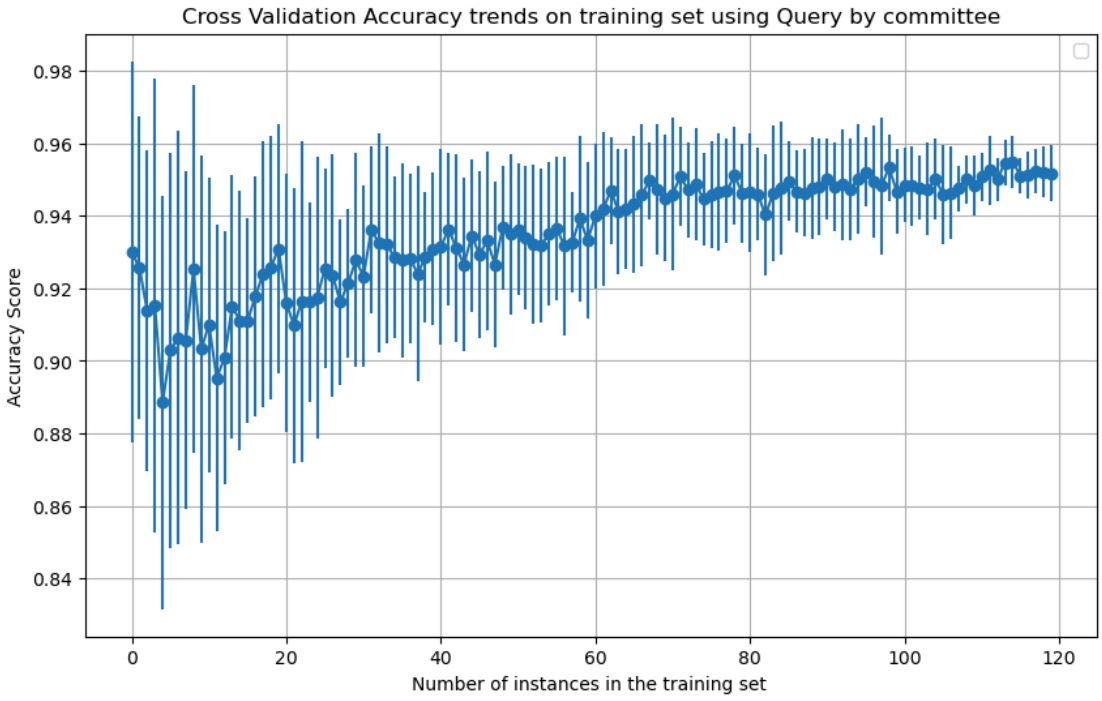
02750 HW 2
Anushka Sinha
Feb 25, 24

Exercise 1 Systematic Evaluation of Heuristic Query Selection Methods

a. **Description of the dataset, base learner, and loss function**
The dataset used in exercise 1 is the Iris dataset. It contains 150 instances, each with five properties, out of which 4 are features and 1 is class label. The four features are sepal_length, sepal_width, petal_length, and, petal_width. There are three class labels: Iris setosa, Iris versicolor, and Iris virginica. These labels were converted to [0,1,2] using a label encoder. The dataset is used to classify the species of 150 instances based on the four features listed above. The base learner used is the RandomForest classifier which is an ensemble method that uses multiple decision trees. Each decision tree produces a prediction and the label with the most votes gets chosen as that instance's label.

