 

Faculty of computer and Artificial Intelligence Cairo university

Assignment #

|  |  |  |
| --- | --- | --- |
| Name | ID | Group # |
| Noor Elden Tariq Mohammed Medhat | 20210426 | ALL |
|  |  |  |
|  |  |  |
|  |  |  |

Supervised by

Eng. Eman Salah Eng. Rana Mohamed

2024-2025

# Table of Contents

* Introduction

Active learning is a machine learning technique where the model is allowed to choose the data from which it learns. Instead of training on a fully labeled dataset, the model actively selects the most informative or uncertain samples to be labeled by an oracle (typically a human annotator). This approach is especially useful when labeled data is scarce or expensive to obtain. By focusing on the most valuable data points, active learning can significantly reduce labeling costs while maintaining high model performance. It is commonly used in applications like image classification, natural language processing, and medical diagnosis. Overall, active learning helps build more efficient and effective models with less labeled data.

# What the difference between Active learning and passive learning

Active learning lets the model choose which data it wants to learn from, focusing on the most useful examples. In contrast, passive learning uses whatever labeled data is available without any selection. The key difference is that active learning is selective, while passive learning is not.

* Datasets

The **Breast Cancer dataset**, often used in classification tasks, contains features computed from digitized images of fine needle aspirates (FNAs) of breast masses. These features describe characteristics of the cell nuclei present in the images. The dataset is typically labeled as either *malignant* or *benign*. It is known to be **imbalanced**, with more samples labeled as benign than malignant, which can bias machine learning models towards the majority class if not handled properly.

The **Iris dataset** is a classic dataset in machine learning and statistics, containing 150 samples of iris flowers from three species: *setosa*, *versicolor*, and *virginica*. Each sample includes four features: sepal length, sepal width, petal length, and petal width. Unlike the breast cancer dataset, the Iris dataset is **balanced**, with an equal number of samples (50) from each species, making it ideal for testing classification algorithms under fair conditions.

# Strategies

# here is a brief description of the three strategies that have been used:

 **Uncertainty Sampling**: This strategy selects the data points where the model is least confident in its predictions. For example, in binary classification, it might pick samples with predicted probabilities close to 0.5. The idea is to label the most uncertain instances to improve the model's understanding.

 **Query by Committee (QBC)**: This method involves training multiple models (a committee) on the current labeled data. The unlabeled instances that the models most disagree on are selected for labeling. The assumption is that disagreement reflects uncertainty and labeling such points helps improve the overall model.

 **Expected Error Sampling**: This strategy selects samples based on the expected reduction in the model's future error. It estimates how much adding the label of a certain instance would reduce the model's generalization error, and picks the one with the highest expected improvement.

# Experiments (methodology)

What process occurred in each round for each strategy for each dataset?

# Results

What are the results in each round for each strategy for each dataset? And The comparison of all strategies with the two datasets for each round

# Conclusion

What is the optimal strategy for each dataset, and after how many rounds and with how many samples can the best accuracy be achieved?

# References

Each paper or link helps you in this project