problem and what business decisions or recommendations can be obtained from the machine learning experiments.

Task 4: Data Preparation

Describe the processing in setting up the machine learning experiments, giving details of pre-processing and filters applied to the data set.

Task 5: Experiment Setup

Describe each experiment set. Visualize and discuss the results from each experiment set and compare the results across the experiment sets in the project.

Project Presentation (Week 15: 24 January 2025 - Friday)

Each team will create an interactive poster using Adobe Spark Page to present the story and outcome of the machine learning project.

PROJECT EVALUATION

The project contributes 20% of your overall grade. The marks distribution of the project is as follows:

- Project Proposal: 2%.
- Final report: 10%.
- Project Presentation: 8%.