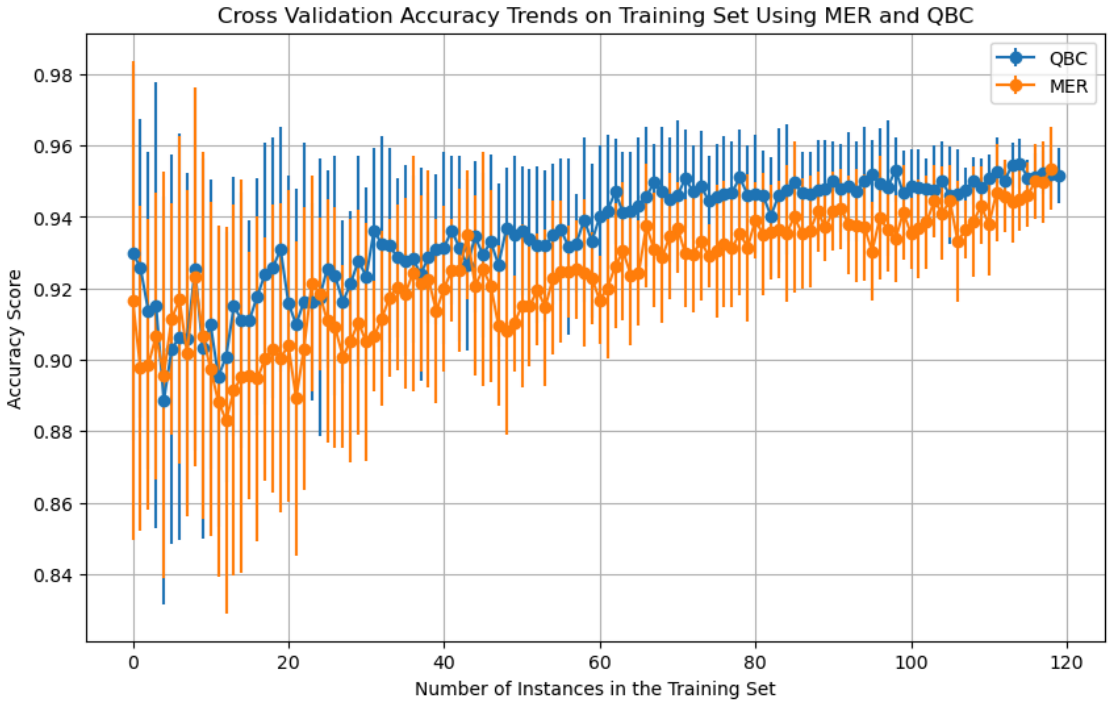
b. **Query by committee: Make sure to concisely describe (i.e., no more than a few sentences), how you constructed and maintained your committee as well as quantified committee disagreement for your learner.**



The committee was constructed with ten RandomForest classifiers. Since RandomForest is an ensemble method, it tends to reduce the correlation between the errors which leads to a diversity in the opinions of the models. The committee disagreement was quantified using hard vote entropy which selects an instance where the votes are closest to being a

