Literature Review

Batch Active Learning for Drug Discovery

rjb255

January 29, 2022

Abstract

1 Introduction

2 Active Learning

There are several schools of thought with regards to active learning. These can be separated into two distinct categories: current data and future predictions. The former of these is computationally cheaper, as will be apparent on discription.

2.1 Current Data

The simplest is applicable to cases in which a certainty is provided with each prediction. By simply selecting the largest uncertainty with the remaining data points allows for this form of active learning. For an unknown, noisy linear funciton (1), a simple funciton for selecting a subsequent data point has been given.

$$f(x) = ax + b + N(\mu, \sigma^2) \tag{1}$$

A second form stems from information theory. Here, the aim is to produce an evenly dispursed x allowing a well informed knowledge base. U [Eis20].

Conversly, a density weighted model has been suggested, as it escapes the introduction of error from outligher (i.e. data points far away from alternative datapoints).

As more complex methods are explored, we stumble across the method of competing hypothesis. This builds upon the [], and attempts to find []. The majority of work here relates to classification, although the same principles apply to regression.

2.2 Estimated Future

These methods attempt to minimise a future attribute to the model. The first of these attempts to

3 Batch Active Learning

Several naive methods are available here. Firstly, getting the top N data points from a model described in Section 2. However, this method does not take into account the equivalence of the data points. This is extremely clear using the highest uncertainty method. It stands to reason that the area which has the highest uncertainty will see this for the data points nearest neighbours. Thus, this singular data point suffers the potential of being surrounded by $N-1$ other data points. The benefit this provides in fitting the model is thus extremely limited, and only slightly greater than if one data point had been chosen. A simple fix would be to simulate the model after 1 iteration, and select the next point from here. By doing this $N-1$ times, a better solution may be found, although this may prove to be computationally very expensive.

4 Drug Data

References

[Eis20] Michael Eisenstein. 'Active machine learning helps drug hunters tackle biology'. In: *Nature Biotechnology* 38.5 (May 2020), pp. 512–514. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0521-4. URL: https://www.nature.com/articles/s41587-020-0521-4.