CM4107 AI Coursework 2

Nirmalie Wiratunga Kyle Martin Sadiq Sani

Outline

Aim of this coursework is to make use of the scikitLearn library to conduct a comparative study on algorithm performance on a sample of datasets. We will be using text classification datasets, which calls for a deep understanding of the importance of data quality and transformation. A secondary aim of this coursework is to explore your understanding and awareness about the topic of "AI matters". Specifically to focus on issues that are important for explainable AI and ethical AI and to critically appraise and draw from relevant literature about these two topics.

Submission deadline: 30th Nov 5pm

Weight Allocation: This coursework comprises of 3 tasks and carries 50% of the weight towards the final grade Weights assigned to each task is an indication of its individual contribution towards the final grade.

Feedback: Students will be given initial verbal feedback in the form of an outline / expected solution one week after submission of Part 1 (14th Dec). A provisional grade will be made available a week after that on Moodle. An outline solution will be posted on 14 Dec. Written feedback and provisional grade for each student will be made on Moodle.

1 Part 1 - Comparative Study using ScikitLearn

The goal is to analyse a collection of text documents and build a text classifier. You will need to use sciktlearn and nltk to extract feature vectors suitable for machine learning and train models to perform text classification. You can also use a grid search strategy to find a good configuration of both the feature extraction components and the classifier.

You should select 2 datasets from the the list provided. You need to select 3 algorithms from the sciktlearn library; two of which must be an Artificial Neural Net and the k-Nearest Neighbour (kNN). The third can be one of your own choices from the sciktlearn library (e.g. Naive Bayes, Support Vector Machines, Decision Trees). It is important that you select and organise your training and test sets appropriately using cross-validation.

Proposed candidate datasets listed below:

- Movie reviews
- Spam Dataset
- 20 Newsgroups

Task (weight 3): Maximum one page (including references) which needs to be organised as follows:

- explain your evaluation strategy and the pipeline that was used to convert the text data into a vector form; and
- present and interpret your results.

2 Part 2 - Explainable AI

What requirements, if any, should be imposed on AI systems and technology when making decisions that directly affect humans? For example, should they be required to make transparent decisions? If so, how?

In this section, you should provide a brief overview of explainable AI in general, including: what is explainable AI? Why is it important? What are some of the legal, ethical and social issues surrounding explainable AI?

You should then focus on the specific qualities that apply to your selected algorithms. This should include references to existing literature.

Task (weight 1): Maximum one page (including references) which needs to be organised as follows:

- A brief overview of explainable AI and its importance.
- The legal, ethical and social implications of explainable AI.
- In what ways can you explain the decisions of your selected algorithms?

Some literature that you may wish to read for ideas on the subject (you can Google and download these from the web). Feel free to explore other sources, but please reference appropriately.

- Sørmo, F., Cassens, J., Aamodt, A.: Explanation in case-based reasoning - perspectives and goals.

  Although CBR focused, this paper presents an excellent overview of explainability in modern AI. Particularly, it examines how to measure different aspects of explainability in an algorithm.

- Miller, Tim et al. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences.

A very interesting paper discussing the human connotations of explainable AI, as well as advocating taking advantage of existing research in other areas like philosophy to inform the next steps for explainability.

- IJCAI/ECAI Workshops on Explainable Artificial Intelligence (XAI)

  This is a collection of papers submitted to IJCAI's workshop on explainability. IJCAI is one of the most prominent AI conferences in the world, and their workshops contain a vast amount of relevant quality papers discussing explainability.

3 Part 3 - Ethical AI

All knowledge and tools, including AI, can be used for good or for bad. This is why it's important to think about what AI is, and how we want it to be used.

In this section you will focus on the ethical aspects of AI. You are to carry out a mini-literature review on the topic and present your findings with focus on What are the ethical issues with AI? Are they the same issues as we have with other artifacts we build and value or rely on, such as music or parks?

