Using Machine Learning and AI in Social Sciences

Peter Ormosi

Norwich Business School, and
Centre for Competition Policy, University of East Anglia

CCP Annual Conference
6 June 2019

Democratisation of ML/AI

Erosion of entry barriers:

- 5 years ago, to do machine learning, had to build everything from scratch
- today, deep learning frameworks like tensorflow (or TF2.0), PaddlePaddle, keras, theano, torch, or caffe to design a neural network for your own application.
- today we have massive word embedding trained data available (glove, word2vec, etc) - 1-100B words
- today: running NN in Google Cloud

Democratisation of ML/AI

Erosion of entry barriers in research - raising the profile of the little guys:

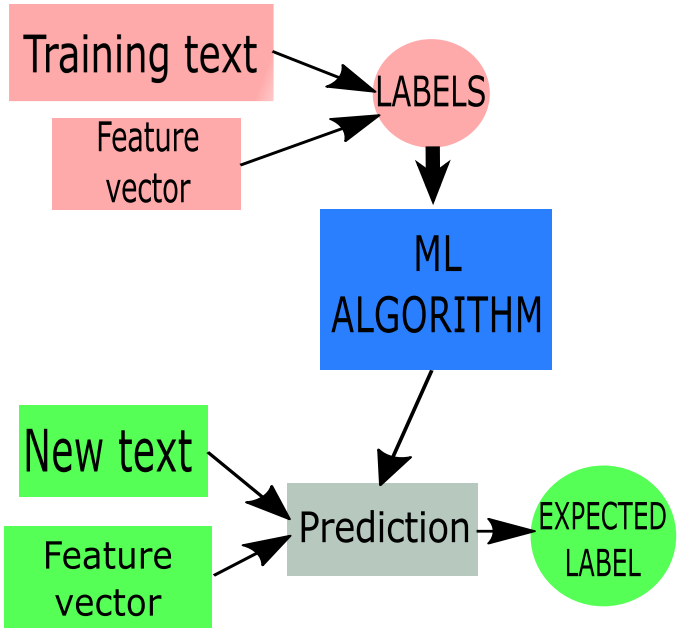
- 5 years ago, large team of RAs would have been needed
- today, much of this can be automated

HOW DO I USE ML?

ML for data collection

- Much of what follows is about **natural language processing**
- Extracting structured quantitative data from unstructured text data
- Two upshots:
  - New avenues for research (new types of data)
  - Automating a lot of human work

Text classification



- Labelling text based on their content
- E.g. relevant patents

Sentiment analysis

- Extracting information from text data
- Analysis of customer reviews
- E.g. Estimating the reputational impact of corporate misbehaviour

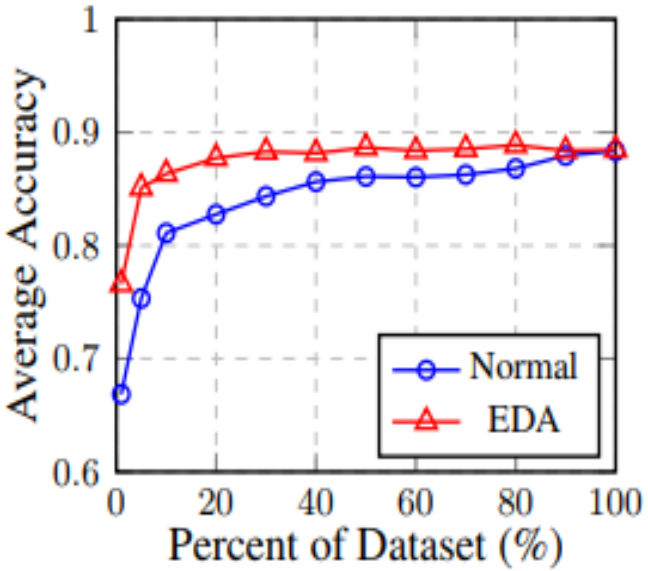
Named entity recognition

Asda could be listed on the stock market after its merger with supermarket rival Sainsbury's was blocked by the competition authorities. Judith McKenna, chief executive of Asda's owner Walmart, has told staff such a listing is being considered. But, she told managers at an event in Leeds - where Asda is based - any listing could "take years".

ORGANISATION **PERSON** **PLACE**

- Useful when working with text data
- Which businesses/brands are mentioned in a text
- LSTM, bidirectional neural network
- Simple frameworks available, e.g. NLTK Python package
- Somewhat related: coreference resolution

Data augmentation



- When we don't have enough data
- Most obvious use for image data
- Applications for text data
  - Synonym replacement
  - Tokenise text by sentences and shuffle sentences
  - Randomised synonym insertion
- Risk of overfitting

IDENTIFYING RENT SEEKING IN LEGISLATION

Rent seeking

"We are suffering from the ruinous competition of a foreign rival who apparently works under conditions so far superior to our own for the production of light that he is flooding the domestic market with it at an incredibly low price; for the moment he appears, our sales cease, all the consumers turn to him, and a branch of French industry whose ramifications are innumerable is all at once reduced to complete stagnation. This rival, which is none other than the sun ..."

Can we identify rent seeking?

