Homework 3

Thomas Zhang (tczhang)

March 28th 2024

1.

(a)

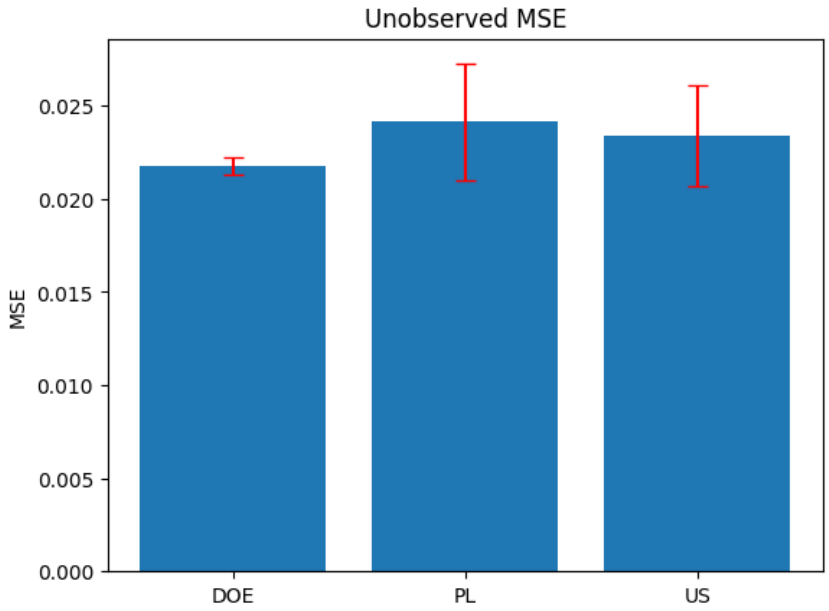

Figure 1: The unobserved MSE (mean squared error). DOE: (E-optimal) Design of Experiments. PL: Passive Learning. US: Uncertainty sampling.

The data set is a regression dataset consisting of two features (numerical floats) and a single label (float). A linear regression classifier from SKlearn was used to predict the label of the samples. There are 100 samples in the dataset.

All quantities were measured after labeling 30 data points. For passive learning and uncertainty sampling, a random 10 data points were chosen to initialize the model. Then 20 cycles of active learning were performed with each cycle adding 1 data point to the labeled set.

DOE was performed using D-optimality. This means choosing the samples that maximizes :

$$|X^T X|$$

To approximate the solution to this maximization problem, a random initial 30 samples were chosen. Then for 20 iterations (20 iterations was chosen so that the "time" the algorithm had to improve the labeled set is equivalent between the active learning methods and DOE), each unlabeled sample was used replace each labeled sample to calculate,

$$|X^T X|$$

. Then the algorithm chooses the unlabeled sample and labeled sample to be replaced that produced the highest D-optimally score to add to the labeled set. If the unlabeled sample produces a better score, then the unlabeled sample is always used to replace sample "i". If the score is worse then the loop terminates because no replacement of a labeled sample with an unlabeled sample will improve the scoring metric.

The results partially matched my expectations. DOE produced the models with lower average MSE than PL but only slightly lower than the MSE of uncertainty sampling. DOE is expected to perform better than uncertainty sampling because DOE allows any of the 30 initial samples to be replaced, while active learning cannot replace the initial labeled samples and can only add more samples to the labeled set. One reason why DOE may have only been slightly better than uncertainty sampling is that the true D-optimal set of points is only slightly better than a random set for this data set. This possibility is reflected by the fact that the passive learning MSE is not much worse than the DOE or uncertainty sampling methods.

The standard deviation of the methods across 10 iterations are as expected with DOE having the smallest standard deviation, and uncertainty sampling having second least standard deviation and passive learning having the most standard deviation. This is expected because each labeled sample can be replaced in DOE. Hence the labeled set should be most similar across all the iterations. I would expect the resulting training set to be very similar across all 10 simulations simulations. Passive learning should have the most standard deviation because it chooses samples randomly. Uncertainty sampling should have less standard deviation than passive learning because it chooses samples in a biased manner, but more standard deviation than DOE, because it can't replace the starting 10 labeled samples.

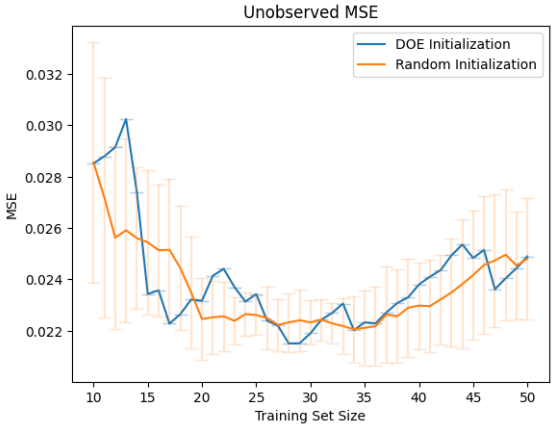

Figure 2: Unobserved MSE. DOE initialization performed with D-optimal metric.

