Homework 2

Thomas Zhang (tczhang)

February 27th 2024

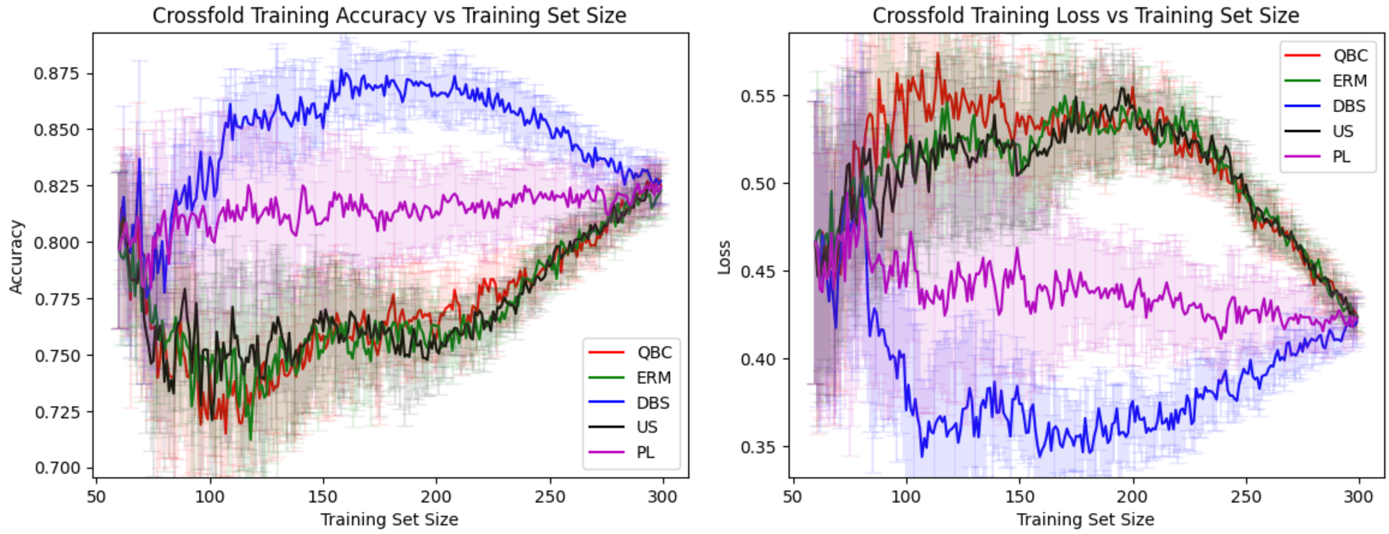

Figure 1: QBC: Querry by Committee. ERM: Expected Risk Minimization. DBS: Density Based Sampling. US: Uncertainty Sampling. PL: Passive Learning. Training set size is number of samples in the training set. 5-fold cross-validation on training data was used to generate graphs.

1. (a) The following data set was used for all three questions. The selected data set is "Heart failure clinical records" from:

   https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records

   There are 300 samples in the dataset. The dataset is a binary classification data set with 12 features. The 12 features are a mix of binary, integer and continuous values.

   There are 5 binary features which are: whether the patient has anaemia, high blood pressure, and/or diabetes. The last two binary features are patients sex and whether they smoke.

   There are 5 integer features: the patients age, the patients CPK enzyme level, the blood percentage leaving the heart each contraction, the blood sodium, and the follow up time in days.

   The final 2 features are continuous: platelets in blood and level of serum creatinine in the blood.

   The class labels are dead and alive.

   The base learner used was a support vector classifier (SVC) with a radial basis function as the kernel. The SVC was used by importing SKlearn's implementation of SVC. The loss function used was hinge loss.

   (b) The committee was created by bootstrapping the training data on each iteration. 50 bootstrapped datasets, each the same size as the labeled dataset, were made. The committee was then formed using the bootstrapped data to train 50 models. The highest disagreement sample was added to the labeled training set. Disagreement was measured using the formula:

$$x*_{KL} = \underset{x\in U}{\operatorname{argmax}} \frac{1}{|M|} \sum_{h\in M} D(P_h(y|\mathbf{x}) \| P_M(y|\mathbf{x}))$$

   where

$$D(P_h(y|\mathbf{x}) \| P_M(y|\mathbf{x})) = \sum_{y'\in \mathbf{Y}} P_h(y'|\mathbf{x}) \log \frac{P_h(y'|\mathbf{x})}{P_M(y'|\mathbf{x})}$$

   is the KL divergence between a single model in the committee P_h and the total average of all committes P_m. P_m is defined as:

$$P_M(y'|\mathbf{x}) = \frac{1}{|M|} \sum_{h\in M} P_h(y|\mathbf{x})$$

   which is the average probability that y' is the true label of $\mathbf{x}$ according to the committee.

