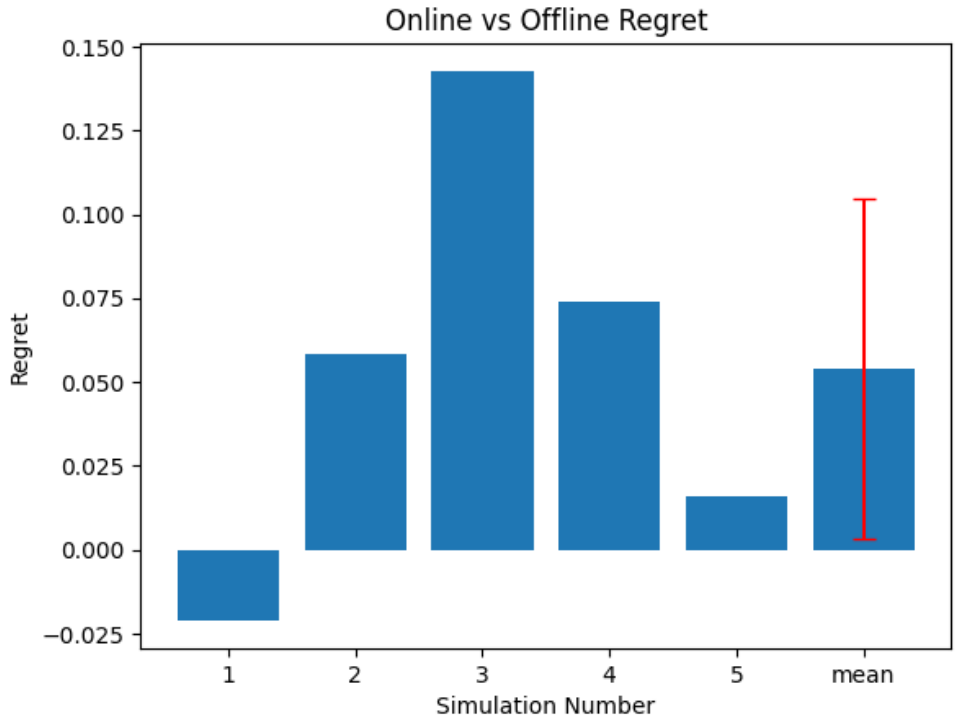
Homework 1

Thomas Zhang

February 2024

1. Question 1:



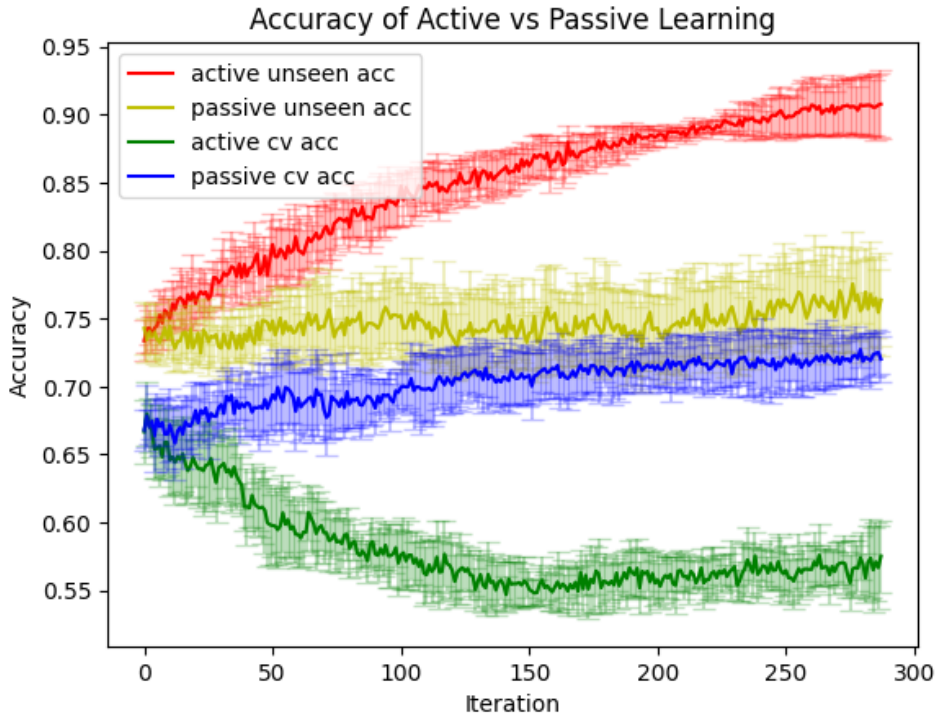
[h]

The model used was a support vector classifier. Loss was calculated with log-loss.