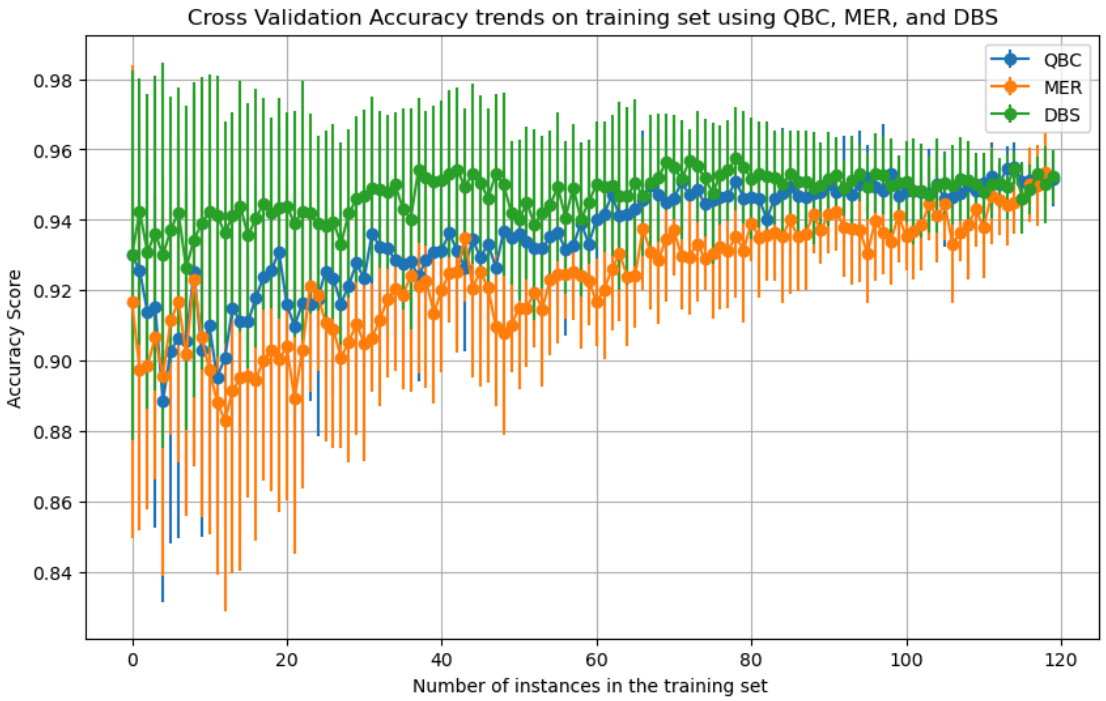
tie. A high hard vote entropy indicates that the models show the greatest disagreement about the true class of that instance. These are the cases that are perhaps the most informative for training because the committee is least convinced about the correct class in these instances.

c. Minimization of expected risk: Make sure to concisely describe (i.e., no more than a few sentences), how you selected the query selection method and provide corresponding details for that learner.



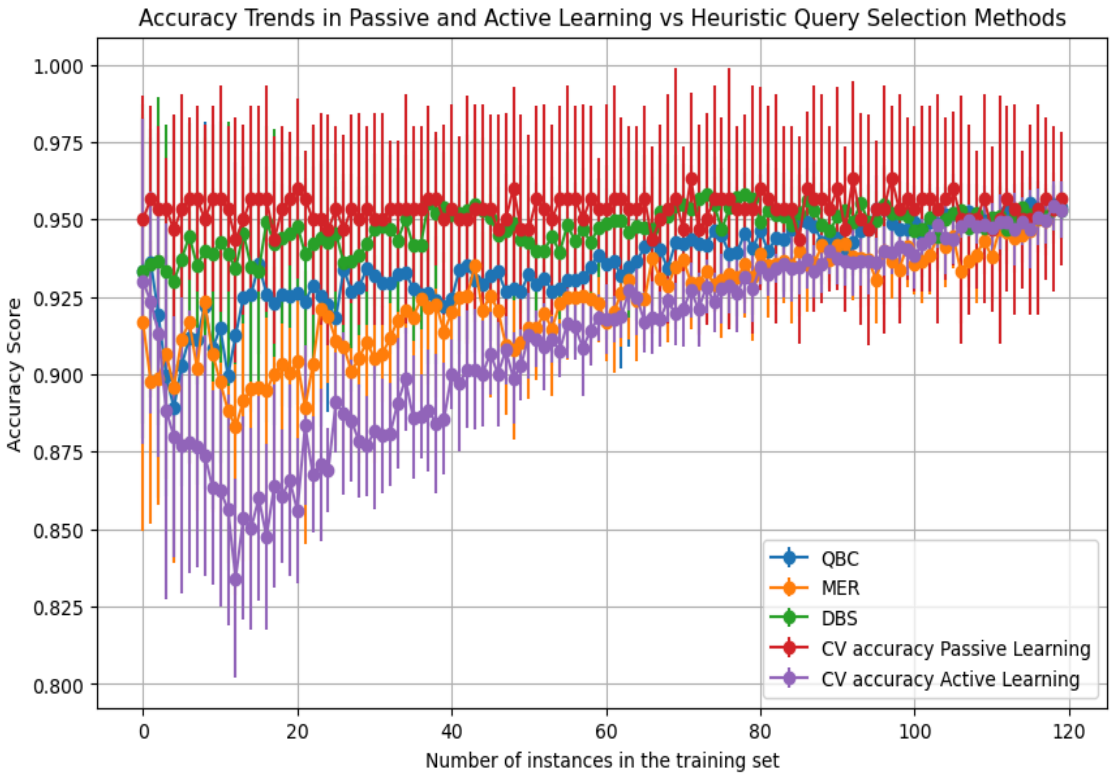
Minimization of expected risk is a query selection method that selects instances which are predicted to reduce generalization error by learning separate models that are each conditioned on one of the possible labels for each instance coming from the unlabeled pool. For each instance in the unlabeled pool, a model is conditioned of every possible label for that instance. The key idea is to observe how much the generalization error of the model would decrease if it were trained on this specific instance with each possible label. The instance that is estimated to show the greatest reduction in the generalization error is added to the training set with the idea that labeling this specific instance will enable the model to generalize better to the new and unseen data.