   (c) The query selection method chosen was minimizing expected risk. The next sample to be labeled was chosen by using the following the formula:

$$\underset{x\in U}{\operatorname{argmin}} \sum_{y\in \mathbf{Y}} P_\theta(y|\mathbf{x}) \Big(\sum_{x'\in U} 1 - P_{\theta+(x,y)}(\hat{y}|x') \Big)$$

   where P_θ is the current model and $P_{\theta+(x,y)}$ is the model after adding the point (x, y) to the labeled data set.

   (d) The query selection strategy was query by committee (same as in part a). The density of a sample x was calculated using the formula:

$$\Big(\frac{1}{|U|} \sum_{x'\in U} sim(x, x')\Big)^\beta$$

where β is a hyper parameter controlling the importance of the weight and $sim(x, x')$ is the l2-norm of the difference between x and x'.

$$sim(x, x') = \|x - x'\|_2$$

For the simulation in this assignment a β value of 0.7 was used.

(f) From the simulation the query by committee and the uncertainty sampling had lower average cross-fold validation accuracy while training. The passive learning strategy had a relatively constant accuracy throughout the simulation. The expected risk minimization and the density based sampling had a slight increase in average accuracy during training. At the end of the simulation all methods ended at roughly the same accuracy.

This somewhat matched my expectations. The query by committee and uncertainty sampling had a lower cross-validation accuracy at first which is expected since the query selection methods add samples that the model struggles with. And the passive learning method had a relatively stable cross-fold validation accuracy throughout the training process.

However the density based sampling was somewhat surprising as it had higher cross-validation accuracy than passive learning during the most of the iterations. The only time density based sampling had lower cross-validation accuracy was at the very start where only a few of the samples from the unlabeled set were added to the training data.

One potential reason the density based sampling method may have had higher average accuracy during training is as follows. In density based sampling, the chance of a unlabeled samples being selected is based on its similarity to other unlabeled samples. Hence the unlabeled sample that is chosen has a high chance of being similar to something that the model has already seen.

Another somewhat surprising behavior was that the cross-fold validation accuracy of expected risk minimization behaved similar to uncertainty sampling and query by committee. This is surprising because expected risk minimization, unlike query by committee and uncertainty sampling, do not take uncertainty into account.

One possible reason this may have occurred is that expected risk minimization, uncertainty sampling, and query by committee are all ultimately trying to make the model generalized. Hence the samples that have high uncertainty and the samples that minimize classification error could be similar.

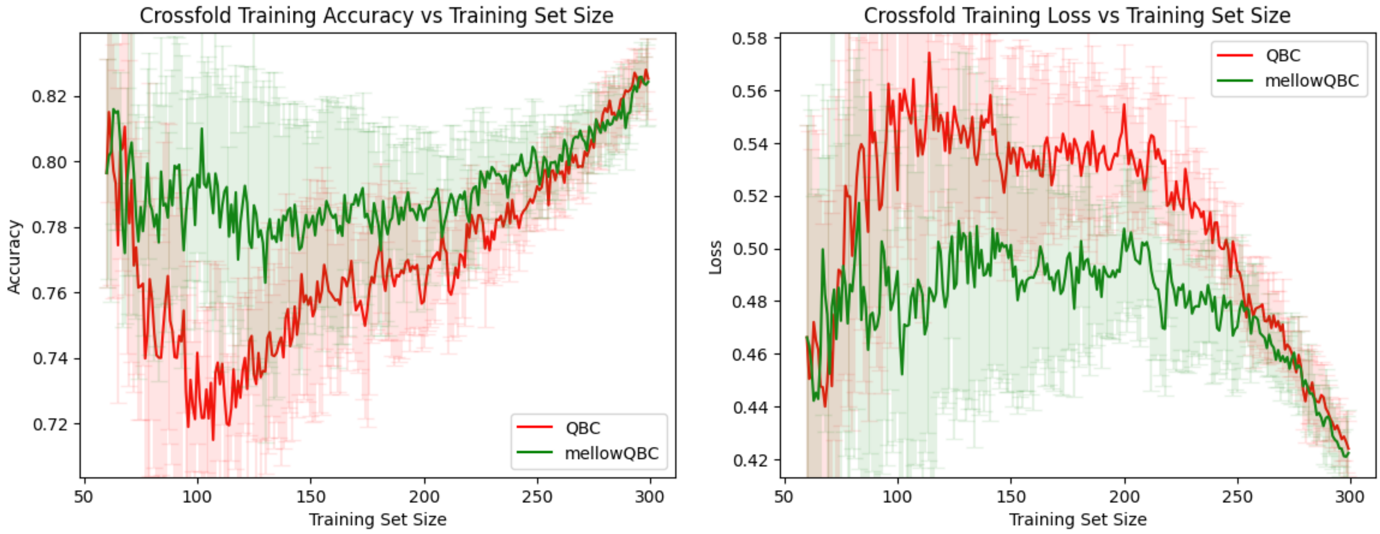

Figure 2: QBC: Aggressive Query by committee. mellowQBC: Mellow Query by Committee. Training set size is number of samples in the training set.5-fold cross-validation on training data was used to generate graphs.

