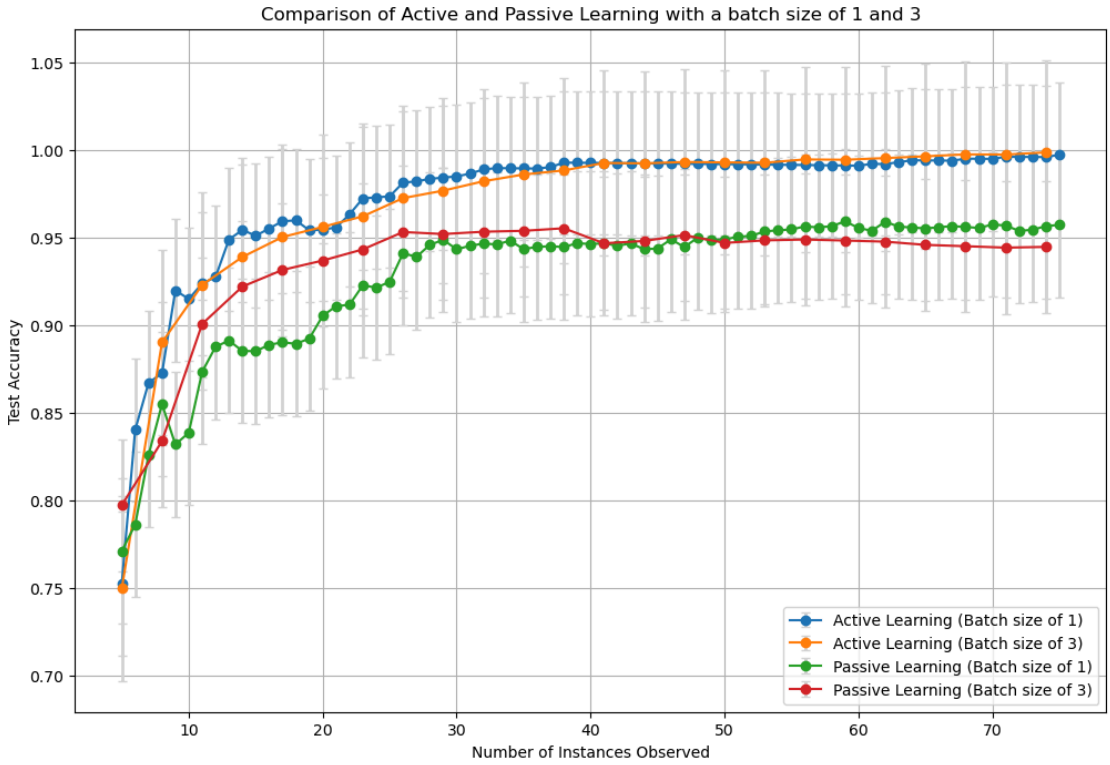
02750 HW 4
Anushka Sinha
April 21, 24

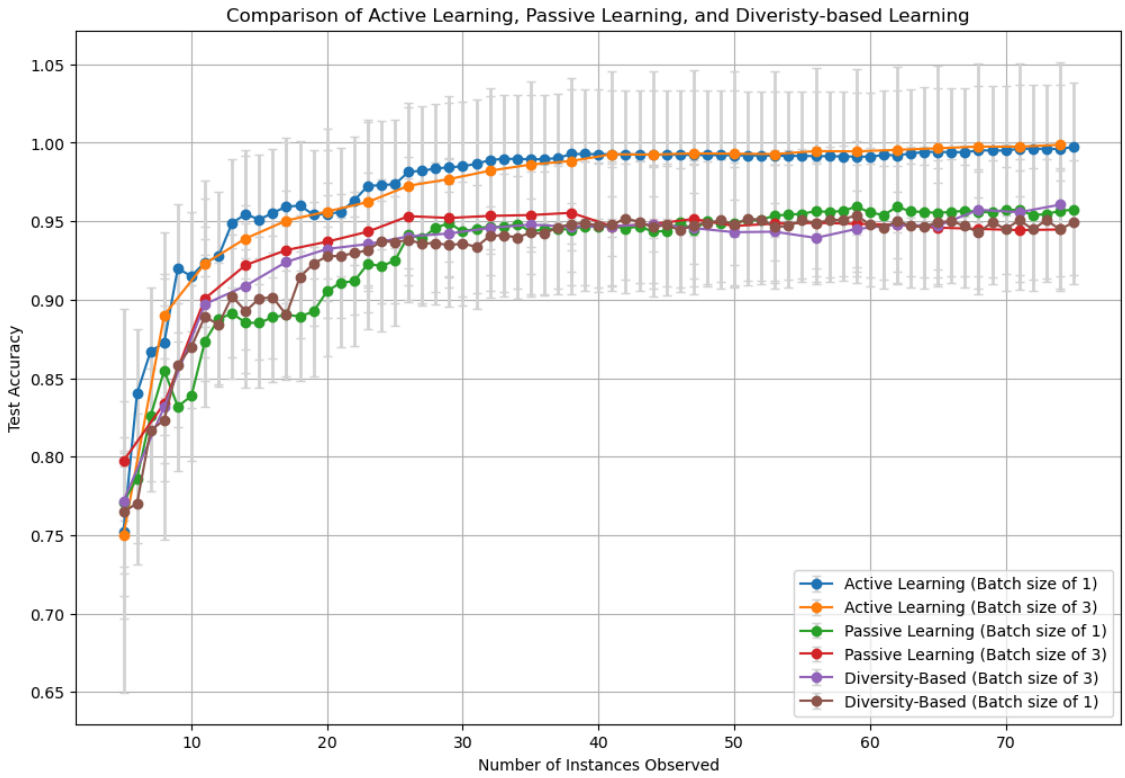
Exercise 1 Batch Selection

a. Make sure to also include comparison against passive learning for a batch size of three (3). Comment on the differences in performance between each method.



The plot above shows comparison of accuracies obtained using Active and Passive Learning approaches for batch sizes 1 and 3. The y-axis represents the accuracy value obtained from testing the model on unseen data. The x-axis represents the number of instances in the training set. According to the plot, active learning, using uncertainty sampling, tends to outperform passive learning for both batch sizes because it focuses on selecting the most informative samples which leads to more efficient learning. For the batch size of 1, the most uncertain sample which was the sample with the highest entropy was added to the train set. For the batch size of 3, the three most uncertain samples which were three samples with highest entropies were added to the train set. The batch size did not seem to make much difference between active learning as we can see that active learning with a batch size of 1 performs quite similar to active learning with a batch size of 3. However, for passive learning, larger batch size requires more iterations to catch up in accuracy, indicating that random selection might not be as efficient in accelerating learning.

b. Implement a batch-size diversity-based sampling method. Describe how you selected the query selection method and provide corresponding details for that method.

