Literature Review

Batch Active Learning for Drug Discovery

rjb255

January 29, 2022

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

0.1 Introduction

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Scores displayed in examples have been based on the entire data set. Although this usually leads to data leakage within machine learning, this is not a concern here as the true comparison comes from testing *intelligent* vs *dumb* learning methods. In both of these cases, the model is kept identical, but the selection process in not. The baseline simply takes the first n entries from the data set, with the *intelligent* method described where required. Three data sets have been used to demonstrate on multiple data sets [][][].

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.

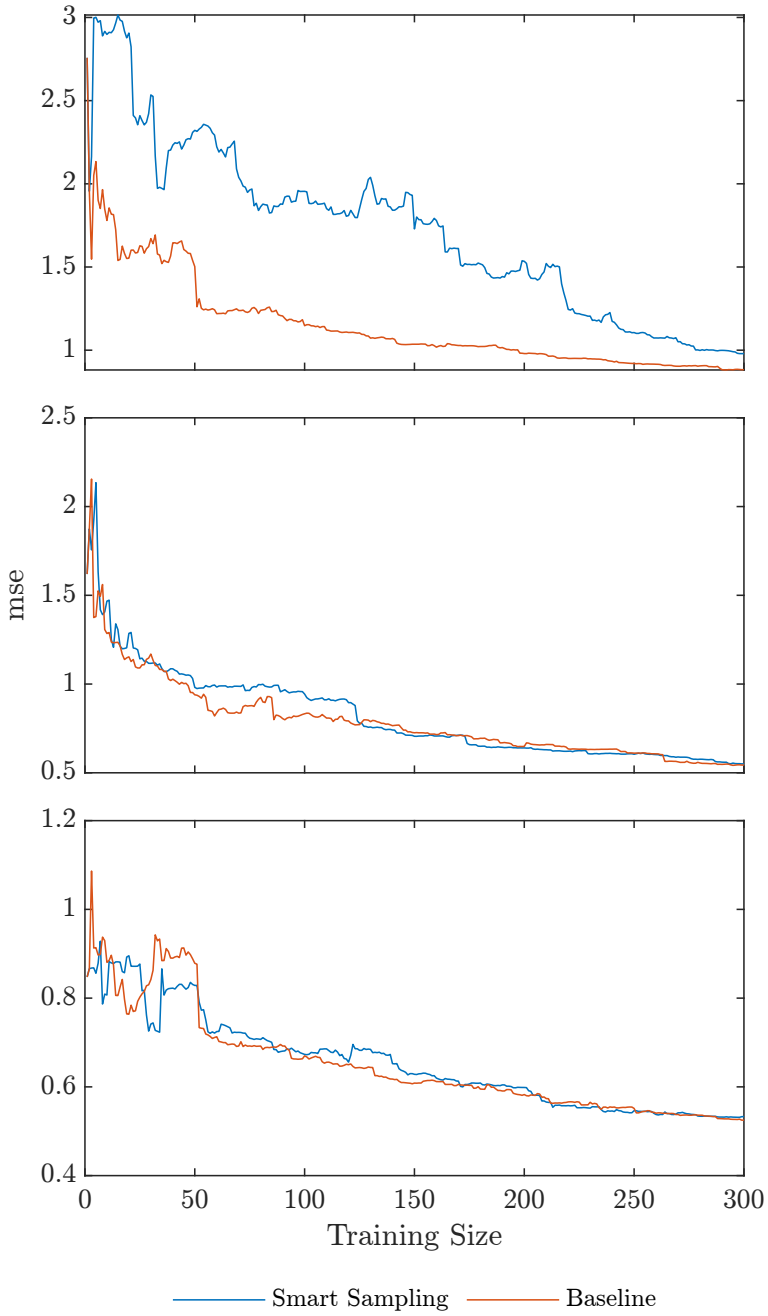
0.2.1 Current Data

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. Using the dataset '', this is demonstrated in [] with the algorithm for deciding the next sample point given in Algorithm 1.

