COMP5318 - Machine Learning and Data Mining

Assignment 2

Due: 29 Oct 2018 5:00PM

This assignment is to be completed in groups of 2 to 3 students. It is worth 20% of your total mark. Your groups can be different from Assignment 1.

1. Objective

The objective of this assignment is to apply machine learning and data mining methods to solve a real problem. You should compare at least three techniques with at least one, not taught in this course (eg. adaboost, random forest, support vector regression, etc).

2. Instructions

2.1 Datasets

In this assignment you can choose one of the following datasets:

- Cifar100, classification, https://www.cs.toronto.edu/~kriz/cifar.html
- Chars74K-EnglishImage, classification, http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/
- MNIST-fashion, classification, https://github.com/zalandoresearch/fashion-mnist
- Adult, classification, https://archive.ics.uci.edu/ml/datasets/adult
- Airline delay, regression, https://www.kaggle.com/giovamata/airlinedelaycauses
- Diabetes, time-series, classification, https://archive.ics.uci.edu/ml/datasets/Diabetes

Note that if the datasets are too big to run, you can consider doing some preprocessing of the data or use part of them to train. However, those should be clearly explained in your report.

2.2 Assignment tasks

a) Choose a data set from the list above.

b) Try different Machine Learning methods (at least 3) and compare their performance. At least one of the techniques you use should be not covered in the course material. To this end, clearly discuss your design choices to achieve higher performance and speed. Design options can be at least of four-fold:

- Choosing an appropriate model and its complexity
- Using preprocessing techniques on the dataset (e.g. clustering, feature extraction, etc.)
- Computer infrastructure (e.g. parallelizing, speeding-up your code, etc.)
- Ease of prototyping (e.g. choice of the programming language and existing libraries)

c) You are expected to fine tune each algorithm and explain why one approach outperforms the other.

d) Since you are expected to use more complex models that have not been discussed in lectures, you can use most external open-source libraries such as: scikit-learn, pandas, Keras, Tensorflow, PyTorch, Theano, Caffe2, or their equivalent in Python 3. Should you require to use any other external library, please post on Canvas.

e) You are allowed to only use Python 3 in this assignment.

3. Report

The report must be organised in a similar way to research papers, and include the following:

- In the **Abstract**, succinctly describe the rest of your report
- The **introduction** section should present the dataset that you chose, discuss its relevance in diverse applications, and give an overview of the methods you used.
- You are expected to include a section on **previous work**, listing successful techniques on similar datasets
- The next section should discuss the **methods** you used. Explain the theory behind each of them and discuss your design choices. This part should at least include preprocessing and machine learning techniques used.
- The **experiment** section displays results and comparisons for the previously introduced methods. Include runtime, hardware and software specifications of the computer that you

used for performance evaluations. You are then expected to include meaningful comments on the results of your experiments, and reflect on design choices.

- In **conclusion**, sum up your results and provide meaningful future work.
- The **references** section includes all references cited in your report, formatted in a consistent way.

3.1 Evaluation metrics

You should compare the algorithms with a 10-fold cross validation exercise.

Classification task: When evaluating different classifiers, include accuracy, precision, recall and confusion matrix.

Regression task: For regression problems, include Mean Square Error (MSE) and Negative Log Likelihood for the predictions (NLL):

$$NNL = -\log p(y_*|D, \mathbf{x}_*) = \frac{1}{2}\log(2\pi\sigma_*^2) + \frac{(y_* - \bar{f}(x_*))^2}{2\sigma_*^2}$$

where y_* is the actual value to be predicted, D is the training dataset, x_* is a query poing, and $\bar{f}(x_*)$ and σ_*^2 are the prediction mean and variance respectively.

3.2 Report layout

Please use the provided MS-word or LaTeX template.

Length: Ideally 10 to 15 pages - maximum 25 pages with [-10] penalty for each additional page after 25.

4. Submissions

The report and code are due on *29 Oct 2018, 5:00 PM*.

4.1 Go to Canvas and upload the following files/folders compressed together as a zip file.

a) Report (a pdf file)

The report should include each member's details (student ID and name)

b) Code (a folder)

Your code (could be multiple files or a project). Do NOT include the dataset. Include a readme file to describe how to run the code.

4.2 Only one student in your group needs to submit the zip file which must be named as student ID numbers of all group members separated by underscores. E.g. "xxxxxxxx_xxxxxxxx_xxxxxxxx.zip"

4.3 Your submission should include the report and the code. A plagiarism checker will be used.

4.4 Clearly provide instructions on how to run your code in the appendix of the report.

4.5 Clearly provide the hyperlinks to the dataset you used, external open-source libraries you used for the analyses, and the version of libraries, e.g., pytorch 0.4.

4.6 Indicate the contribution of each group member.

4.7 A penalty of MINUS 20 (twenty) percent per each day after the due date. Maximum delay is 5 (five) days, after that assignments will not be accepted.

4.8 Remember, the due date to submit them on Canvas is **29 October 2018, 5:00PM**.

5. Marking scheme

Category	Criterion	Marks	Comments
Report [80]	Abstract [3] -problem, significance, methods, results and conclusions		
	Introduction [5] - What's the problem you intend to solve? - Why is the problem important?		
	Previous work [10] - Previous relevant methods used in literature		
	Methods [25] - Theory on different techniques compared - Pre-processing - Design choices		

	Experiments and Discussion [25] - Experiments, camparisons and evaluation - Meaningful discussion of results and design choices - Relevant personal reflection		
	Conclusions and future work[3] - Meaningful conclusions based on results - Meaningful future work suggested		
	Presentation [5] - Grammatical sentences, no spelling mistakes - Good structure and layout, consistent formatting - Appropriate citation and referencing - Use graphs and tables to summarize data		
	Other [4] - At the discretion of the marker: for impressing the marker, excelling expectation, etc. Examples include fast code, high accuracy, etc.		
Code [20]	Attempts to speed up the program [7]		
	Code runs and classifies within a feasible time [7]		
	Well organized, commented and documented [6]		
Penalties [-]	Badly written code: [-20]		
	Not including instructions on how to run your code: [-20]		
	Late submission: [-20] for each day late		
	Not using Python3 [-100]		

	No contribution to the group work: [-100]		
	Long report penalty [-10] for each additional page after 25 pages.		

Note: Marks for each category is indicated in square brackets. The minimum mark for the assignment will be 0 (zero).