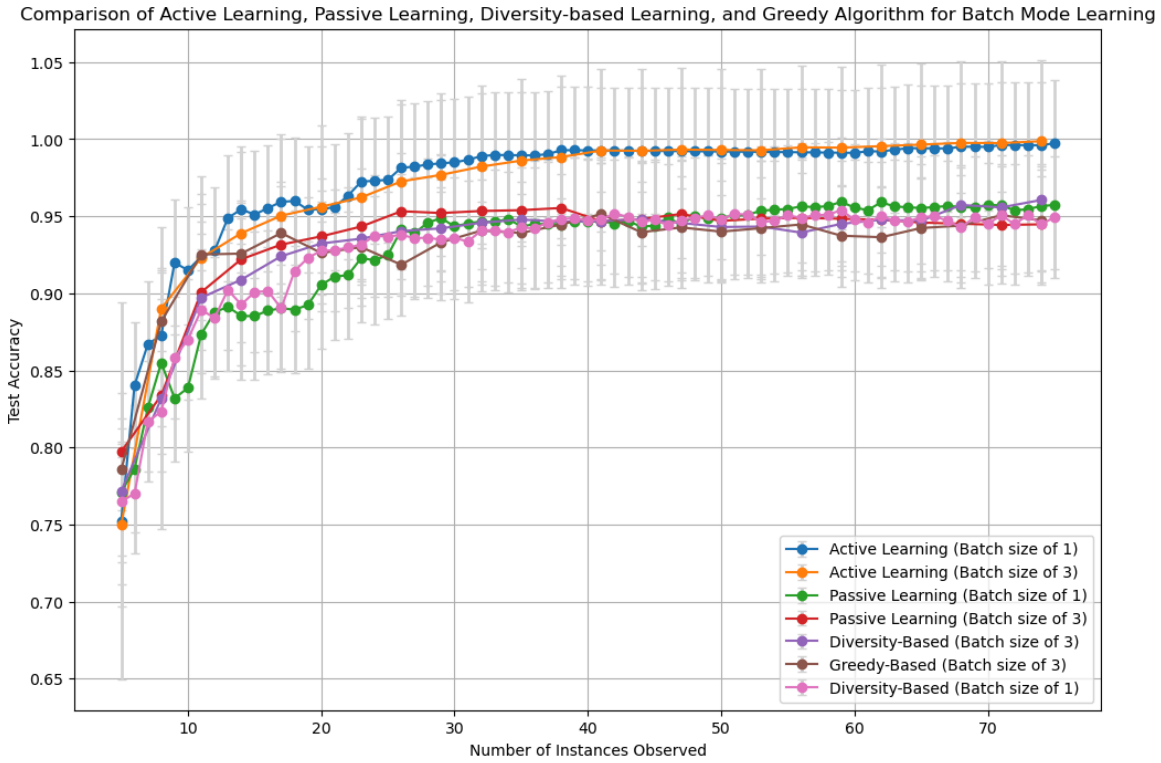


To implement a diversity-based learning method, k-means clustering was done on the train set iteratively with the number of clusters being the batch size. After obtaining the clusters, one random sample from each cluster is added to the train set. This approach ensures that each selected batch contains a diverse set of examples which reduces redundancy and also enhances the diversity in the training set.

According to the plot above, the batch size does not make much of a difference in the performance of the diversity-based approach as we can see that accuracies obtained on the unseen data using both batch sizes are similar. The performance of the diversity-based approach follows a similar curve as passive learning for both the batch sizes. The uncertainty-sampling approach for both batch sizes performs better than the diversity-based approach.

