

Literature Review
Batch Active Learning for Drug Discovery

rjb255

January 29, 2022

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

0.1 Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

0.2 Active Learning

There are several schools of thought regarding 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 description.

0.2.1 Current Data

Uncertainty Sampling

The simplest is applicable to cases in which a certainty is provided with each prediction. Settles [Set09] suggests selecting the data point with the largest uncertainty according to the current model. For an unknown, noisy linear function (1), a simple linear model is given in (2) where $\alpha \sim N(\hat{\alpha}, s_{\alpha}^2)$ and $\beta \sim N(\hat{\beta}, s_{\beta}^2)$, and α and β are independent of (x, y) . By noting the non-independence of α and β , a new distribution can be found for $g(x)$. Note that since αx and β are both normal distributions, $g(x)$ is normal and a function of x .

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

$$g(x) = \alpha x + \beta \tag{2}$$

Where X is the set of x that can be measured, the next data point, x_{next} , is given in (3). As addressed by Settles [Set09], this can be extended to any

probabilistic model.

$$x_{\text{next}} = \operatorname{argmax}_X [s_g(X)] \quad (3)$$

Settles [Set09] also notes the use of information theory for probabilistic models (4), where y_i refers to all possible categorisations for x . This derives from the principle that the greatest entropy requires the most information to encode, and thus the least certain. However, Settles [Set09] fails to address non-probabilistic models in this instance, instead converting such models into probabilistic ones.

$$x_{\text{next}} = \operatorname{argmax}_x \left[- \sum_i P(y_i|x) \ln P(y_i|x) \right] \quad (4)$$

Broad Knowledge Base

A second form stems from information theory. Here, the aim is to produce an evenly dispersed x allowing a well-informed knowledge base. There are two paths to proceed: density and nearest neighbours.

The former of these requires a definition of density in a sparsely populated space. As an analogy, the density of a gas appears well-defined, it becomes non-smooth once the volume defined over is comparable to the distance between particles. Thus, a new definition is required.

Alternatively, nearest neighbour requires little explanation. x_{next} is the unlabelled data point furthest from any labelled data point.

Density Hotspots

Conversely, a density weighted model has been suggested, as it escapes the introduction of error from outlier (i.e. data points far away from alternative data points). Settles and Craven [SC08] suggest (5). This can be broken down into two parts: a function for selection, ϕ_A , and a function for similarity, sim . The former arises from another method described in this section. The latter requires a function to describe the similarity between data points.

$$x_{\text{next}} = \operatorname{argmax}_x \left[\phi_A(x) \times \left(\frac{1}{U} \sum \text{sim}(x, x_i) \right)^\beta \right] \quad (5)$$

Regions of Disagreement

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. By minimising the region of disagreement between various models, a finer fit may be achieved.

0.2.2 Estimated Future

These methods attempt to minimise a future attribute of the model. This works by predicting changes given with the inclusion of more data.

Expected Model Change

0.3 Batch Active Learning

Several naive methods are available here. Firstly, getting the top N data points from a model described in Section 0.2. 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.

0.3.1 Cluster-Margin

0.4 Drug Data

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Bibliography

- [SC08] Burr Settles and Mark Craven. “An Analysis of Active Learning Strategies for Sequence Labeling Tasks”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '08. Honolulu, Hawaii: Association for Computational Linguistics, 2008, pp. 1070–1079.
- [Set09] Burr Settles. *Active Learning Literature Survey*. University of Wisconsin-Madison Department of Computer Sciences, 2009. URL: <https://minds.wisconsin.edu/handle/1793/60660>.