Homework 2: Active Learning + GNNS

Naomi Derel 325324994, Gili Cohen 326280815, Renana Shachak 213920010

01.08.2024

1 Active Learning

1.1 Pipeline Structure

The pipeline we used for active learning is as follows:

1. Load the dataset and split it into train and test indices. Find the features and labels from the dataset.
2. For a set number of iterations:
    - Train a model on the labeled set.
    - Use the model to predict the labels of the unlabeled set, and calculate an accuracy score.
    - Select a set number of samples from the unlabeled set to be labeled, using one of the selection methods as set.
3. Output the accuracy scores for each iteration. Alternatively, output the final trained model.

1.2 Uncertainty-Based Selection

The measure we picked for uncertainty-based selection is entropy.

The entropy of a distribution is a measure of the uncertainty in that distribution, and so it is a classical measure for our case.

We used it by calculating the entropy of the output of the model for each sample in the pool, and defining them as an estimate for the uncertainty of the model on that sample. We then selected the samples with the highest entropy to be labeled next, since they are the samples the model is most uncertain about.

1.3 Custom Selection

The measure we picked for custom selection is the density selection measure we saw in class.

the density selection is based upon the intuition that the model will benefit the most from samples that extend the vector space the model has seen so far. We used it by calculating the distance of each sample in the pool from the samples the model has seen so far, and then selecting the samples with the lowest density.

For defining the density measure, we tried two approaches (where x_i are the samples the model has seen so far, and x is the sample we are calculating the density for):

1. The lowest sum of distances from the samples the model has seen so far: $\sum_{i=1}^{n} distance(x_i, x)$.
2. A density measure based on the sum of exponential distances from the samples the model has seen so far: $\sum_{i=1}^{n} \exp(-distance(x_i, x))$. This is based on a Gaussian kernel, and is a common measure for density.

We then selected the samples with the highest density to be labeled next, which turned out to be:

1.4 Parameter Comparison

We compared the results of multiple combinations of models, iterations, and budget per iteration.

Random Selection

	Random Forest	SVC	Logistic Regression
X Iterations + Y Budget	10	10	10
X Iterations + Y Budget	10	10	10
X Iterations + Y Budget	0.8	0.7	0.9

Table 1: Results of Parameter Comparison for Random Selection

Uncertainty-Based Selection

	Random Forest	SVC	Logistic Regression
X Iterations + Y Budget	10	10	10
X Iterations + Y Budget	10	10	10
X Iterations + Y Budget	0.8	0.7	0.9

Table 2: Results of Parameter Comparison for Uncertainty-Based Selection

Custom Selection

	Random Forest	SVC	Logistic Regression
X Iterations + Y Budget	10	10	10
X Iterations + Y Budget	10	10	10
X Iterations + Y Budget	0.8	0.7	0.9

Table 3: Results of Parameter Comparison for Custom Selection

Analysis and Discussion

We can see that the best results were achieved with the – model, with - iterations and a budget of - samples per iteration. The accuracy score was -.