The results mostly matched my expectations. In general the loss of the online model exceeded that of the offline model. This is expected because the offline model sees all of the data at once. The offline model should have lower loss than the online model which can only see a portion of the data.

The one unexpected instance is that the first iteration which has negative regret. This may be attributed to the dataset being small and the random split of the data happening make the active model have higher loss than the offline model. This may also be the reason why the regret values have small regret magnitude.

2. Question 2:



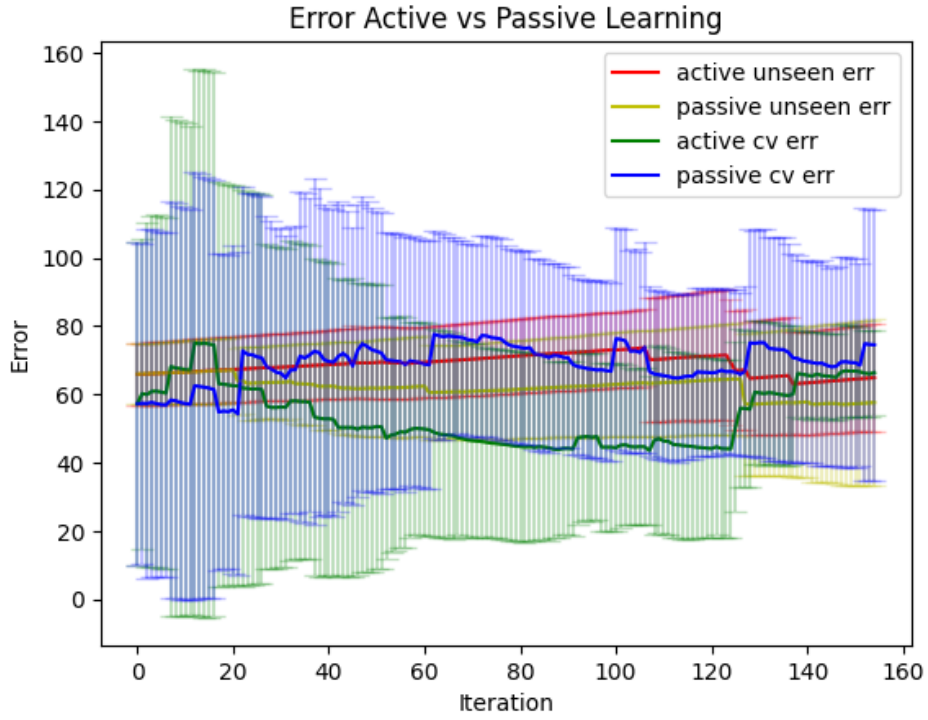
b) The model used was a Random Forest classifier. Entropy was used to calculate uncertainty and the most uncertain point was chosen each iteration. The formula for entropy is

$$\text{Entropy: } \boldsymbol{x}_H^* = \underset{\boldsymbol{x}\in\mathcal{U}}{\text{argmax}} - \sum_{y\in\mathbb{C}} P(y|\boldsymbol{x}) \log P(y|\boldsymbol{x})$$

c) The graphs somewhat matched my expectations. The active learning accuracy for uncertainty sampling was larger than the accuracy of passive learning for both the test data. This is expected because the active sampling chooses the samples that the model is most uncertain about. This means the model will learn on data points it struggles with which allows it to increase in accuracy faster than passive learning. Passive learning adds a random point which causes the model to learn much slower.

The unexpected behavor is that the, the active cross-validation accuracy drops compared to the passive cross-validation accuracy. One possible reason for this is that the newly added points run contradictary to the existing model, hence the model accuracy on the training data drops until it receives more data points.

3. Question 3:



b) The model used was linear regression with root-mean squared error. The formula for uncertainty used was variance and the most uncertain point was chosen each iteration. The equation for variance is:

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

where σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=i}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}\hat{e}_i^2$$

c) The data somewhat reflects the expected behavior. The error for the active learning cross-validation is lower than that of the passive learning cross-validation. Additionally the cross-validation error of the active learning seems to have less variance the cross-validation error of the passive learning as the number of iterations increase. This makes sense because we are no longer randomly selecting data points.

Something that doesn't make sense it that the passive unseen error is slightly lower than the active unseen error. This is unexpected because the error for the active learning method should decrease faster than the error for the passive learning method. This may be happening due to the relatively few simulations done.

Another unexpected behavior is that the error doesn't seem to have a decreasing trend, but seems to stay relatively constant. The error is also expected to decrease as the number of iterations increase. One reason this might be happening is that the dataset is very small, hence the model reaches close to the minimum loss very quickly regardless of whether active or passive learning is used.