(b)

The model used to generate the MSE scores was a linear regression model. In particular LinearRegssion from the python package SKlearn was used.

Uncertainty sampling was simulated with two forms of initializing the initial observed set. Each form of initialization was simulated 10 times. The first initialization method was DOE with D-optimality (maximizing $|X^T X|$ of the observed dataset), the second was randomly. The DOE method implementation is as described in 1a and was allowed to perform 50 replacements. The initial size of the observed set was set to 10 instances. Uncertainty sampling was run for 40 iterations for each simulation with each iteration moving one unobserved sample to the observed set.

The results somewhat match my expectations. The initial average MSE across the 10 simulations appears to be the same for both random initialization and DOE. This is unexpected because DOE is expected to produce an initial observed dataset that can train a more generalized learner (model). However, something of note is that the standard deviation of the DOE initialization is 0 or very close to 0. This makes sense because the initial observed points from DOE should be very similar each simulation. Thus an aggressive uncertainty sampling strategy will also choose the same points to add to the observed data set each simulation.

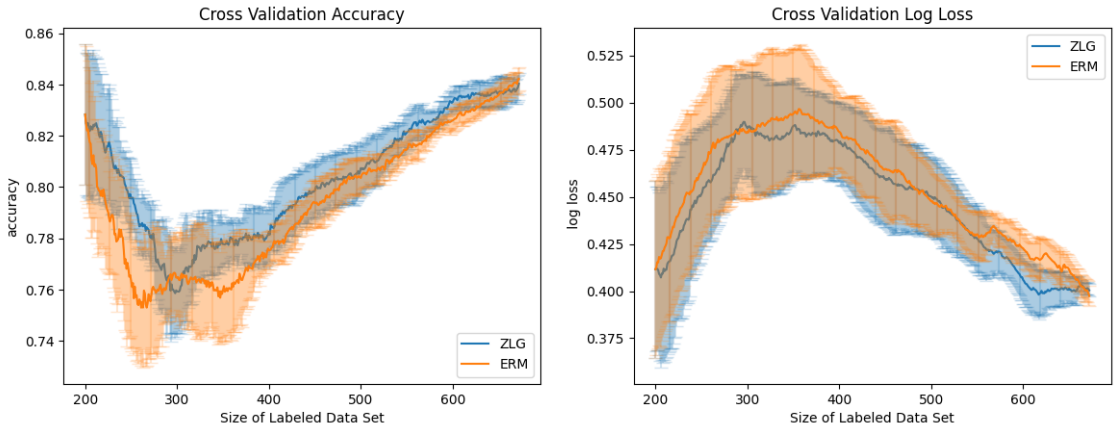

Figure 3: Cross validation accuracy and log loss. ERM (expected risk minimization)

2.

The dataset is a classification dataset with 2 classes (0 and 1). The data set has 8 features that are floats. The learner used for these simulations is a logistic regressor. Specifically the logist regressor implemented in SKlearn.

Two methods were used to perform active learning, ZLG and ERM. ZLG was implemented using the following pseudocode:

Require: L, U, W
  while More labeled data required **do**
    Compute Harmonic $f_u = -\Delta_{uu}^{-1}\Delta_{ul}f_l$
    Find best querry k $\underset{k}{\text{argmin}}\hat{R}(f^{+x_k})$
    Query point x_k and recieve label y_k
    Add (x_k, y_k) to L and remove x_k from U.
  end while

L is the labeled samples, U is the unlabeled samples, W is the weight matrix representing a graph where each node is a sample and $W_{i,j}$ is the weight of edge connecting sample i to sample j. The weight matrix is a nxn matrix where n is the total number of samples, unlabeled and labeled. The values in W are calculated using the formula

$$W_{ij} = \exp(-\frac{1}{\sigma^2}\sum_{d=1}^{m}(x_{id} - x_{jd})^2)$$

where m is the number of features for the samples (i.e. $x_i \in \mathcal{R}^m$). σ was chosen to be the standard deviation of the combined labeled and unlabeled features. The weight matrix is then filtered by thresholding W with some value t. A t value of 0.01 was selected for these simulations to generate a relatively well connected graph.

Δ is the laplacian matrix. The laplacian matrix is calculated using the formula:

$$\Delta = W - D$$

where D is a diagonal matrix with $D_{i,i} = \sum_{j=1}^{n} W_{i,j}$. Δ_{uu} is the portion of the laplacian matrix where the rows and the columns correspond to unlabeled samples. $Delta_{ul}$ corresponds to the portion of the laplacian matrix where the rows correspond to unlabeled samples and the columns correspond to labeled samples. $\hat{R}(f^{+x_k})$ is the risk function after adding unlabeled sample x_k to the labeled data set. Risk was calculated using the formula:

$$\hat{R}(f^{+x_k}) = (1 - f_k)\hat{R}(f^{+(x_k,0)}) + f_k\hat{R}(f^{+(x_k,1)})$$

$$\hat{R}(f^{+(x_k,y_k)}) = \sum_{i=1}^{p} min(f_i^{+(x_k,y_k)}, 1 - f_i^{+(x_k,y_k)})$$

where p is the number of unlabeled samples.

