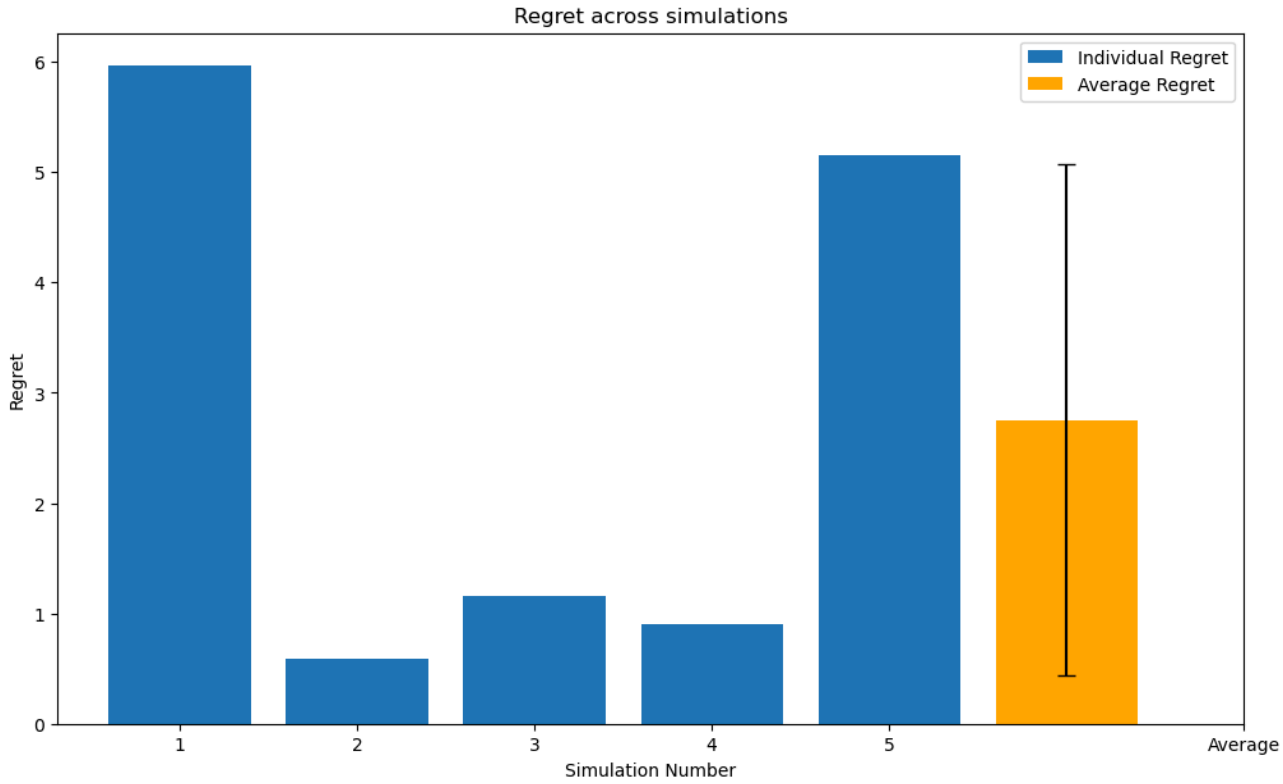
02750 HW 1
Anushka Sinha
Feb 8, 24

Exercise 1: Offline vs Online Learning:

a. A plot showing the average and standard deviation of the regret



b. Describe the results as well as discuss whether the results matched your expectations and explain your reasoning.

The plot above shows the regret obtained from a classification task across five simulations with different seeds where x-axis represents the simulation number and the y-axis represents the regret obtained for that simulation.

