MIT 6.862 Applied Machine Learning Project Proposal

Matthew West, Fall 2019

Question

This is a supervised learning project where the questions of interest relate to medical diagnostics, investigating how supervised learning can be used in classification and particularly in determining the presence of breast cancer from a dataset derived from mammogram images. The first part of the project will involve looking at both traditional machine learning (Naive Bayes, SVM, etc) and more contemporary deep neural networks for classification, and to determine which approach performs best using a naive metric such as accuracy, thus providing some basic insight into which approach might be favourable for the nature and size of dataset available. Given the relatively small size of the dataset, it is anticipated that deep networks won't confer a significant benefit in classification, though it is possible that more data will become available, ideally including the original images from which the chosen dataset is derived.

Following this, it will be of interest to extend the use of accuracy, bringing in alternative metrics related to precision and recall, and performing a comparison between the different algorithms as before, but with a set of new potential metrics. This will also involve an exploration of which types of metric are preferable for medical diagnostics, where we often care much more about minimizing false, giving comparatively little weight to false positives.

Data

The dataset for the first part of this project at least will be the Wisconsin Breast Cancer Dataset [2]. This was collected by a group at the University of Wisconsin in the early 1990's, and is a structured dataset with 10 features computed from a set of digitized images of breast tissue. There are 569 instances, with 357 labelled as malignant (M), and 212 as benign (B). This dataset thus presents a binary classification problem, making it amenable to many metrics of success that incorporate precision and recall.

Beyond this, there are a number of other candidate datasets that could be incorporated into the project, but each presents a challenge in comparing to the Wisconsin dataset. Most of them are raw images only and typically labelled in terms of regions of interest, thus presenting both an image segmentation and classification problem simultaneously. While it would be fruitful to compare Convolutional Neural Networks to traditional machine learning classifiers on data of this variety, there is another layer of complexity inherent in extracting appropriate features to optimize classifier performance, such that they would provide a fair comparison to CNN's.

Among these potential datasets are DDSM, the Digital Database for Screening Mammography, and its curated sub-dataset with expert mammographer ground truth validation, Curated Breast Imaging Subset (CBIS-DDSM). [4] These present an image segmentation problem, so there would be a considerable amount of additional work in processing the data before it can be used purely for investigation of classification. Another promising candidate dataset is the CAMELYON dataset, which was made public for a grand challenge in pathology posed by the Department of Pathology at Radboud University Medical Center. [5]

Motivation and Previous Work

Machine and deep learning methods have already received significant attention within medical imaging diagnostics, a well publicised example of which was DeepMind's use of deep learning in diagnosing retinal disease. [1] More specifically, similar methods have been used in breast and prostate cancer imaging diagnostics. [6, 3]

The motivation for this kind of work is clear, as medical diagnostics is becoming increasingly automated and incorporating the use of statistical and machine learning methods. There are a number of practical and ethical issues in using machine learning in a clinical setting, even when being utilized in tandem with a trained clinician, so it is important to give proper treatment to classification problems specific to the medical domain. Another desirable feature of models used in this way is interpretability, which some sophisticated models struggle with, so it may be the case that for a diagnostic tool it becomes worth trading performance in some metric for an increased degree of interpretability.

Computational Resources

I will be running and evaluating the models for this project primarily on my local machine, and perhaps using external cloud GPU services (AWS, etc) if the models become sufficiently complex. My computer has an Intel i7 with 32GB of RAM and an Nvidia GTX 1650 GPU, with 4GB of GDDR5 memory, and I anticipate that this will be sufficient for training most of the models within a reasonable time frame, given the size of the dataset. I will use the programming language Python 3 and a selection of machine learning libraries and frameworks such as scikit-learn and TensorFlow.

Project Plan

- **Exploratory Data Analysis:** Get to know the dataset and demonstrate that classification is even a reasonable thing to attempt given the problem statement. Set up data pipeline to make classification possible. Identify useful features and reduce to the most useful ones if necessary. (By 10/14)
- **Comparison of traditional ML on selected feature set:** Compare different algorithms using cross-validation to determine accuracy on validation data. (By 10/24)
- **Beyond Accuracy: Precision and Recall** Explore alternative metrics of success and determine which metric captures the desired behaviour of a tool used in medical diagnostics. Re-evaluate previous algorithms using selected metrics. (By 11/10)
- **Beyond Wisconsin (time allowing):** Incorporating alternative datasets with raw images allowing for comparison between CNN's and traditional ML on structured data. (By 11/30)

References

[1] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan ODonoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342, 2018.

[2] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[3] S Larry Goldenberg, Guy Nir, and Septimiu E Salcudean. A new era: artificial intelligence and machine learning in prostate cancer. *Nature Reviews Urology*, page 1, 2019.

[4] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4:170177, 2017.

[5] Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, et al. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *GigaScience*, 7(6):giy065, 2018.

[6] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.