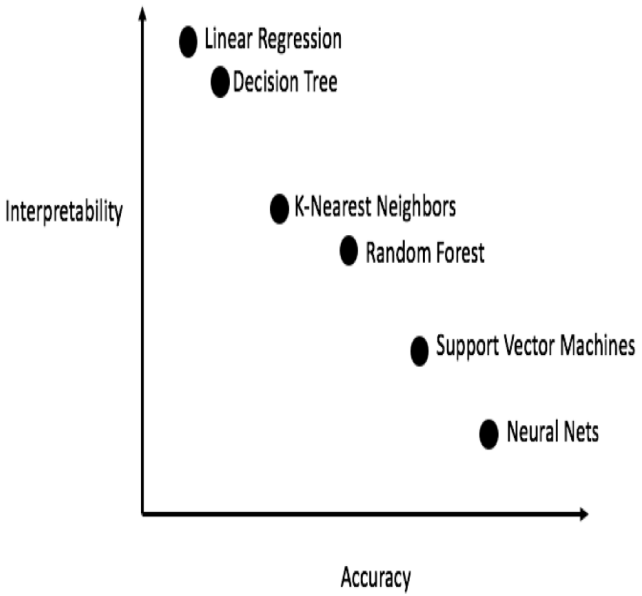
- Text classification problem. Has a piece of law been lobbied?
- Would a human be able to do this?
- We have some labelled data, plus features (e.g. date, subject, length)
- We use an LSTM model to predict the incidence of lobbying
- We then run this model on the complete corpus of US bills and get a propensity of lobbying/rent seeking in each calendar quarter.

What are the main issues?

We are currently achieving 90% accuracy but there are issues:

- Interpretability: Can we interpret our results? What drives whether a law has been lobbied?
- Can we transfer this knowledge to other areas?
- How could we have more labelled data? This could improve accuracy.
- Do we need an end-to-end solution?

Interpretability of machine learning models



- NN (and even SVM) can be like a black box.
- We have input layers, and output layer, parameters, and their estimated weights
- but we don't know which features contributed most to these weights.
- What can we do?
  - LIME? (Ribeiro et al 2016)
  - layerwise relevance propagation? (Bach et al 2015)

Transfer learning

We have labelled data for the US, but what about applying it to other English speaking jurisdictions?

- Instead of training a new NN for the new corpus of laws we could
  - Take an NN trained on a different domain/task
  - Adapt it for own domain

Active learning

What can we do when we don't have enough labelled data?

- labelling is very costly
- with active learning we start off with a small number of labels
- then new labels added for most important (most important for the classification problem) datapoints.

For a review: Miller et al (2018) Active Learning Approaches for Labeling Text

End-to-end vs multiple stages

- Are we right in pursuing an end-to-end solution?
- End-to-end learning replaces all stages with a single neural network
- Image example
- Text analogue:
  - Which part of the law is relevant
  - In the relevant part what content is important

- It is also probably a good idea to think if we should handle each problem separately or think at domain (space) level solutions. Idea from Zamir et al (2018)
- Lots of people (and a lot of money) is interested in analysing legal texts.
- We tend to try to solve each problem separately (run into insufficient labelled data issues)
- Also, we tend to just go for an end-to-end solution.
- How about pre-training a model on the whole corpus of the law.
- And for individual tasks with smaller labelled data we can use transfer learning.

Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3712-3722).

Other uses of ML in Social Sciences

ML for model selection and inferences

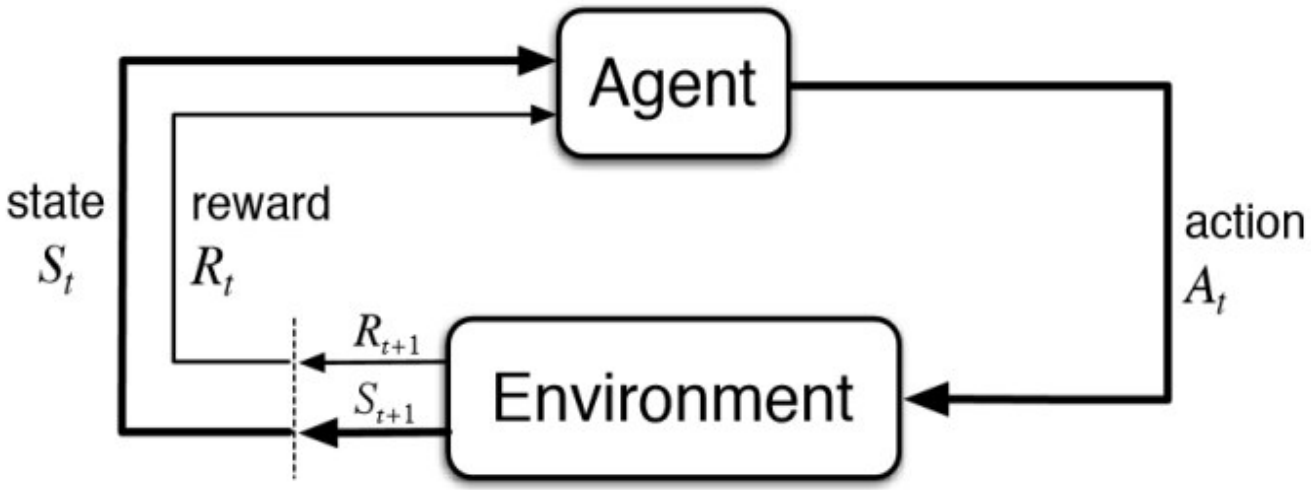
When there is a dimensionality problem

- Increasingly used in economics
- Regularized Linear Regression: Lasso, Ridge, and Elastic Nets
- Regression trees, random forests

ML for causal inferences

- Orthogonalisation and cross-fitting
- ML for heterogeneous (wrt observables) treatment effects
- Big questions whether these ML estimates have properties such as asymptotic normality
- Example: merger retrospectives

Reinforcement learning



- Algorithmic collusion
- Rational v boundedly rational consumers
- Decision making in Law (mutually informative)

Conclusion

- ML is increasingly used in Social Sciences.
- In economics there is of course the question of how the profession relates to data-driven rather than theory-drive solutions.
- Obstacles to innovation in social sciences.
- On my level:
  - Discovering the data science literature
  - Collaboration outside of academia