d. Density-based sampling: Make sure to concisely describe (i.e., no more than a few sentences), how you computed pairwise similarity between instances as well as the query selection strategy (i.e., function ϕ described in class)



Density-based sampling is a query selection strategy that, unlike other query selection methods, considers the representation or density of instances in the sample. For every instance in the unlabeled pool, its similarity with every other instance was calculated using the cosine similarity function which is done by taking the dot product of the two vectors and dividing it by the product of their magnitudes. The mean similarity was calculated over the unlabeled pool and is called the cosine similarity score. Using a beta parameter, one can control the importance given to this criteria. The beta in this assignment was set to 1. The ϕ used was the hard vote entropy score of the query by committee method. The instance selected to be added to the train set is the one with highest product of its hard vote entropy score and its cosine similarity score which ensures that the model receives an instance with high representation or density in the dataset and is also the one that the committee is most uncertain about.

e. Plot



f. Compare the results between parts 1b-1e. Discuss whether the results matched your expectations and explain your reasoning.

The plot above displays the comparison of average cross-fold accuracy trends with standard deviation of passive and active (uncertainty sampling) with the heuristic query selection methods such as query by committee, minimization of expected risk, and density based sampling. The x-axis represents the number of instances added to the training set and y-axis represents the accuracy score.

Based on the plot above, density-based sampling does the best in terms of maintaining a higher accuracy score compared to other heuristic methods. This observation matches my expectation since I expected the density based sampling to perform the best initially by achieving a higher accuracy score and show the least amount of drop in the accuracy score throughout the simulation. Since the density-based sampling algorithm attempts to choose instances that are similar to each other, it allows the model to learn faster which is evident in the faster rise in the accuracy score. The accuracy score drops slightly whenever it encounters an instance that is not too similar to the other instances that were selected before.