Algorithm 1: Uncertainty Sampling Selection

Data: X_{known}, Y_{known}, X_{unknown}
Result: Next X to label
model = BayesianRidge();
model.fit(X_{known}, Y_{known});
standard_deviation = model.standard_deviation(X_{unknown});
return *max(standard_deviation)*



As addressed by Settles [Set09], this can be extended to any probabilistic model.

$$x_{\text{next}} = \underset{X}{\text{argmax}} \left[s_{g(X)} \right] \tag{1}$$

Settles [Set09] also notes the use of information theory for probabilistic models (2), 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.

$$x_{\text{next}} = \underset{x}{\text{argmax}} \left[-\sum_i P(y_i|x) \ln P(y_i|x) \right] \tag{2}$$

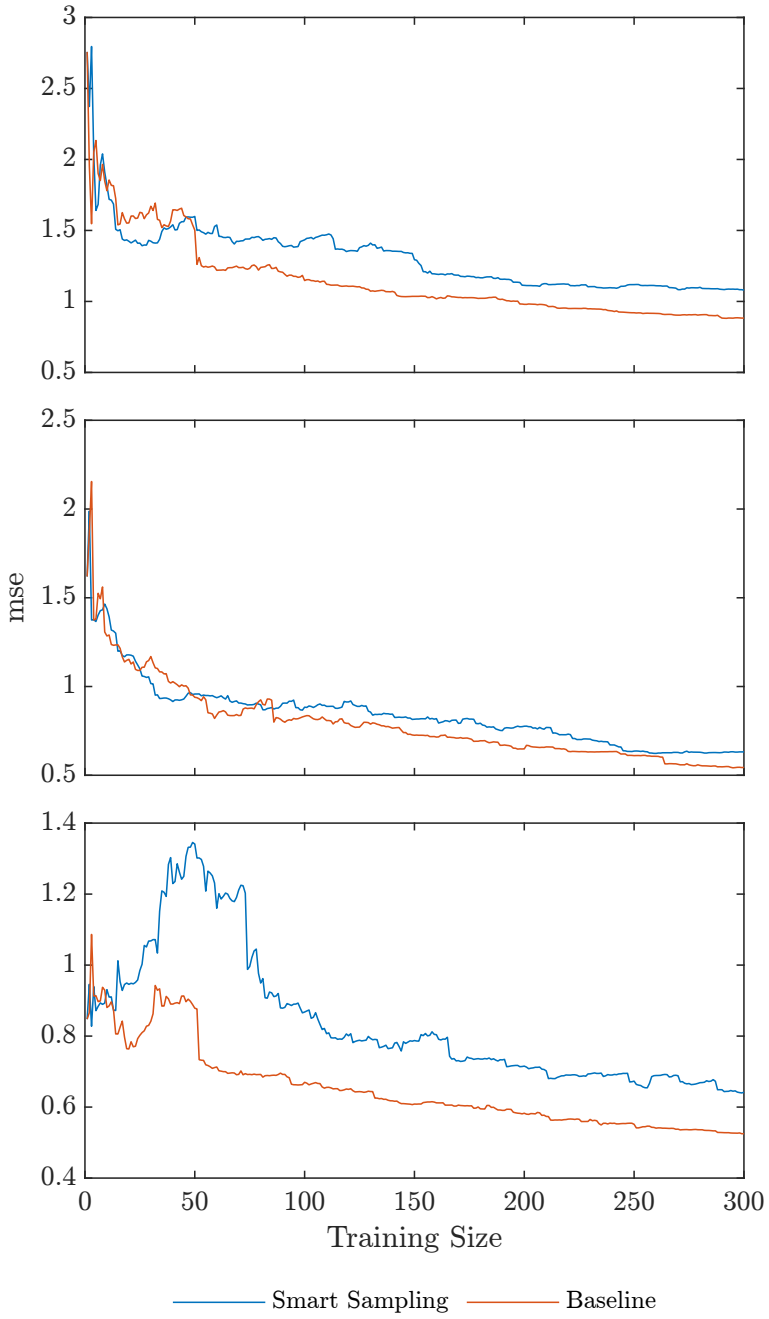
Broad Knowledge Base

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.

$$x_{\text{next}} = \underset{x}{\text{argmax}} \left(\sum \frac{1}{\text{sim}\,(x, x_i)} \right) \tag{3}$$



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 (4) which 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.

$$x_{\text{next}} = \underset{x}{\text{argmax}} \left[\phi_A(x) \times \left(\frac{1}{U} \sum \text{sim}\left(x, x_i\right) \right)^{\beta} \right] \tag{4}$$

Settles and Craven [SC08] admits that sim is open for interpretation. For simplicity, the average distance

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.