Task (weight 1): Maximum one page (including references) which needs to be organised as follows:

- What were the key ethical issues that you consider to be important and in particular what role do you think that AI should play in society?
- What might be the risks from AI that we need to be aware of?
- What about machine morality; When computers make errors are they to blame?

Some literature that you may wish to read for ideas on the subject (you can Google and download these from the web). Feel free to explore other sources, but please reference appropriately.

- Patience Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics, (2018) in Ethics and Information Technology 20(1)19-26. An earlier version appeared the AISB 2012 proceedings below and the AAAI Spring Symposium Ethical and Moral Considerations in Nonhuman Agents. Both AI and ethical systems are cultural artefacts, so whether AI is a moral subject (can take responsibility or be something we're obliged to) is a matter of choice, not something that needs to be "discovered." https://link.springer.com/article/10.1007/s10676-018-9448-6
- Joanna J. Bryson, Mihailis E. Diamantis, and Thomas D. Grant, Of, For, and By the People: The Legal Lacuna of Synthetic Persons. Artificial Intelligence and Law 25(3):273–291 [Sep 2017]. A discussion by AI and law as to whether it would be a terrible idea to make something strictly

AI (in contrast to an organisation also containing humans) a legal person. They provide examples and discusses issues about where legal personhood has already been overextended.

- Robert H Wortham, Andreas Theodorou, Joanna J Bryson, Improving Robot Transparency: Real-Time Visualisation of Robot AI Substantially Improves Understanding in Naive Observers. In The 26th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2017. An earlier version was in the IJCAI-2016 Ethics for Artificial Intelligence Workshop.
https://goo.gl/MCtcqR

- Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan, Semantics derived automatically from language corpora contain human biases. Science 356 (6334):183-186 [14 Apr 2017]. Be sure to also look at the supplement, which gives the stimuli and shows similar results for a different corpus and word-embedding model. Meaning really is no more or less than how a word is used, so AI absorbs true meaning, including prejudice. This is demonstrated empirically and shows machine learning replicates our biases in AI. Open access version
http://randomwalker.info/publications/language-bias.pdf

- David Gunkel and Joanna J. Bryson (eds.), Introduction to the Special Issue on Machine Morality: The Machine as Moral Agent and Patient, Philosophy Technology, 27(1):5–8, March 2014. This derived from The Machine Question: AI, Ethics and Moral Responsibility, David J. Gunkel, Joanna J. Bryson and Steve Torrance, (eds.). A symposium proceedings published by The Society for the Study of Artificial Intelligence and Simulation of Behaviour, 3-5 July, 2012.
https://link.springer.com/article/10.1007/s13347-014-0151-1

- Just an Artifact: Why Machines are Perceived as Moral Agents, with Philip P. Kime, in the proceedings of The Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI '11).
http://www.cs.bath.ac.uk/ jjb/ftp/BrysonKime-IJCAI11.pdf

4 How to Submit

Please follow these guidelines:

- Submit by the due date

- The write-up including figures, tables and references must use the IEEE style file template provided (cm4017-cw2-template.doc - you can download this from the coursework section).

- All files should be submitted as a single zip file named using your surname-firstname-matriculation.

- Use the dropbox to submit your coursework. This consists of your write-up using the Word template file (max 3 pages); plus any ipynb files that evidence your coding effort.
- Please also use Turnitin to evidence that you submission contains non plagarised material.

5 How grades are aggregated from CW1 & CW2

Figure 1: Aggregation of grades from CW1 and CW2.

				CW2				
		A	**B**	**C**	**D**	**E**	**F**	**NS**
	A	A	A	B	B	D	D	D
	B	A	B	B	C	D	D	D
	C	B	B	C	C	D	D	D
CW1	**D**	B	C	C	D	D	E	E
	E	D	D	D	D	E	E	E
	F	D	D	D	E	E	F	F
	NS	D	D	D	E	E	F	NS