**ML Assignment -2**

**Group-11**

Archana - 2019H1030519P

Sanchita-2019H1030520P

Yashita - 2019H1030521P

The key idea behind active learning is that a machine learning algorithm can achieve greater accuracy with fewer labeled training instances if it is allowed to choose the data from which is learns.

An active learner may ask queries in the form of unlabeled instances to be labeled by an oracle (e.g., a human annotator). Active learning is well-motivated in many modern machine learning problems, where unlabeled data may be abundant but labels are difﬁcult, time-consuming, or expensive to obtain.

Active learning systems attempt to overcome the labeling bottleneck by asking queries in the form of unlabeled instances to be labeled by an oracle. In this way, the active learner aims to achieve high accuracy using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data

**Q1. Implement stream-based and pool-based scenarios.**

**Stream-based :**

The key assumption is that obtaining an unlabeled instance is free (or inexpensive), so it can ﬁrst be sampled from the actual distribution, and then the learner can decide whether or not to request its label. This approach is sometimes called stream-based or sequential active learning, as each unlabeled instance is typically drawn one at a time from the data source, and the learner must decide whether to query or discard it.

**Pool-based:**

assumes that there is a small set of labeled dataLand a large pool of unlabeled data U available. Queries are selectively drawn from the pool,which is usually assumed to be closed (i.e., static or non-changing), although this is not strictly necessary. Typically, instances are queried in a greedy fashion, according to an informativeness measure used to evaluate all instances in the pool

The main difference between stream-based and pool-based active learning is that the former scans through the data sequentially and makes query decisions individually, whereas the latter evaluates and ranks the entire collection before selecting the best query

**Q2. Implement the following query strategy frameworks:**

These strategies are used for analysing the informativeness of the unlabeled data.

* **Uncertainty sampling**

There are four types of Uncertainty Sampling:

Least Confidence: difference between the most confident prediction and 100% confidence

Margin of Confidence: difference between the top two most confident predictions

Ratio of Confidence: ratio between the top two most confident predictions

Entropy: difference between all predictions, as defined by information theory

We have used margin of confidence in our implementation.

**Test Accuracy:** 90.2%

* **Query by committee (Vote entropy, KL divergence)**

The QBC approach involves maintaining a committee C = {θ(1),...,θ(C)} of models which are all trained on the current labeled set L, but represent competing hypotheses. Each committee member is then allowed to vote on the labeling of query candidates. The most informative query is considered to be the instance about which they most disagree.

**Test Accuracy :** 90%

* **Diversity sampling**

In this approach, we pick points of each class/type. We first cluster the unlabeled data(no. of clusters based on the budget). Then query few points from each cluster and fit the model with the newly labeled points. To select points from each cluster we have used the difference between distances from cluster centroid as a measure for uncertainty.

**Test Accuracy :** 88.88%

**Q3. Devise and implement a cluster-based strategy for data labeling, given a limited budget. It is a requirement that no. of wrong labelings is minimized.**

**Clustering-based**

This is a heuristic-based approach. The data is clustered using KMeans where k can be decided based on the budget. The unlabeled is clustered clusters are then labeled using the majority of these queried labels. This fully labeled data is now fit to the classifier(Random Forest).

For choosing the representative points, we have implemented two methods:

* Picking points randomly(based on the budget)
* Picking points near the cluster centroids( as they would give a good representation of the entire cluster)

**Results:**

The dataset that we have used is the “Wine dataset”.

K-means with k=2 was used for clustering the training data(train split=75%)

**Case 1:** **Random Sampling**

Train Accuracy: 69.17%(rightly labeled data)

**Test Accuracy:**  75.33%

**Case 2:** **Points near centroids**

Train Accuracy: 55.63%(rightly labeled data)

**Test Accuracy:** 57.33%