COMP5318 - Machine Learning and Data Mining

Assignment 2

Due date: 5^{th} of June 2017, 5PM

This assignment is to be completed in groups of 2 to 3 students. It is worth 20% of your total mark. To facilitate group allocation, please fill the following form: `https://docs.google.com/spreadsheets/d/13o4shdPQv5cv7W7EZvTnLv8KnOWRi3zA3zhVaqliwC0/edit?usp=sharing`

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:

- Cifar10 (vision dataset), classification, `https://www.cs.toronto.edu/~kriz/cifar.html`
- SVHN (vision dataset), classification, `http://ufldl.stanford.edu/housenumbers/`
- Covtype, classification, `https://archive.ics.uci.edu/ml/datasets/Covertype`
- 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 because not all datasets are of same complexity, performance marks will be scaled accordingly.

2.2 Assignment tasks

1. Choose a data set from the list above.
2. 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)

3. You are expected to fine tune each algorithm and explain why one apporach outperforms the other.

4. 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, Theano, Caffe, or their equivalent in other languages. Should you require to use any other external library, please post on Edstem.

5. You are allowed to use one of the following languages:
Python, Cython, Matlab, R, C/C++, Julia or Java.

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 cites 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):

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

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

4 Submissions

The report and code are due on the 5^{th} of June 2017, 5PM.

1. Go to eLearning and upload the report (.pdf) and the code compressed together as a zip file. Do NOT include the dataset.

2. Only one student 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_xxxxxxxx_.zip".

3. Your submission should include the report and the code. A plagiarism checker will be used. Clearly provide instructions on how to run your code in the appendix of the report.

4. Clearly provide the hyperlinks to the dataset you used and external open-source libraries you used for the analyses.

5. Indicate the contribution of each group member.

6. There is no special format to follow for the report but please make it as clear as possible and similar to a research paper.

7. A penalty of MINUS 7 (seven) points per each day after the due date. Maximum delay is 7 (seven) days, after that assignments will not be accepted.

8. Remember, the due date to submit them on eLearning is 5^{th} of June 2017, 5PM.

5 Marking scheme

Category	Criteria	Marks	Comments
Report [80]	Abstract [3] •Problem, significance, methods, results and conclusions. Introduction [5] •What is the problem you intend to solve? •Why is this 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, comparisons 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, using LaTeX, etc.		
Code [20]	•Attempts to speed up the program •Code runs within a feasible time •Well organized, commented and documented		
Penalties [−]	•Badly written code: [−20] •Not including instructions on how to run your code: [−30] •Late submission of report: [−7] for each day late up to seven days •Not contributing to the group work: [−100]		

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