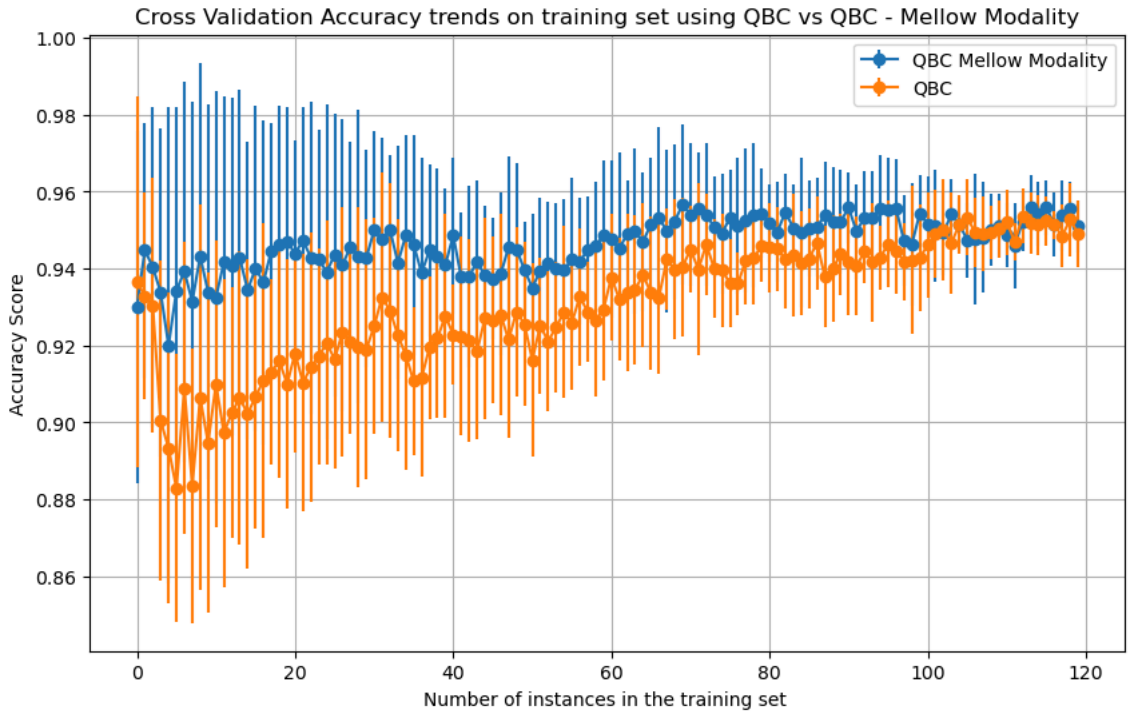
The plot also shows that minimization of expected risk continues to have the lowest accuracy score which goes against my expectation since I expected this method to minimize the overall risk of incorrect predictions. Minimization of expected risk attempts to minimize the expected generalization error which is a measure of how well a model

will perform on the unseen data and since it is one of the most important property to optimize on while performing active learning, it was expected that minimization of expected risk will produce the highest accuracy scores. Few reasons for this observation could be that since the implementation of minimization of expected risk was focused on optimization to reduce the running time, it may have missed a computational step, or that the first twenty percent of the instances added to the train set is not representative of the overall distribution of dataset which could cause the model to make poor selection of instances.

The plot above also shows that passive learning continues to maintain a high accuracy score throughout the simulation which matches my expectation since such a training simulation is prone to overfitting.

Exercise 2 Aggressive vs Mellow

Make sure to concisely describe your approach to convert it to a mellow modality.
The query by committee method was converted to a mellow modality by modifying the hard vote entropy function. Instead of choosing the instance that produces the highest entropy score to be added to the train set, the method involves considering a pool of instances above a certain entropy threshold. The entropy score for the instances that are larger than a certain threshold was stored in a list which indicates that all of these instances are valuable enough to be added to the train set for the model to learn. From this list of instances, a random instance is chosen to be added to the train set to add some variability and mellowness into the selection process.



Compare it to the original version of the query selection method.

