School of Electrical Engineering and Computer Science

CSI5155 – Machine Learning

Project Fall 2020

Total: 40 marks

This project involves the analysis of health care data, as obtainable from https://archive.ics.uci.edu/ml/machine-learning-databases/00296/

Specifically, the dataset contains the information about patients with diabetes in 130 hospitals in the USA for the years 1999 to 2008. The study consists of approximately 100,000 patients with 55 features, some containing missing values.

The original study, as published in https://www.hindawi.com/journals/bmri/2014/781670/, looks into the impact of haemoglobin (HbA1c) measurement, which refers to the average level of blood sugar over the past 2 to 3 months, on hospital re-admission rates. The authors used multivariable logistic regression in their study. In this course project, our aim is to assess the value of machine learning algorithms, and notably supervised and semi-supervised learning techniques, when applied to this data.

A: Feature engineering, supervised learning & evaluation of results [20]

As a first step, you should explore the data to obtain a general understanding of the problem domain. Detailed descriptions of the features may be found in Table 1 of the above-mentioned paper. You next step would be to assess how to handle missing values, and to decide on other feature transformations, if need be. Of course, this step also involves calculating the levels of class imbalance in our dataset, as we move towards supervised learning.

There are numerous ways to explore this data, and you should include two different supervised learning tasks in your analysis.
Task 1: The first task you should focus on is to predict the outcomes, in terms of patient re-admissions. This is a multi-class learning problem, with three class labels {no, readmitted within 30 days, readmitted after 30 days}.
Task 2: The choice of the second task is your own. For instance, an angle would be to explore the data by using gender as class label. Alternatively, building models by using age ranges, or the admissions source, as class labels, could also provide us with additional insights.

Be sure to include at least one algorithm from the different families we covered in this course: trees, linear models such as neural networks and support vector machines, distance-based algorithms, Bayesian approaches and ensembles.

In medical domains, we are interested in sensitivity (recall) and specificity (negative recall). The evaluation of the results should also explore the overall accuracy, as well as metrics such as the f-measure (trade-off between precision and recall) and the runtime. In addition, you should draw the ROC curves and determine whether there are any statistical differences between the performances of the algorithms.

B: Semi-supervised learning & evaluation of results [10]

Semi-supervised learning, where we address the scenario where most class labels are unknown, is a very common technique used in real-world applications. In this approach, the aim is to combine a small amount of labeled data with a large amount of unlabeled data during training. Semi-supervised learning is based on the observation that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy. (Note that we will cover this topic early in November.)

During this part of the project, you are required to implement three different semi-supervised approaches for the hospital re-admissions task (Task 1). You should test various levels of unlabelled data, notably 0% (fully supervised - baseline), 10%, 20% 50%, 90% and 95%. Again, we are interested in the sensitivity and specificity of results. The evaluation of the results should also include the accuracy, f-measure, and the runtime. In addition, you should draw the ROC curves and determine whether there are any statistical differences between the performances of the algorithms.

C: Source code, final report, and project demonstration [10]

Submit your source code (or a link to a GitHub repository), as well as a final report in PDF that explains (i) the steps you followed, (ii) the results you obtained, and (iii) the lessons learned. You are also required to demonstrate your project during an individual Zoom meeting, to be scheduled during the week of 7 December 2020.

Link to resources on semi-supervised learning

1. Original survey paper by Zhu:
https://minds.wisconsin.edu/bitstream/handle/1793/60444/TR1530.pdf?sequence=1

2. Recent survey by Van Engelen and Hoos:
https://link.springer.com/article/10.1007/s10994-019-05855-6

3. Two semi-supervised techniques in Scikit Learn: https://scikit-learn.org/stable/modules/label_propagation.html#