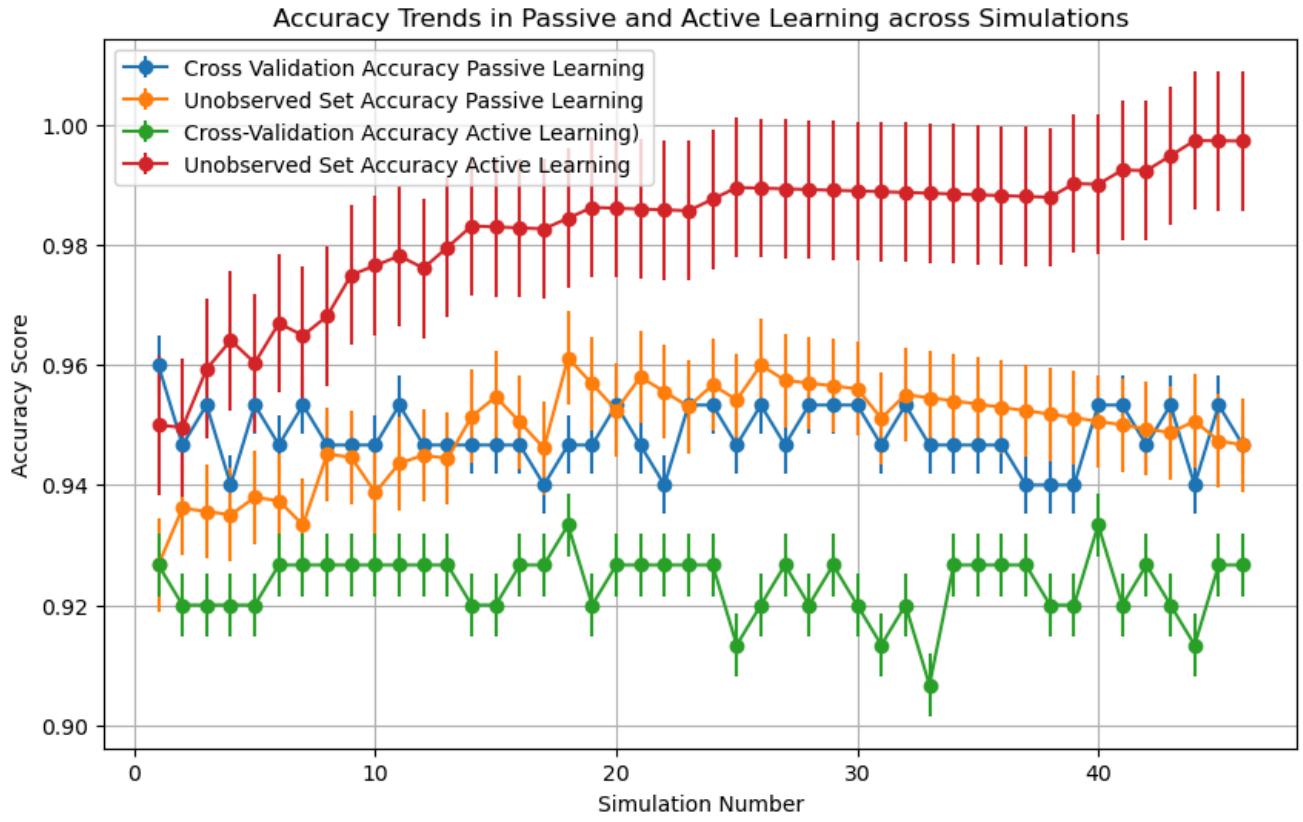
Based on the plot above, the regret obtained for each simulation was positive. There is a wide variability in the regret obtained for each simulation where the first and last simulations result in the highest regret. Since the regret allows us to infer how well the simulation would have performed (in terms of total errors) if the learning algorithm had received the entire data at the beginning, a positive value for regret implies that the online setting did worse than the offline setting for this particular dataset.

In the online setting, the model does not receive the wide range of data and therefore is not able to learn for the diverse data points that are added later on which leads to a higher cumulative loss. The results did match my expectations because the

dataset used in this problem is fairly small and incrementing the size of the data that is being received by the model to train on by 10% does not capture the underlying distribution in the dataset or the variability in the labels which is causing the model to perform poorly compared to the model in the offline setting. For example, during certain iterations in the online setting, incrementing the data by 10% was not capturing all the three labels present in the iris dataset which could result in poor predictions and therefore higher cumulative loss.

Exercise 2: Passive vs Active Learning (Classification)

A plot showing accuracy scores of Passive and Active Learning on a Classification Task



Make sure to concisely describe (i.e., no more than a few sentences), how you estimated uncertainty for that learner.

Entropy was used to estimate uncertainty of each test data point (data points that are not in the training set) using this following equation:

$$x^*_H = \text{argmax}_{x \epsilon u} \; - \; \Sigma_{y \epsilon u} \, P(y|x) \log P(y|x)$$

For this classification task, the instance/data point that has the highest entropy was chosen to be added to the sample data since that data point represents the most uncertainty about the labels.

