Advanced Machine Learning
Coursework Assignment

Dr Douglas McIlwraith

Tuesday 28th March, 2017

This document outlines the requirements for the coursework which accompanies the Advanced Machine Learning course, February/March 2017. This coursework accounts for 60% of your total grade, with the remaining 40% taken from your participation in the weekly tutorials.

This coursework is to be performed individually although you are free (and encouraged!) to discuss your approaches with contemporaries. The purpose of this exercise is to give you the opportunity to apply relevant techniques introduced within the lectures, and to delve deeper into them – both to reinforce and to check your understanding.

Please choose ***only one of the following problems.*** Each problem has a different objective, however, there should be enough scope to choose and execute a technique (or several) in order to achieve the specified objective. The problems are listed below in order of increasing difficulty/involvement and are all taken from *kaggle.com* – where further information along with the datasets themselves can be found.

Easy: The Titanic Dataset

In this problem, we ask you to investigate the sinking of the Titanic. You must try to predict if a passenger is likely to survive given information about them (class of travel, gender etc.) Pay attention to the evaluation metric [1]. Submit your classified test set to Kaggle regularly. Design features. How do subsequent approaches improve classification accuracy?

> The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.
>
> One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew.

> Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.
>
> In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy. (*https://www.kaggle.com/c/titanic*)

Medium: House Price Regression

Here we ask you to investigate the Ames residential homes dataset [2]. You are given many descriptors of residential homes and their final price. It is your job to build a regressor to predict the final price of a home. In this case, the evaluation metric is the RMS of log error [3], i.e.

$$RMSLE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(log(p_i+1) - log(t_i+1)\right)^2} \tag{1}$$

where p_i is the prediction price and t_i is the ground truth price for data point i. [1]

> Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.
>
> With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home. (*https://www.kaggle.com/c/house-prices-advanced-regression-techniques*)

Using a method (or methods) of your choosing, predict the final price of houses in the test set. Pay attention to the evaluation metric and submit to Kaggle regularly. What is your best RMSLE? Why does (or doesn't) your approach work?

Hard: Digit Recognizer using the MNIST Dataset

For this project, you will attempt to perform digit classification using the MNIST dataset [5]. This dataset contains 70,000 samples of hand written digits between 0 and 9. Images are 28px by 28px, with each pixel containing an

[1] This is used because it is equivalent to penalising the ratio between predictions, not the absolute difference. Observe that $log(p_i+1) - log(t_i+1) = log(\frac{p_i+1}{t_i+1})$ Consequently a \$50,000 dollar error with respect to a \$500,000 house will be penalised by the same amount as a \$500,000 error on a house that sold for \$5M. Virtual counts are introduced to allow for true values of 0.

intensity value between 0 and 255 inclusive.

Models are to be trained on a 42,000 sample dataset, and tested on a 28,000 sample test set. Approaches are evaluated for accuracy over the test set. Details of the submission procedure can be found online [4] - submit regularly!

> MNIST ("Modified National Institute of Standards and Technology") is the de facto hello world dataset of computer vision. Since its release in 1999, this classic dataset of handwritten images has served as the basis for benchmarking classification algorithms. As new machine learning techniques emerge, MNIST remains a reliable resource for researchers and learners alike.
>
> In this competition, your goal is to correctly identify digits from a dataset of tens of thousands of handwritten images. Weve curated a set of tutorial-style kernels which cover everything from regression to neural networks. We encourage you to experiment with different algorithms to learn first-hand what works well and how techniques compare. (https://www.kaggle.com/c/digit-recognizer)

Researchers have been working on this particular dataset for some time and as such there's a wealth of research available. Yann Le Cun *et al.* [5] have provided an overview of results dating back to 1998.

What techniques did you choose, how do you rank against previous efforts? Can you reproduce the research of others? What would you do next if you had more time? This is your opportunity to shine! Can you obtain results which are 'state of the art?'

Method of Assessment

The method of assessment will be a written, technical report. The report should be aimed an audience of contemporaries and should not be overly long. Be succinct and accurate.

Discuss your approach, what libraries did you use? What algortihms do they implement for learning? Why does your overall approach work? Do you have any hypotheses? State them. Prove/disprove them. What work would you like to carry out in the future?

Marks will be awarded based upon (and in order of the most points available):

1. A clear, concise report detailing your work and providing evidence that you understand the applied methods.
2. Your selection and understanding of the most appropriate techniques for the problem at hand.

3. Demonstration of an improving approach, as defined by the evaluation metric in each competition.

Good luck!

References

[1] http://bit.ly/2nvFFp9.

[2] D. D. Cock, "Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project," *Journal of Statistics Education*, vol. 19, no. 3, 2011.

[3] http://bit.ly/2nckqpe.

[4] http://yann.lecun.com/exdb/mnist/index.html.

[5] http://bit.ly/2nqU5Wc.