Active Learning Strategies ModAL

Maymounah– 20190657
Ali Adel Ali – 20190337

22/4/2023

—

Selected Topics in AI_2

—

Ta/ Salah

Datasets

1- IRIS dataset:

The dataset consists of measurements of the sepal length, sepal width, petal length, and petal width of 150 iris flowers belonging to three different species: Setosa, Versicolor, and Virginica.

2- Wine dataset

It contains the results of a chemical analysis of wines grown in the same region in Italy, but derived from three different cultivars. The dataset consists of 178 samples, with 13 chemical features for each sample.

3- IRIS with Unbalanced

An unbalanced dataset is a dataset where the class distribution is not even or equal, meaning that one or more classes have significantly more instances than the others. This can cause problems for machine learning algorithms, as they may be biased towards the majority class and perform poorly on the minority class.

Strategies We Used:

1- Random sampling:

random sampling is a basic strategy for selecting data points to query the oracle.
Random sampling in modAL involves randomly selecting a batch of unlabeled data points from the pool of available data and presenting them to the oracle for labeling. The size of the batch is a parameter that can be tuned to balance the need for exploring the space of possible data points and exploiting the most informative ones.

2- Uncertainty Sampling:

Uncertainty sampling is a common strategy for active learning in which the algorithm selects the data points with the highest uncertainty for labeling. In modAL, uncertainty sampling can be implemented using a variety of uncertainty measures, such as entropy or margin-based measures.

3- Margin Sampling

Margin sampling is an uncertainty-based active learning strategy that selects the data points that have the smallest difference between the top two predicted class probabilities. In other words, it selects the data points for which the model is most uncertain about which class they belong to.

4- Entropy Sampling

Entropy sampling is an uncertainty-based active learning strategy that selects the data points with the highest entropy

5- Query-By-Committee with PCA

is a model disagreement-based active learning strategy that selects the data points for which the predictions of the committee of models disagree the most. In other words, it selects the data points for which the models are the most uncertain about their class labels.

Steps For strategy:

1- Read the dataset
2- Apply PCA to have a good analysis on it
3- Split the data to train/ validate/ test
4- Initialize an active learner object with specific strategy and a simple machine learning model
5- Find the performance for the model before apply querying
6- Apply active learning technique by querying on the validation dataset and save our performance after each query.

IRIS Dataset:

Strategy	Estimator	# of queries	Initial accuracy	Final Accuracy
Random sampling	RandomForestClassifier	54	98%	100%
Uncertainty Sampling	KNeighborsClassifier	20	33.3%	90%
Margin Sampling	RandomForestClassifier	20	83.3%	97.3%
Entropy Sampling	RandomForestClassifier	20	64.6%	96%
Query-By-Committee with PCA	KNeighborsClassifier	20	98.3%	85.39%

The best Accuracy get when use the Random sampling because the random sampling It is an unbiased sampling method, meaning that each sample in the dataset has an equal chance of being selected for the training or testing set.

Wine Dataset:

Strategy	Estimator	# of queries	Initial accuracy	Final Accuracy
Random sampling	RandomForestClassifier	54	35.3%	97.7%
Uncertainty Sampling	KNeighborsClassifier	20	33.1%	96.1%
Margin Sampling	RandomForestClassifier	20	52.8%	98.3%
Entropy Sampling	RandomForestClassifier	20	47.1%	98.3%
Query-By-Committee with PCA	KNeighborsClassifier	20	98.3%	85.39%

The best Accuracy get when use the margin and entropy sampling because the goal of both of them select the most informative samples for labeling by the human annotator

IRIS with Unbalance:

Strategy	Estimator	# of queries	Initial accuracy	Final Accuracy
Random sampling	RandomForestClassifier	54	100%	90%
Uncertainty Sampling	KNeighborsClassifier	20	97.1%	92.6%
Margin Sampling	RandomForestClassifier	20	100%	88%
Entropy Sampling	RandomForestClassifier	20	100%	89.3%

The best Accuracy get when use the Random sampling