

Machine Learning for medical diagnostics

Supervisor: Joël Tabak, College of Medicine and Health j.tabak@exeter.ac.uk

Background

Machine Learning is a branch of artificial intelligence that is behind new technologies such as self-driving cars or unlocking an iPhone using face recognition. As the name suggests, Machine Learning models learn from data to perform these tasks, as opposed to being given formal instructions as in traditional software development.







Machine Learning has enormous potential in medical diagnostics. For example, in cancer diagnosis, images of biopsy samples are examined individually by a pathologist who then determines if a tumour should be classified as benign or malignant. It may take years for human pathologists to develop the expertise to do this with good accuracy. Such classification depends on many features of the cells (such as size and shape of the cell nucleus) and how these features combine in ways that are difficult to represent in traditional two-dimensional graphs. While it is difficult for us to see patterns in multi-dimensional data, Machine Learning algorithms are designed to handle such data. It is hoped that with increasing access to large high-quality medical data sets, it will be possible to develop Machine Learning algorithms that automate and enhance medical diagnostics.

There is currently a shortage of data scientists with machine learning experience. This project will give you good grounding in machine learning, which employers are seeking.

Research Project

The group will choose a biomedical research question to address, from a depository of databases. Examples of such problems include: classification of tumour cells from images, predicting if a patient will experience an epileptic seizure using brain activity recordings, determining which gene mutations among hundreds are responsible for a type of cancer, predicting whether a patient will develop diabetes using genetic and environmental factors, etc.

These datasets are publicly available on sites such as [kaggle.com](https://www.kaggle.com/), Google's online community for data scientists and machine learners, which provide classification challenges. The datasets include both the features to be used by the machine learning model, and the classification that the model must learn.

219		Breast Cancer Proteomes Dividing breast cancer patients into separate sub-classes kajot updated 3 years ago (Version 3)	healthcare	CSV 5.4 MB Other	</> 78 9 55k
109		MRI and Alzheimers Magnetic Resonance Imaging Comparisons of Demented and Nondemented Adults Jacob Boysen updated 2 years ago (Version 1)	healthcare neurology health scie... + 2 more...	CSV 12.5 KB CC0	</> 16 1 30k
38		Cuff-Less Blood Pressure Estimation Pre-processed and cleaned vital signals for cuff-less BP estimation Mohammad Kachuee updated 2 years ago (Version 5)	healthcare health	Other 4.6 GB Other	</> 12 9 12k
13		Parkinson Disease Spiral Drawings Hand drawing data using Digitized Graphics Tablet Data Set Team AI updated 2 years ago (Version 1)		Other 3.7 MB CC0	</> 0 0 4k
18		HCC dataset Hepatocellular Carcinoma Dataset mrsantos updated 4 months ago (Version 5)	healthcare hospitals survival an... medicine	Other 2 MB CC4	</> 4 1 6k
9		Raman spectroscopy of Diabetes Raman Spectroscopy to Screen Diabetes Mellitus with Machine Learning Tools Edgar Guevara updated a month ago (Version 7)	healthcare nutrition health + 2 more...	CSV 1.3 MB Other	</> 1 0 3k

After choosing a research question, the group will train machine learning models to correctly determine a diagnosis. Once an accurate model is developed, the group will analyse how the trained models work. This last step is often neglected in machine learning projects, but it is extremely important. If we do not understand how the model determine a diagnosis, we may not be able to convince doctors and patients to accept it. Additionally, understanding how the network classifies data will give us useful information about which features of data are most important for determining the diagnosis. For instance, what combination of cell features is a good predictor of malignancy? This information could provide useful insight into the disease for doctors.

In previous years, students have chosen to predict the risk of coronary heart disease from routine general practice medical data. They developed an algorithm that was more accurate than the criteria currently used by family doctors. Last year, students developed algorithms to assess the severity of Covid-19 cases in hospital. We will choose together a problem that is both potentially useful to address and of current relevance.

Research Methods

Each dataset contains records of samples, typically hundreds or thousands. Each record consists of a number of features (eg nucleus size, roundness, symmetry, etc.) and a true diagnostic (eg benign, malignant) made by a human expert. The data will be divided into a training set and a test set. The training set will be used to train the machine learning model

to perform the diagnostic. To do so, the model predictions are compared to the true diagnostic and the difference is the model error. Model parameters are varied until the error is minimised. Once this is done, the model has “learned” the training set. The capacity of the trained network to generalise to other data will then be tested by evaluating the predictions of the network on data that it has not seen so far: the test set.

Model analysis will determine how the model uses the data to make its prediction. For simple models such as logistic regression, we will use the model parameters to reconstruct how the model performs its diagnostic. For more complex models this might be cumbersome, so we will use a different strategy. We will instead evaluate the importance of each feature to the diagnostic.

All dataset manipulation, model training and evaluation will be done using Python modules such as numpy, pandas, and scikit-learn.

Program of research:

Week 1
<ul style="list-style-type: none"> Undertake training on the use of python and machine learning
Weeks 2-3
<ul style="list-style-type: none"> Select a biomedical dataset and research question, and undertake a brief literature review of the disease area, treatments, patient outcomes and current diagnostic methods. Get familiar with the selected dataset – Data exploration is an important step before proceeding to machine learning.
Weeks 4-6:
<ul style="list-style-type: none"> Train simple machine learning models (logistic regression, decision trees) and assess their performance.
Week 7-8:
<ul style="list-style-type: none"> Train more complex machine learning models (neural networks, random forests) and assess their performance.
Week 9-10:
<ul style="list-style-type: none"> Analyse how the best models make their prediction.

Skills

These skills will be used and developed during the project:

- Python programming
- Data manipulation and visualisation
- Machine learning and optimisation

Resources and reading

- Introduction to Machine Learning with Python, by Andreas Muller and Sarah Guido.
- Andrew Ng series of videos on machine learning

https://www.youtube.com/playlist?list=PLLssT5z_DsK-h9vYZkQkYNWcltqhIRJLN