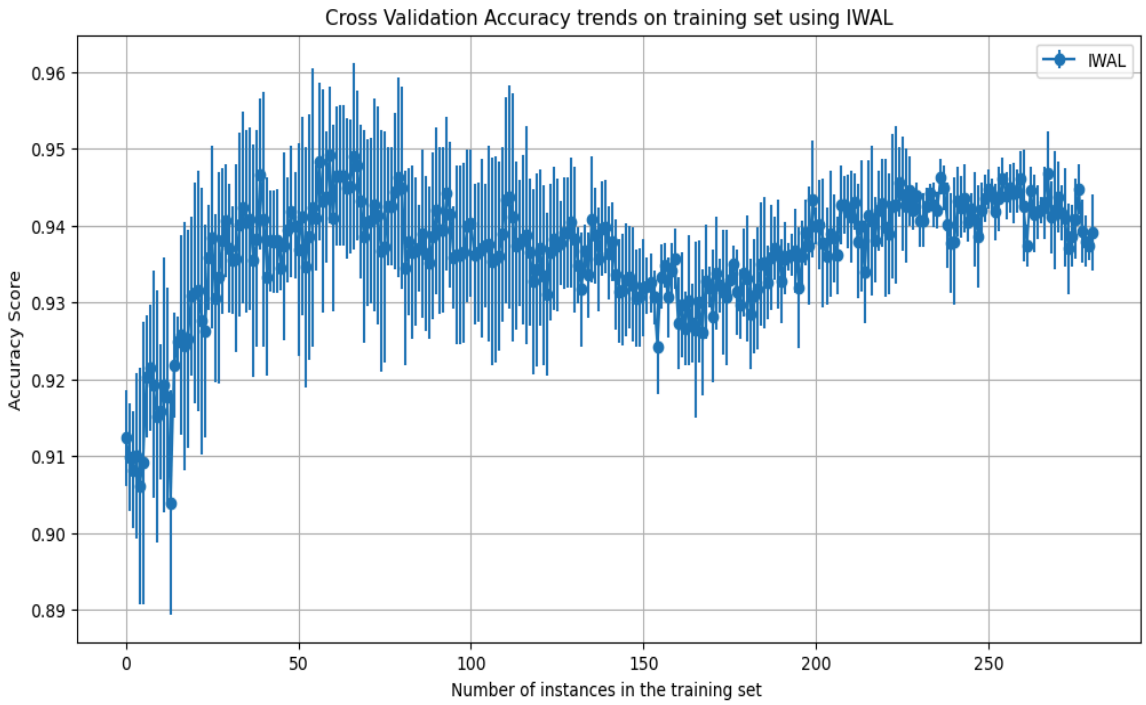
The plot above displays the comparison of average cross-fold accuracy trends with standard deviation of query by committee (QBC) and query by committee - mellow modality (QBC- mellow modality). The x-axis represents the number of instances in the training set and the y-axis represents the average accuracy score. Based on the plot above, both forms of QBC display a similar trend which is an overall increase in the accuracy score as the number of instances in the training set increases. However, the mellow modality of QBC performs better in terms of resulting in a higher accuracy score throughout the simulation of 10 seeds than the non-mellow modality of QBC. This observation matches my expectation since a mellow method finds the instances in the disagreement regions of the models in the committee and a random instance out of those instances is added to the train set. The randomness allows for a reduced risk of overfitting in the mellow modality compared to the original modality of QBC which tends to increase the accuracy score. The mellow QBC also shows a more stable performance in terms of maintaining a relatively similar accuracy score throughout the simulations compared to the non-mellow QBC which shows an increase of 0.4 as more instances are added in the training set.