Both active learning strategies used 30% of the data as an initial labeled set. For each simulation, the active learning method was used to move a single sample from the unlabeled set to the labeled set until no unlabeled samples remained.

The results were as expected. Both ZLG and ERM had a dip in cross-validation accuracy and a increase in cross-validation loss at first. This is expected because both ZLG and ERM attempt to minimize risk and as a result will tend to add samples that generalize the model. Hence the cross-validation performance of the learner should decrease at first since the first samples ZLG and ERM add will tend to be samples that contain information the learner has yet to see.

3. (c)

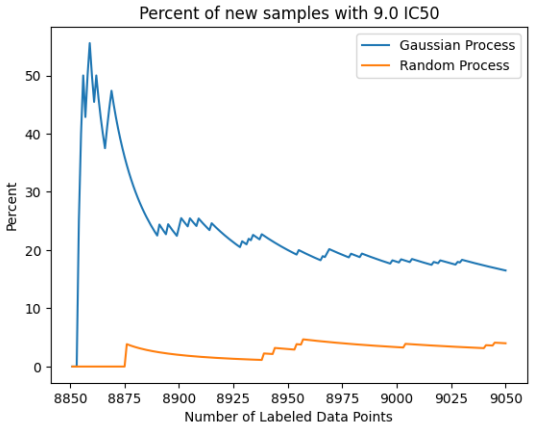

Figure 4: Percentage of samples added with binding affinity 9.0 pIC50

(d)

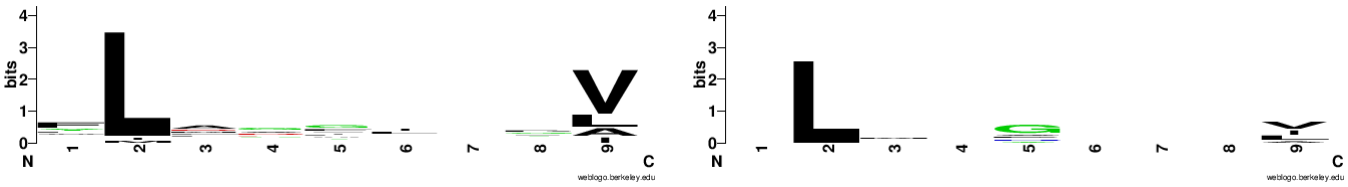

Figure 5: Left: The sequence logo of the samples with binding affinity 9.0 pIC50 retrieved by Gaussian processes. Right: The sequence logo of the samples with binding affinity 9.0 IC50 retrieved by random process.

(e) The dataset is a MHC class I binding affinity prediction dataset. There are 9052 samples in this dataset. The features used for the simulations is a one-hot encoded peptide sequence. Since the peptides are 9 amino acids long and there are 20 amino acids, the resulting feature vector is of length 180. The label is the binding affinity measured via pIC50.

80 percent of the samples were randomly selected to be placed in the initial labeled dataset.

For gaussian processes a radial basis function (RBF) was used as the kernel. A sparse gaussian process regressor implemented by the python package Gpy was used to model the labeled data. The model was optimized using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm.

Using the trained model, the mean and variance of the unlabeled samples were predicted. Then the mean and the variance were used to compute the upper confidence bound for each unlabeled sample. The sample with the maximum upper confidence bound was then chosen to be moved from the unlabeled data set to the labeled data set. The upper confidence bound was calculated using the following formula:

$$\text{UCB} = \mu + \beta_0 * \sigma^2$$
$$\beta_0 = 2 * \ln(\frac{\pi^2 n}{0.6})$$

where $n =$ the total number of labeled and unlabeled samples. μ is the mean of an unlabeled sample and σ^2 is the variance as predicted by the gaussian process.

The results were as expected. The gaussian process selects more samples that have 9.0 binding affinity than the random process. This makes sense because the gaussian process predicts the mean and variance of each unlabeled sample. Then it uses that info to choose the sample with the highest upper confidence bound. Assuming the gaussian process models the data well, this will tend to reflect the samples with higher pIC50 values.

This also results in a sequence logo that is more informative because we have selected more samples with 9.0 binding affinity. While both the gaussian process sequence logo and the random process logo have a large "L" amino acid for the 2nd position, the gaussian process sequence logo has a larger distinction for the amino acids in the 9'th position ($V > L > A$). For the random process "Y" appears in the 9'th position but it is less certain than the "V" in the gaussian processes 9'th position.