**Name: Ahmed Khalaf Mohammed ID: 20190028**

**Name: Youssef Ahmed Hamed ID: 20190627**

**Report**

**Introduction & setup**

Active learning (AL) is a learning model that helps improve model accuracy by selecting data points for labeling, rather than relying on randomly selected data points. AL algorithms aim to increase the accuracy of models by repeatedly selecting the most informative samples from a large, unlabeled dataset, which are then classified by an expert or the model itself. In this report, we work on three strategies for AL: margin sampling, entropy sampling, and least confidence sampling in three datasets: MNIST, Wisconsin breast cancer, and iris.

MNIST dataset:

The MNIST dataset consists of 70,000 grayscale images of handwritten numbers, divided into 60,000 training and 10,000 test samples. In this experiment, we initialize a small, labeled set of 50 data points and set the number of instances per query to 20. We then evaluate the three AL strategies in this data set.

Wisconsin breast cancer dataset (unbalanced dataset):

The Wisconsin breast cancer dataset consists of 569 breast cancer patient samples, with 30 features describing characteristics of the cell nucleus. In this experiment, we randomly select 10 samples as the initial labeled data and set the number of instances per query to 10. We then evaluate the three AL strategies in this dataset.

Iris Dataset:

The Iris dataset consists of 150 samples of iris flowers, with four features describing the characteristics of each sample. In this experiment, we randomly select 5 samples as the initial labeled data and set the number of instances per query to 5. We then evaluate the three AL strategies in this dataset.

**Results**

# At MNIST dataset with 3 Active Learning Strategy

# Margin Sampling Strategy

# 

# Entropy Sampling Strategy

# 

# Least Confident Sampling Strategy

# 

# Plotting the results

# Training accuracy

# 

# Testing accuracy

# 

# At Breast Cancer Wisconsin dataset(unbalanced dataset) with 3 Active Learning Strategy

# Margin Sampling Strategy

# 

# Entropy Sampling Strategy

# 

# Least Confident Sampling Strategy

# 

# Plotting the results

# Training accuracy

# 

# Testing accuracy

# 

# At Iris dataset with 3 Active Learning Strategy

# Margin Sampling Strategy

# 

# Entropy Sampling Strategy

# 

# Least Confident Sampling Strategy

# 

# Plotting the results

# Training accuracy

# 

# Testing accuracy

# 

# Discussion of the results

# At MNIST dataset

# As shown in the results of MNIST dataset, Margin Sampling performed the best among the three strategies in the MNIST dataset, with a test accuracy of 93.2% after 20 rounds of querying. Margin sampling selects samples close to the decision boundary between two classes, making it effective in selecting the most uncertain samples. Entropy sampling and least confidence sampling also performed well on this dataset, with test accuracy of 88.7% and 90.9%, respectively, after 20 rounds of querying. Entropy Sampling selects samples with high uncertainty or uncertainty, while Least Confident Sampling selects samples about which the model is less confident.

# At Breast Cancer Wisconsin dataset(unbalanced dataset)

# As shown in the results of Breast Cancer Wisconsin, margin sampling, Entropy Sampling and least confidence sampling were performed similarly on the Wisconsin breast cancer dataset, with a test accuracy of 97.4% after 10 rounds of query. The results indicate that all three strategies are effective in identifying the most informative samples for labeling this dataset.

# At Iris dataset

# As shown in the results of Iris dataset, margin sampling, entropy sampling and least-confidence sampling were performed in the same way on the iris dataset, with a test accuracy of about 96.7% after 5 rounds of query. The results indicate that all three strategies are effective in identifying the most informative samples for labeling this dataset.

# Conclusion

# In this report, we evaluated three strategies for AL: margin sampling, entropy sampling, and least confidence sampling in three datasets: MNIST, Wisconsin breast cancer, and iris. The results indicate that the effectiveness of these strategies varies across the data set. Margin sampling performed best on the MNIST dataset, while all three strategies performed similarly well on the Breast Cancer Wisconsin dataset and the Iris dataset.

# Overall, this paper emphasizes how crucial it is to choose an AL approach based on the data set being employed. All three methods work well at finding the most useful samples for labelling, although depending on the dataset, their effectiveness may differ. To improve the quality of their models, researchers should thoroughly assess various AL techniques on their data sets.