In the implementation of the diversity-based approach, a random sample from each cluster was added to the train set and therefore, there is a chance that the chosen random sample only ensured diversity but did not add much information in terms of variability to the model which led to its slight poor performance compared to the active learning using uncertainty sampling. There is also a chance that k-means clustering with n = batch size was not an appropriate clustering method and therefore the clusters did not represent the patterns present in the data.

c. Hoi et. al.



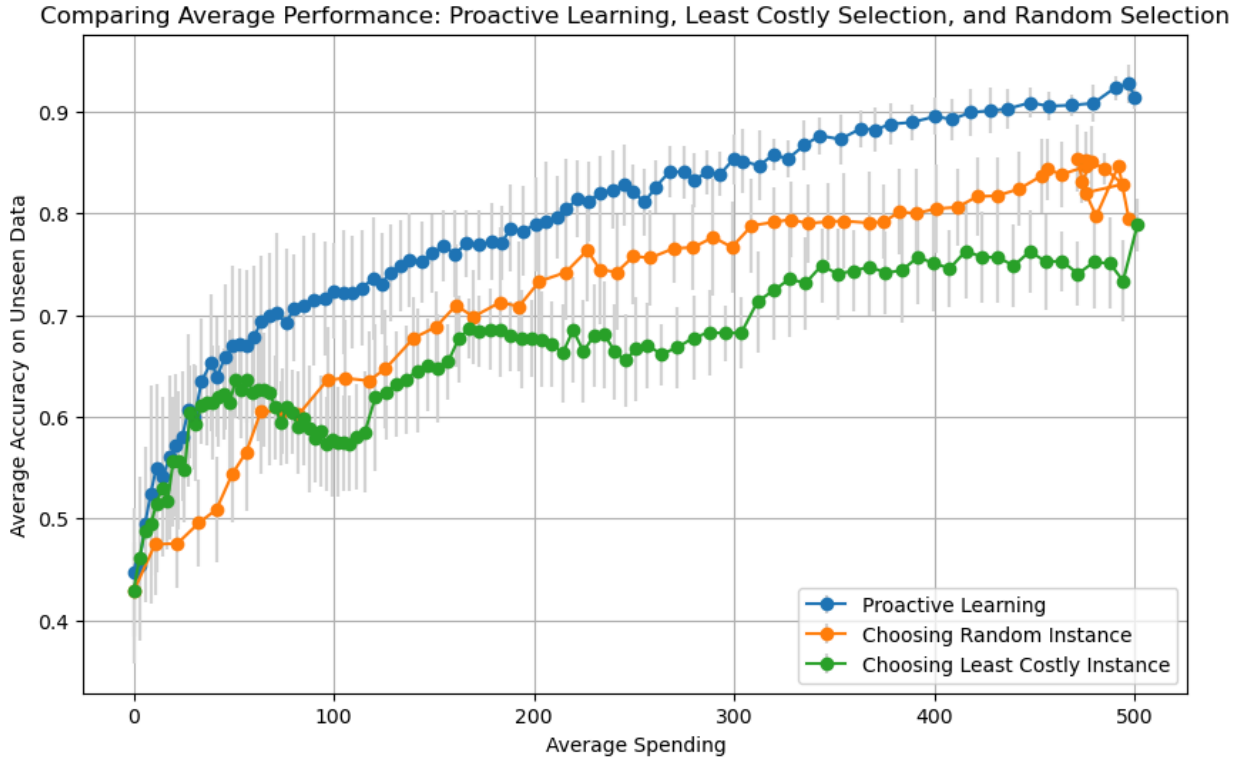
d. Compare the results between part 1a, 1b, and 1c. Discuss whether the results matched your expectations and explain your reasoning.

