

TRINITY COLLEGE DUBLIN
School of Computer Science and Statistics

Week 5 Assignment

CS7CS4/CSU44061 Machine Learning

ASSIGNMENT: WRITE A SHORT GROUP PROJECT PROPOSAL

Your task is to pick a project topic and write a short description of your proposed project. There are two main types of project:

1. *Application project* (most common). Pick an application that interests you, and explore how best to apply learning algorithms to solve it.
2. *Algorithmic project*. Pick a problem or family of problems, and develop a new learning algorithm, or a novel variant of an existing algorithm, to solve it.

Projects can of course also combine elements of applications and algorithms. It's often a good idea to either pick an application area that you're interested in or a machine learning technique that you want to explore more. Try not to be over-ambitious though - its far too easy to end up spending a lot of time on a project. Bear in mind that the project is worth 30% of the overall marks and 2-3 people are working together on it, so use that to calibrate the amount of time and effort you put into it.

Your project must include the following components:

1. Gathering of raw data. For example, by scraping it yourself off the web (from twitter, news etc) or taking it from open data sources (many countries, including Ireland have open data initiatives with e.g. live city data, social and health data). Do *not* use pre-packaged data from Kaggle, UCI or the like - its important to get experience of working with raw data.
2. Pre-processing/cleaning up of the raw data and feature engineering/selection.
3. Model selection. Try to use more than one significantly different type of model e.g. a logistic regression and kNN classifier (but not logistic regression and SVM classifier, unless the SVM classifier is kernalised).
4. Performance evaluation. This must include comparison against a reasonable baseline and, importantly, a critical discussion/reflection on results.

DOCUMENTS

For the project you need to generate two documents. The first is a short project proposal, which is this assignment. Then later you need to write the project report. Both the project proposal and project report should be submitted on Blackboard.

The project proposal document should give the title of the project, the full names of the team members and their student IDs and a 300-500 word description of what you plan to do. The description should include:

1. Motivation: What problem are you tackling?
2. Dataset: What data will you use and how will you collect it?
3. Method: What machine learning techniques are you planning to apply or improve upon?
4. Intended experiments: What experiments are you planning to run? How do you plan to evaluate your machine learning algorithm?

As a guide, the final report should be about 5 pages long (including figures and tables)
- excessively long reports may be penalised as its important to learn to present results succinctly yet clearly. Generally it should consist of the following sections:

1. Introduction (about 0.5 page)

- Explain the problem and why it is interesting.
- State what the input and output are, e.g. “The input to our algorithm is an {image, amplitude, patient age, rainfall measurements, grayscale video, etc.}. We then use a {SVM, neural network, linear regression, etc.} to output a predicted {age, stock price, cancer type, music genre, etc.}.”

2. Dataset and Features (about 1 page)

- Describe your dataset. How/where did you gather the data, how many data points are there, what preprocessing did you do, what normalization or data augmentation. Show some examples from your dataset.
- Discuss how you mapped the data to features. Explain the rationale.

3. Methods (about 1 page)

- Describe the learning algorithms that you used. For each algorithm, give a short description (a paragraph) of how it works - we are looking for your understanding of how these machine learning algorithms work. This is especially important if you use methods not covered in the module since using methods in a black box fashion is terrible practice and you need to clearly communicate that you have not done that.

4. Experiments/Results/Discussion (about 2-3 pages)

- Give details about what parameters and hyperparameters you chose, and how you chose them. Be sure to use cross-validation to select hyperparameters and to state the number of folds used. How many training/validation/test examples did you use? Explain whether you think you may have overfit your training set and what, if anything, you did to mitigate that. For model parameters did you use gradient descent or some other algorithm?
- Before you describe your results explain what your primary metrics are: accuracy, precision, AUC, etc.
- For results, you want to have a mixture of tables and plots. If you are solving a classification problem, you should include a confusion matrix and AUC/ROC curves. Include performance metrics such as precision, recall, and accuracy. For regression problems, state the average error. Make sure to refer to and discuss the figures/tables in your main text throughout this section (no standalone plots please!). Your plots should include legends, axis labels, and have font sizes that are legible when printed.
- Be sure to compare performance against a reasonable baseline - this is mandatory.
- You should have a mix of both quantitative and qualitative results.

5. Summary (100-200 words)

- Summarize your report and reiterate key points e.g. which algorithms were the highest performing, why you think that some algorithms worked better than others.

6. Contributions. This doesn't contribute to the 5 page limit.

- Describe what each team member worked on and contributed to the project. It's important to give a decent level of detail e.g. what code was written by whom, which experiments were carried out by whom, who wrote each part of the report. Each team member needs to initial this section (electronic initials are fine).
7. Include a github link (or similar) to the code written as part of the project.

NOTES

- *Deep learning.* If you decide to work on a deep learning project, please make sure that you use other material you learned in the class as well. For example, you might set up logistic regression or SVM as a baseline for evaluation. Also, be aware that training deep learning models can be very time consuming (don't underestimate this!) and make sure you have the necessary computing resources.
- *Using methods not covered in class.* It's ok to use ML methods not covered in the class but you must take care to properly demonstrate in your report that you really do know how the methods work – it's far too easy to just use tools in a black box way but that's really poor practice and is a good way to haemorrhage marks.
- *Pre-prepared datasets.* While we don't want you to have to spend much time collecting raw data, the process of inspecting and visualizing the data, trying out different types of pre-processing, and doing error analysis is an important part of machine learning. We therefore ask that you do not use pre-prepared datasets (e.g. from Kaggle, the UCI machine learning repository, etc.).
- *Replicating results.* Replicating the results in a paper or online post can be a good way to learn. However, instead of just replicating previous work also try using the technique on another application or do some analysis of how each component of the model contributes to final performance.
- *Team working.* The project involves a team because working with raw data often involves a decent amount of effort to gather and clean up the data. A team allows this work to be shared and so leave enough time to focus on the more interesting stuff e.g. results and discussion. However, we understand that problems can arise when working in a team. With that in mind:
 - Do keep a record of the work carried out by the individual team members. You'll need to detail this in the final report and it will help if any problems arise.
 - If team related problems do occur during the course of the project, do contact us as early as possible. We can, for example, re-organise teams or take other steps but those are often best done relatively early during the course of the project.