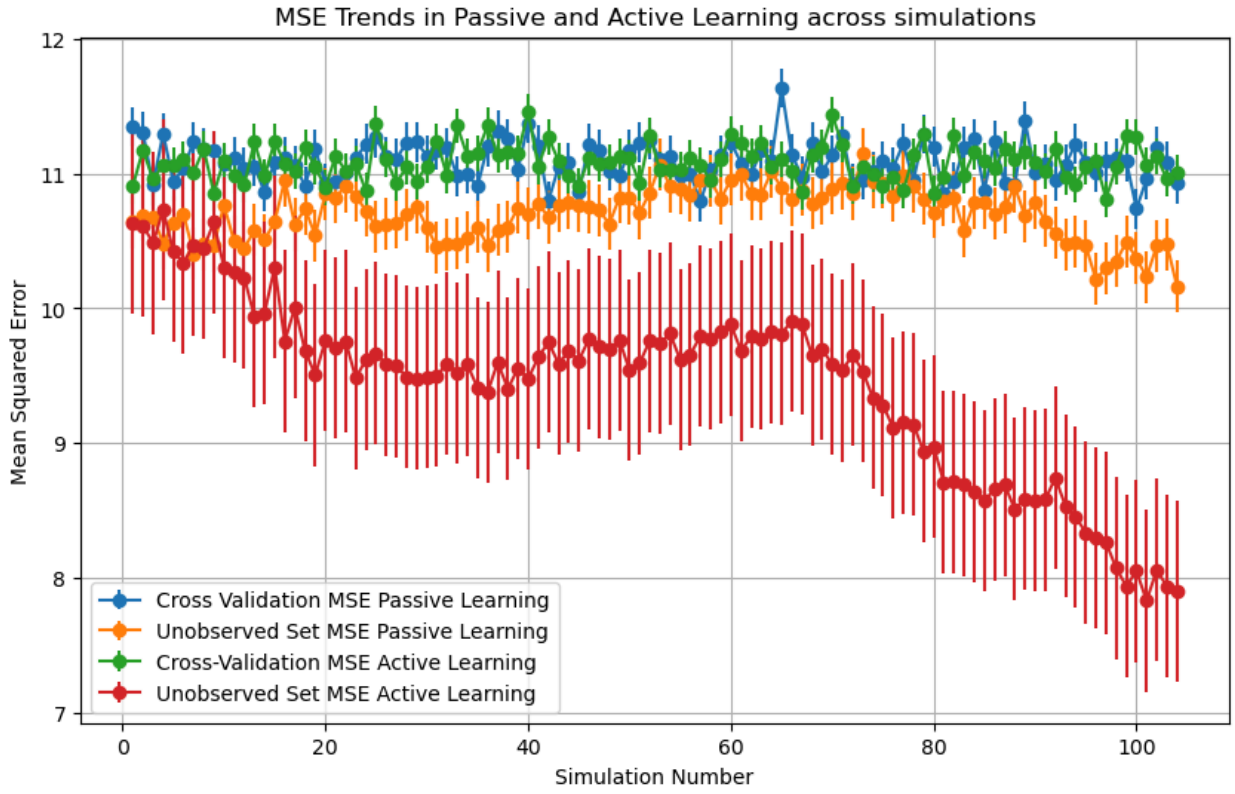
Compare the results between part 2a and 2b. Discuss whether the results matched your expectations and explain your reasoning.

The plot above shows the trends in the accuracy score in passive and active learning where the x-axis represents the simulation number and the y-axis represents the accuracy score. The results obtained from part 2a are the accuracy scores obtained from cross validation on the training set and the accuracy score obtained from passive learning on the unobserved data shown by blue and orange curves respectively. The results obtained from part 2b are the accuracy scores obtained from cross validation on the training set and the accuracy score obtained from active learning on the unobserved data shown by green and red curves respectively.

Based on the graph above, there is some discrepancy between the accuracy score trend of the cross validation from passive and active learning which goes against the expectation since the accuracy score obtained from the cross validation using active learning is lower than the accuracy score obtained from the cross validation using passive learning. It is also clear that the accuracy score obtained on the unseen data using active learning is much higher than the accuracy score obtained on the unseen data using passive learning. This observation is in line with my expectations since the instance that was added in each iteration was the instance with the most entropy (most uncertainty about the labels). Adding an instance to the training set with the most uncertainty allows the model to learn better, which is reflected in the accuracy scores obtained on the unobserved data in active learning.

Exercise 3: Passive vs Active Learning (Regression)

A plot showing Mean Squared Error of Passive and Active Learning on a Regression Task



Make sure to concisely describe (i.e., no more than a few sentences), how you estimated uncertainty for that learner.

Variance was used to estimate the uncertainty for each data point in the test dataset in this regression task. For each test point, variance was calculated using this equation:

$$Var(\hat{y}) = \sigma^2 x(X'X)^{-1}x'$$

σ^2 was calculated by using the predictions and the true labels of the seen (training) data. After that, for each test point, variance was calculated which represents the variances of the prediction which also represents the uncertainty about the prediction. The test point with the highest variance was added to the training set in each iteration so that it would improve the model the most, assuming that the selected instance is not an outlier.

Compare the results between part 3a and 3b. Discuss whether the results matched your expectations and explain your reasoning.

The plot above shows the trends in the Mean Squared Error (MSE) in passive and active learning where the x-axis represents the simulation number and the y-axis represents the MSE. The results obtained from part 3a are the MSE obtained from cross validation on the training set and

the MSE obtained from passive learning on the unobserved data shown by blue and orange curves respectively. The results obtained from part 3b are the MSE obtained from cross validation on the training set and the MSE obtained from active learning on the unobserved data shown by green and red curves respectively.

According to the graph above, the MSE obtained from the cross validation using active learning is higher than the MSE obtained from the cross validation using passive learning, which goes against my expectation. Furthermore, it is evident that the MSE obtained from active learning on unseen data is significantly lower than the MSE obtained from passive learning on unseen data. Although it shows some overlap in the earlier simulations, the active learning curve consistently goes lower past 10 simulations. This observation aligns with my expectation since the instance with the highest entropy which is also the greatest degree of label uncertainty was added to the training data in each iteration. The MSE obtained on the unobserved data in active learning shows that the model learns better when an instance with the highest level of uncertainty is added to the training set compared to the passive learning.