The greedy-based algorithm iteratively calculates the marginal gain $f(S \cup \{x\}) - f(S)$ to build a subset S of size 3 using the submodular function f(S) presented in the hoi et.al. According to the plot above, the greedy-based algorithm for batch size of 3 performs similarly to the diversity-based approach for batch size of 3 as it shows similar trends in average accuracy values on the unseen data. The greedy-based algorithm starts with higher accuracy values; however, the accuracy value drops around when 30 instances have been added to the training set and resembles the curve of passive and diversity-based learning.

These results don't match my expectations. Given that the greedy-based algorithm selects a subset of samples with the greatest marginal gain, I expected the greedy-based algorithm to outperform diversity-based and passive learning algorithms. One possible reason could be that the subset with largest marginal gain does not always yield the subset with most informative samples which could then lead to a poorer performance of the greedy-based algorithm compared to the active learning algorithm. Since choosing random samples from each cluster during the diversity-based approach might not yield the most uncertain samples, it makes sense that the greedy-based algorithm performs similarly to the diversity-based approach.

Exercise 2 Proactive Learning

Comment on the differences between the performance of the selection strategies. Discuss whether the results matched your expectations and explain your reasoning.



The plot above shows the comparison of average accuracy on unseen data using three selection methods: (i) choosing least costly instance, (ii) choosing random instance, and (iii) choosing instances using a proactive learning-cost sensitive active learning approach. The y-axis represents the accuracy obtained on the unseen data and the x-axis represents the average amount of spending so that it does not exceed the budget of 500.

According to the plot above, method 3 which is the proactive learning approach outperformed both method 1 and method 2 as it maintains a higher accuracy on the unseen data throughout the course of the simulation. This result matches my expectations. Since the proactive learning approach uses utility based selection by iteratively selecting an instance with the highest utility, it makes sense that it outperformed the methods that focus solely on random selection or cost-based selection. In the proactive learning implementation, the utility quantifies the desirability or value of selecting a specific instance for labeling and adding to the training set. It takes two factors into account: (i) Information gain (entropy score) and (ii) Cost. The information gain measures the entropy of class probabilities where higher entropy indicates a greater uncertainty in the model. Entropy was calculated for each instance and was also normalized using the maximum entropy. The Cost represents the cost associated with labeling an instance. Costs were also normalized using the maximum cost. The normalized costs were subtracted from the normalized entropies to find the instance with the highest utility so that we can get the highest amount of variability/information for a lower cost. Therefore, since this implementation of proactive learning uses utility based selection, it optimizes the balance between information gain and cost and ends up performing better than random selection and cost-based selection.