MIT 6.862 Applied Machine Learning Project Pre-Proposal

Matthew West, Fall 2019

Machine Learning for medical diagnostics: How does traditional ML compare to deep learning on a small breast cancer dataset?

This is a supervised learning project where the question of interest will be one of diagnostics; how can supervised learning be used to detect the presence of breast cancer in mammograms, and what type of algorithm might be optimal for this dataset? In particular, the project will look at both traditional machine learning (Naive Bayes, SVM) and more contemporary deep learning for classification, and attempt to discern which approach might be favourable for the nature and size of dataset available. The dataset in question is the Wisconsin Breast Cancer dataset, which poses a binary classification problem (malignant/benign) and has 10 features computed from a set of 559 images, with 212 malignant and 347 benign examples. [1] If time allows, and if the full image dataset becomes available, it may be of interest to extend to using CNN's on the raw images rather than extracted features. It will also be of interest to explore which algorithm is superior when paying more close attention to precision and recall, as false positives may be considerably more or less common than false negatives depending on the algorithm.

References

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.