2. (a) The underlying query selection was query by committee. Mellow sampling was done by first converting the disagreement utility scores into a probability distribution by normalizing by the sum of disagreement utility scores. This results in a probability, p_i, for each $x_i \in U$. The probabilities are then used to randomly sample a x_i' from the unlabeled set U with probability of selecting x_i' equal to the probability p_i.

   The mellow and aggressive active learning models perform similarly for the most part. Both experience a dip in accuracy as the training begins that rises again after more of the data is added to the labeled training set. This is as expected since the mellow and aggressive sampling techniques follow the same selection algorithm. It is expected for them to perform similarly.

   The most notable difference between the two methods is that the mellow strategy had less of an initial accuracy drop compared to the aggressive strategy. This is most likely because the aggressive strategy is always choosing the data point which the model struggles the most with while the mellow strategy chooses samples that the model struggles with but not necessarily the sample that the model struggles the most with.

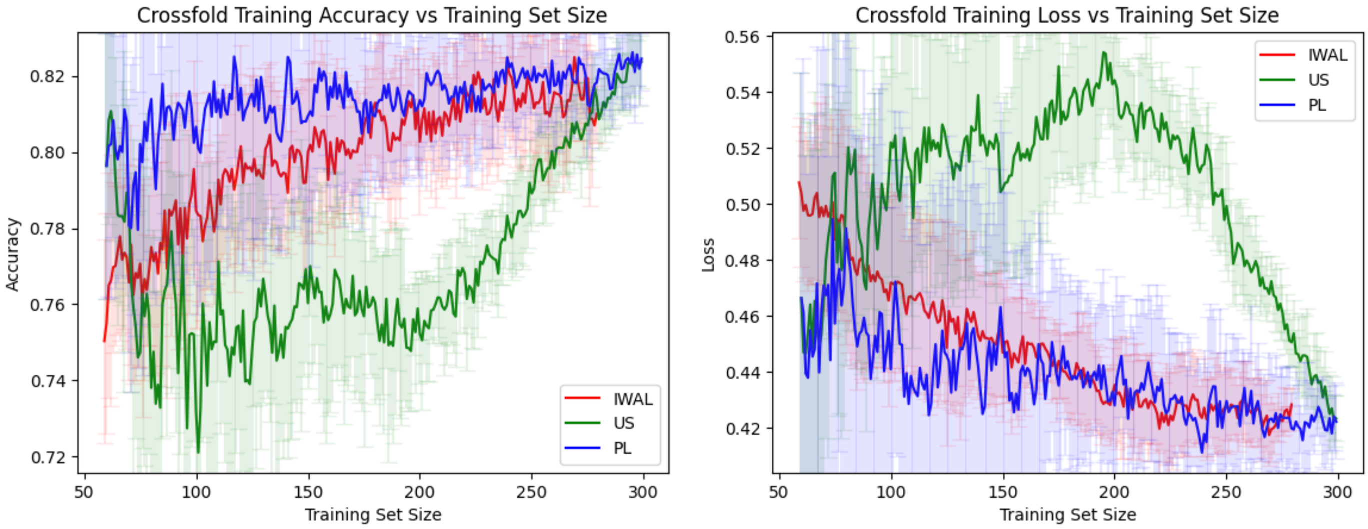

Figure 3: IWAL: Importance Weighted Active Learning. US: Uncertainty Sampling. PL: Passive Learning. Training set size is number of samples in the training set. 5-fold cross-validation on training data was used to generate graphs.

3. IWAL was implemented with a bootstrapping rejection sampling subroutine to determine pt. Log-loss was used to calculate loss for IWAL. A minimum sampling probability of $p_{min} = 0.05$ was used. At a high-level, the bootstrapping IWAL algorithm assigns higher sampling probability to unlabeled samples that the model is unsure about and lower sampling probability to the samples that the model is consistent on.

   The results were as expected for the most part. The passive learning method maintains a relatively stable accuracy across all iterations. The accuracy of uncertainty sampling dropped significantly at the start this method adds data points that the model struggles with.

   Unexpectedly, the cross-validation accuracy of IWAL did not appear to drop like uncertainty sampling. Instead it started at a slightly lower accuracy than the other sampling methods and had an increasing trend until it roughly plateaued at the same accuracy as uncertainty sampling and passive learning.

   One potential reason why the initial accuracy of IWAL was lower is that the samples are weighted in IWAL where as the samples are not weighted in the other two methods.

   One potential reason that IWAL did not have an initial downward trend is that IWAL doesn't affect the order of samples being added to the labeled set. It only assigns a probability of labeling an unlabeled sample and adding it to the training data. Hence IWAL can be thought of as a passive learning strategy with a chance of not adding a sample. Uncertainty sampling on the other hand affects the order of unlabeled samples being added because it always adds the sample that the model had the most difficulty classifying (the most ”uncertain”).