One way of achieving this, especially in a regression model where boundaries are not quite so distinct, is to declare n models $M = \{m_1, \ldots, m_n\}$. Combining these allow for a model $\hat{m}$ to be defined with prediction $\hat{y}$, being the mean prediction of M, $\frac{1}{n} \sum y_i$ and a sample standard deviation $\hat{s}$ defined as the sample standard deviation of y_i. This standard deviation can be used as a measure of the disagreement between the models. Thus, using a method as in Section 0.2.1.

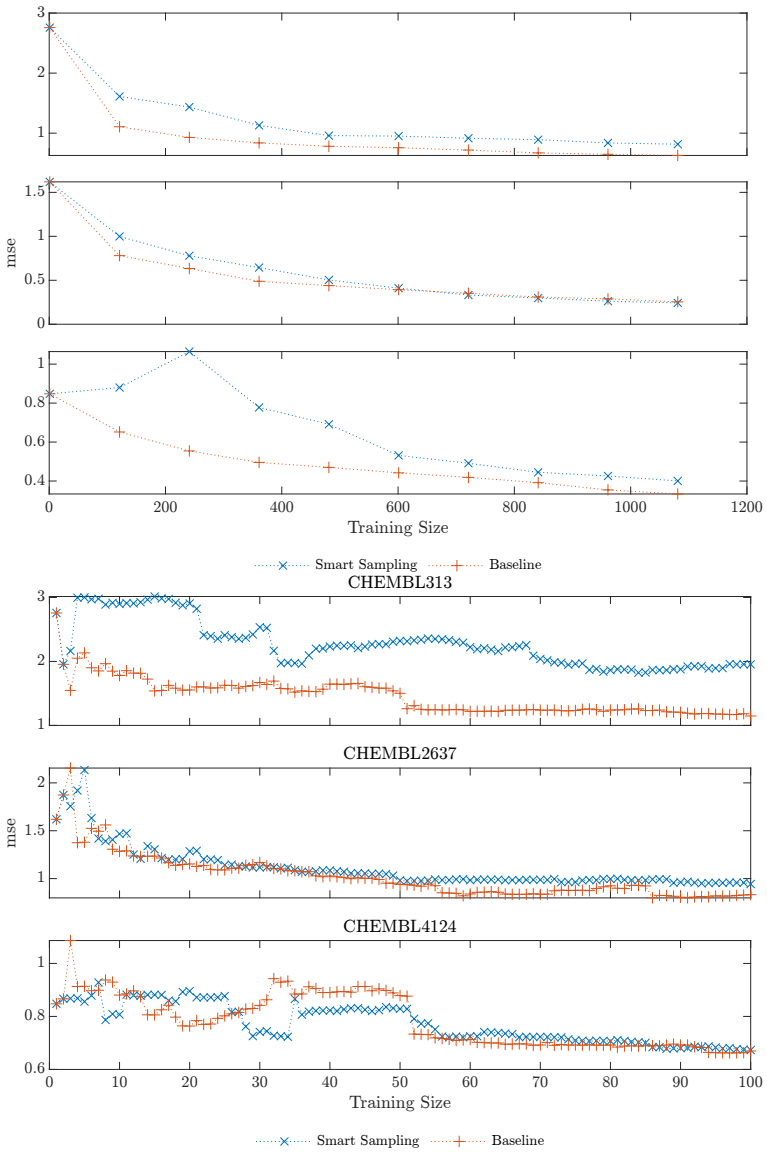
0.2.2 Estimated Future

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. Each method in Section[] has been modified to demonstrate this weakness.



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

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

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`.