MIT 6.862 Applied Machine Learning
Project Progress Report

Matthew West, Fall 2019

1 Introduction

This is the progress report for my MIT course 6.862 project, which aims to explore classification in the context of medical diagnosis, and specifically in datasets derived from images of breast cancer. The focus thus far has been on the establishment of a data pipeline and exploratory data analysis, before comparing multiple candidate machine learning models on the basis of classification accuracy.

This report focuses on the Wisconsin Breast Cancer dataset, [2] an analysis of which first appeared in Street et al., which used a decision tree-based algorithm to classify the data with approximately 97% accuracy. [7, 6] As outlined in the proposal, it was decided that comparison between traditional machine learning on structured data derived from images and the raw images themselves was a desirable direction to take the project in. To this end, I have contacted an original author of this paper in an attempt to acquire the corresponding raw images, though this has not proved fruitful as of yet. There remains the option to extend the analysis to further datasets outlined in the proposal, though such a comparison still won't be straightforward and will introduce substantial roadblocks in the form of data extraction and image segmentation.

The report is structured as follows: Following this introduction, an account of the progress so far will be given regarding the data, methods, results, and discussion, followed by a brief outline of what will be done with the time remaining, and outlining a specific plot that will appear in the final report.

2 Progress

2.1 Data

The data used in the initial portion of the project thus far is the Wisconsin Breast Cancer dataset, collected by a team at the University of Wisconsin-Madison in the early 1990's. The dataset uses features derived from a set of fine-needle aspirate (FNA) images of breast tissue in patients with either malignant or benign tumours. It is a structured dataset with 569 instances and 30 features, an ID number, and a label, with 212 malignant examples and 357 benign examples. There are 10 base features (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension), and for each of these the mean, standard error, and largest/worst example of these features are included over each image, resulting in 30 features in total.

The dataset was checked for missing values, before dropping a non-explanatory ID column, an unexplained column of all NaN values, and separating the labels from the feature vectors.

2.2 Method

The machine learning problem at hand is one of classification. If we denote our hypothesis by $\mathrm{h}(x;\lambda)$ where the feature vector for each tumour instance is denoted as $x^{(i)}$ and λ denotes a specific vector of parameters that determine a hypothesis, our problem is to identify the optimal λ subject to minimising

$$\frac{1}{n}\sum_{i=1}^{n} L(\mathrm{h}(x;\lambda), y^{(i)}),$$

where $L(g, a)$ is the loss function chosen for a given algorithm, providing a measure of how far away the guess from the hypothesis g is from the actual instance label, a. Specifically, this will be evaluated using 5-fold cross validation, as minimising training error will not be a relevant or interesting metric. Selected hypotheses from many such algorithms will be compared to see which of these minimises validation error and is therefore the preferred model in terms of classification accuracy. Multiple algorithms have been investigated so far, with promising results from xgboost, naive Bayes and logistic regression, as seen in table 1. These were mainly implemented using scikit-learn in Python 3 on my local machine.

A principle component decomposition was done on the dataset, allowing 2 or 3 principle components to represent a substantial portion of the variation across all 30 features. This allows for an intuitive understanding of how separable the dataset is, as two or three principle components can be plotted and understood visually, something that isn't readily doable for higher dimensional data. It is also possible to directly train and test models on the principle component representation of the data, though one loses the interpretability for which features are most important in predicting the outcome variable.

2.3 Results

The first result was, having used PCA to represent the data in both 2 and 3 dimensions, it was confirmed that the dataset would be somewhat separable, if not perfectly linearly separable. This was achieved by confirming that the PCA plots showed natural clusters of malignant and benign instances in the reduced latent space.

Furthermore, it was determined preliminarily that tree-based methods such as xgboost are performing better than other algorithms, with xgboost consisently obtaining 96% classification accuracy without too much parameter tuning or feature selection. An additional bonus of using tree-based methods is that they allow some quantitative insight into which features are most important in classification, which appear to be those derived from texture and concave points.

Algorithm	5-fold accuracy set	5-fold accuracy (%)
xgboost	[0.930, 0.957, 0.992, 0.965, 0.973]	96 ± 2
Naive Bayes	[0.922, 0.922, 0.956 0.947, 0.956]	94 ± 2
Logistic regression	[0.930, 0.939, 0.973, 0.947, 0.965]	95 ± 2

Table 1: Table of cross-validated results from selected classification algorithms

2.4 Discussion

Getting rather good results using default hyperameters can be straightforward for some datasets. Marginal performance increases then come from the selection of algorithms together with the set of hyperparameters tuned for that algorithm. Without exploring this vast solution space comprehensively, it has still been possible to obtain classification accuracies close to that initial benchmark described in the orginal paper. One insight that has helped to overcome the 'noise' associated with the accuracy metric across this solution space was the decision to use 5-fold cross validation when investigating accuracy.

A key lesson learnt thus far was how to think about exploratory data analysis in a systematic way, such as by using PCA to represent the data. This paralleled nicely with the content covered in class on autoencoders, but simplifying it to a linear combination of original features. This has allowed me to abstract conceptually from thinking about raw features to latent spaces, which is an invaluable tool in applied machine learning.

3 Looking Ahead

An outline for the next month leading up to the conclusion of the project is given in this section.

- An investigation into precision and recall. In the context of medical diagnostics, the use of machine learning algorithms is controversial due to the direct impact of such diagnostics on human survival. Therefore, perhaps unlike many other domains in which machine learning is applied, special consideration must be given to precision and recall, as false positives are in general much less serious than false negatives. This presents a different set of potential metrics, and some of these will be investigated to see if a different algorithm is preferred when we take the domain-specific selection of metric into account. Deadline: 11/14
- It will hopefully be possible to acquire a set or perhaps representative subset of the original Wisconsin Breast Cancer images from the researchers that first collected this data. If this is the case, then the focus of this milestone will be to investigate the use of convolutional neural networks (CNN's) on these raw images, and compare performance on any relevant metrics. It is of particular interest to me to see if there are advantages to using CNN's in the small dataset regime, despite not expecting a substantial increase in resultant accuracy (e.g. interpretability, automated feature extraction). I identified this as "time-allowing" in the project proposal because of the potential time commitment of pre-processing these larger datasets for use in CNN's and also deriving structured datasets from these similar in form to the Wisconsin dataset, for the purposes of making this comparison. Deadline: 12/01

4 Projected Results Section

The key result of this initial stage of the project will be a comparison of accuracies on the Wisconsin dataset using one model for each of the algorithms selected for comparison, by means of a violinplot or boxplot. The horizontal axis will be discrete in the different algorithms, and the vertical axis will be classification accuracy.

Each of these models will have been tuned to select the optimal set of any relevant hyperparameters, using 5 or 10-fold cross-validation. This plot allows the reader to assess visually which algorithm performs better on this dataset, as well as how much variance there is in each estimate of accuracy across the set of folds used for validation. I envision including a set of similar figures that concern performance not just in terms of accuracy, but also on other metrics that take into account sensitivity and specificity.

References

[1] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, 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] Olvi L Mangasarian. Mathematical programming in neural networks. *ORSA Journal on Computing*, 5(4):349–360, 1993.

[7] W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical image processing and biomedical visualization*, volume 1905, pages 861–870. International Society for Optics and Photonics, 1993.

[8] 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.