MiniProject 1: Machine Learning 101

COMP 551, Fall 2021, McGill University

September 14, 2021

Please read this entire document before beginning the assignment

Preamble

- **Lead TA's**; Raymond Chua, Devin Kreuzer
- This mini-project is due on **September 28** at 11:59pm EST. Late work will be automatically subject to a 20% penalty, and can be submitted up to 5 days after the deadline. No submissions will accepted after this 5 day period.
- This mini-project is to be completed in groups of three. All group members should be familiar with all steps involved in the mini-project, to avoid potential failure of the entire project due to one member's failure to do their work. All members of a group will receive the same grade. It is not expected that all team members will contribute equally to all components. However every team member should make integral contributions to the project.
- Submit your assignment on MyCourses as a group. You must register your group on MyCourses and any group member can submit. See MyCourses for details.
- You are free to use libraries with general utilities, such as matplotlib, numpy, scipy, pandas and sklearn for Python, with the exceptions clearly stated below.

Background

In this miniproject, you will be exploring two datasets. The goal is to gain experience in deploying basic supervised machine learning techniques to tackle a real-world data science problem. In particular, the project encourages you to explore preprocessing of the data, the effect of hyper-parameters, size of the dataset, and performing model selection. You are encouraged to explore techniques you have learned in class to visualize the data and thereafter form a hypothesis about possible patterns in the data.

Preprocessing

Your first task is to acquire the data, analyze it, and clean it (if necessary). You will use two datasets in this project, outlined below.

- **Dataset 1 (Adult dataset)**: This dataset presents several attributes of different individuals and the prediction task is to determine whether someone makes over 50K a year. Download and read information about the dataset here.
- **Dataset 2 (Your choice!)**: Select any dataset from UCI or related to your own research. We suggest selecting a dataset of appropriate size (not too small or too large) such that the experiments can be conducted effectively and efficiently.

The essential subtasks for this part of the project include:

1. Download the datasets. *Hints: For clarity, in the Adult dataset, adult.data contains the training/validation data and adult.test contains the test data.*
2. Load the datasets into Pandas dataframes or NumPy objects (i.e., arrays or matrices) in Python.
3. Clean the data. You should remove instances that have too many missing or invalid data entries.

4. Convert discrete variables into multiple variables using one-hot encoding. For an example on how to do this, check out ”Encoding categorical features” in the scikit-learn documentation.

Experiments

In this part, you will compare two supervised learning frameworks, namely *K-nearest neighbours* (KNN) and *decision trees*, to predict whether the income of an adult exceeds $\$50K/yr$. A similar analysis should be performed for the second dataset. The specific subtasks for this part include:

1. Implement and perform 5-fold cross validation on the training/validation data (for the *Adult* dataset, this data is contained in the *adult.data* file) to optimize hyperparameters for both models. **Your implementation for cross-validation should be from scratch. You should not use existing packages for cross validation.** Report the mean of the training and validation metrics for the given hyperparameters.

2. Sample growing subsets of the training/validation data and repeat step 1. We want to understand how the size of a dataset impacts both the training and validation error.

3. Take the best performing model (the one with the best performance on 5-fold cross validation) and apply it on the test set (in the *Adult* dataset, this is the *adult.test* file). This is an unbiased estimate of how your model would perform on new/unseen data.

4. [**Optional**] Go above and beyond! Examples: different normalization techniques or other ways of handling of missing data (search “data imputation” techniques). Employ more sophisticated techniques for hyper-parameter search. Engineering new features out of existing ones to get a better performance. Investigate which features are the most useful (e.g., by correlating them with your predictions or removing them from your data)?

5. Analyze your findings; how did the choice of the various hyper-parameters impact generalization? How about the size of training data? If any of these findings do not agree with your expectation, you can form hypotheses and further investigate them.

Deliverables

You must submit two separate files to MyCourses **(using the exact filenames and file types outlined below)**:

1. **code.zip**: Your entire code, which should consist of a jupyter notebook file (.ipynb), and additional python files (.py); **the notebook should contain the main body of your code, where we can see and easily reproduce the plots in your report.**

2. **writeup.pdf**: Your (max three pages) project write-up as a pdf (details below).

Project write-up

Your team must submit a project write-up that is a maximum of three pages (**single-spaced, 11pt font or larger; minimum 1 inch margins, an extra page for references/bibliographical content can be used**). We highly recommend that students use LaTeX to complete their write-ups. You have some flexibility in how you report your results, but you must adhere to the following structure and minimum requirements:

Abstract (100-250 words)

Summarize the project task and your most important findings. For example, include sentences like “In this project we investigated the performance of two classification models, namely k-nearest neighbours and decision trees, on predicting if the income of an adult exceeds $\$50K/yr$ from various factors, such as age, sex, nationality, etc...”, “We found that the k-nearest neighbour regression approach achieved worse/better accuracy than decision trees and was significantly faster/slower to train.”

Introduction (5+ sentences)

Summarize the project task, the two datasets, and your most important findings. This should be similar to the abstract but more detailed. You should include background information and potential citations to relevant work, if any, (e.g., other papers analyzing these datasets).

Datasets (5+ sentences)

Very briefly describe the datasets and how you processed them. How did you handle the missing data? If you have come up with new new features to get better results, you should explain it here. Present your efforts for better understanding of the data, e.g. through visualization plots.

Results (7+ sentences, possibly with figures or tables)

Describe the results of all the experiments as well as any other interesting results you find. Elements we expect to see:

1. Comparing performances between KNN and decision trees
2. Revealing how changing hyperparameters affects performances for both models
3. Describe how reducing the amount of data impacts results

Discussion and Conclusion (5+ sentences)

Summarize the key takeaways from the project and possibly directions for future investigation.

Statement of Contributions (1-3 sentences)

State the breakdown of the workload across the team members.

Evaluation

The mini-project is out of 10 points, and the evaluation breakdown is as follows:

- Completeness (2 points)
  - Did you submit all the materials?
  - Did you run all the required experiments?
  - Did you follow the guidelines for the project write-up?
- Correctness (4 points)
  - Are you cross-validation schemes implemented correctly?
  - Are your models used/implemented correctly?
  - Are you visualizations informative and visually appealing?
  - Are your reported accuracy close to (our internal) reference solutions?
- Writing quality (2.5 points)
  - Is your report clear and free of grammatical errors and typos?
  - Did you go beyond the bare minimum requirements for the write-up (e.g., by including a discussion of related work in the introduction)?
  - Do you effectively present numerical results (e.g., via tables or figures)?
- Originality / creativity (1.5 points)
  - Did you go beyond the bare minimum requirements for the experiments?
  - within the context of producing the required results did you propose a creative idea?
  - **Note:** Simply adding in a random new experiment will not guarantee a high grade on this section! You should be thoughtful and organized in your report in explaining why you performed an additional experiment and how it helped in evaluating your hypothesis.

Final Remarks

You are expected to display initiative, creativity, scientific rigour, critical thinking, and good communication skills. You don't need to restrict yourself to the requirements listed above - feel free to go beyond, and explore further

You can discuss methods and technical issues with members of other teams, but **you cannot share any code or data with other teams**.