Exercise 3 IWAL Algorithm

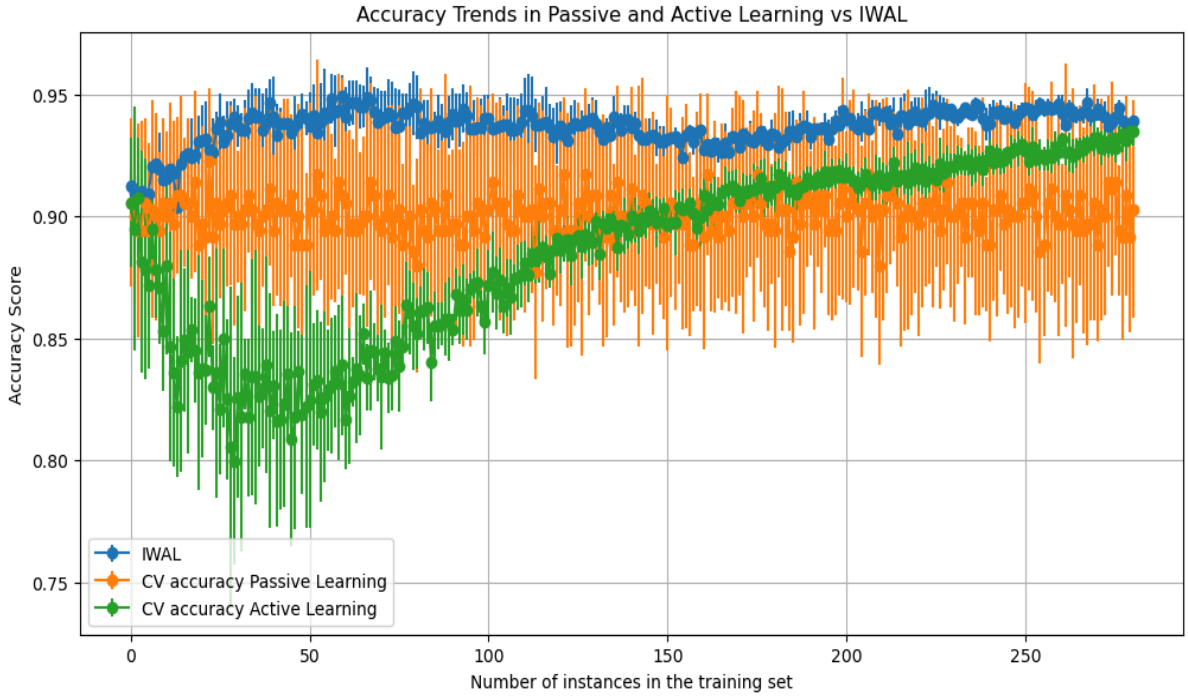
a. **Plot of IWAL**

Description of the dataset, base learner, and loss function

For this binary classification task, the ionosphere dataset was chosen from the UCI repository. This dataset contains the radar data collected from a system in Goose Bay, Labrador. It consists of radar signals returned by the ionosphere and the aim of the dataset is to classify these radar signals as “good” or “bad”. The “good” radar signals are characterized by the presence of some specific type of structure in the ionosphere whereas the “bad” radar signals are recognized by the absence of any specific structure and then passing through the ionosphere. The size of the dataset is 351 instances, each with 35 properties, out of which 34 are features/attributes used for classification of the radar signals. The 35th property is the class label for each instance which was originally “g” for good and “b” for bad but was converted to a [0,1] labels using a label encoder. The base learner used is the RandomForest classifier which is an ensemble method that uses multiple decision trees. Each decision tree produces a prediction and the label with the most votes gets chosen as that instance’s label. The log loss function was used as the loss function in the rejection threshold subroutine.



b. Plot of IWAL, passive and active (uncertainty sampling)



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

The plot above displays the comparison of average cross-fold accuracy trends with standard deviation of IWAL vs Passive and Active Learning (uncertainty sampling). The x-axis represents the number of instances in the training set and the y-axis represents the average accuracy score. Based on the plot above, the IWAL algorithm outperforms both passive learning and active learning using uncertainty sampling by maintaining the highest accuracy score throughout the simulation across 10 different seeds. The passive learning maintains a consistent accuracy score throughout the simulation whereas the active learning (uncertainty sampling) shows a significant improvement in its accuracy score after at least 100 instances were added in the train set.

These observations match my expectation since IWAL is a mellow method which accounts for the sampling bias using importance weighting to return a near optimal model. Instead of just relying on the most uncertain instance to improve the performance which could potentially lead to selection bias, the IWAL algorithm incorporates importance weights for the instances that are chosen to be included in the training set. These importance weights are higher for the instances that are underrepresented in the dataset and therefore ensures that the training set is not just a subset of the dataset that produces the most amount of uncertainty but is in fact a more representative subset of the overall distribution of the dataset. Using this compensation method for the underrepresented instances in the dataset, it allows the model to reduce sampling bias, thereby increasing the accuracy score throughout the simulation.