02-450 Homework 1

Katrina Liu
xiaol3@andrew.cmu.edu
February 20, 2022

1:

(a) I used the Gaussian Naive Bayes model from the sklearn package for my base learner and measured uncertainty with the least confident measure, where

$$x^* = \arg\max_{x \in \mathcal{U}}(1 - \max_{y \in \mathcal{C}} P(y|x)).$$

(b) I estimated the density/weighted-uncertainty with information density

$$x^* = \arg\max_{x \in \mathcal{U}} \phi_A(x)(\frac{1}{|\mathcal{U}|} \sum_{x' \in \mathcal{U}} sim(x, x'))^\beta$$

where uncertainty criteria ϕ_A is defined to be the least confidence measure of

$$\phi_A(x) = 1 - \max_{y \in \mathcal{C}} P(y|x),$$

similarity measure is the reciprocal of the exponent of euclidean distances between instances as

$$sim(x, x') = \frac{1}{e^{||x - x'||}},$$

and importance of weight as

$$\beta = 1.$$

(c) I choose to implement the expected error reduction model using expected 0-1 loss, where

$$x^* = \arg\min_{x \in \mathcal{U}} \sum_i P_\theta(y_i|x)(\sum_{x_u \in \mathcal{U}} 1 - \arg\max_{y \in \mathcal{C}} P_{\theta+(x,y_i)}(y|x_u).$$

(d)

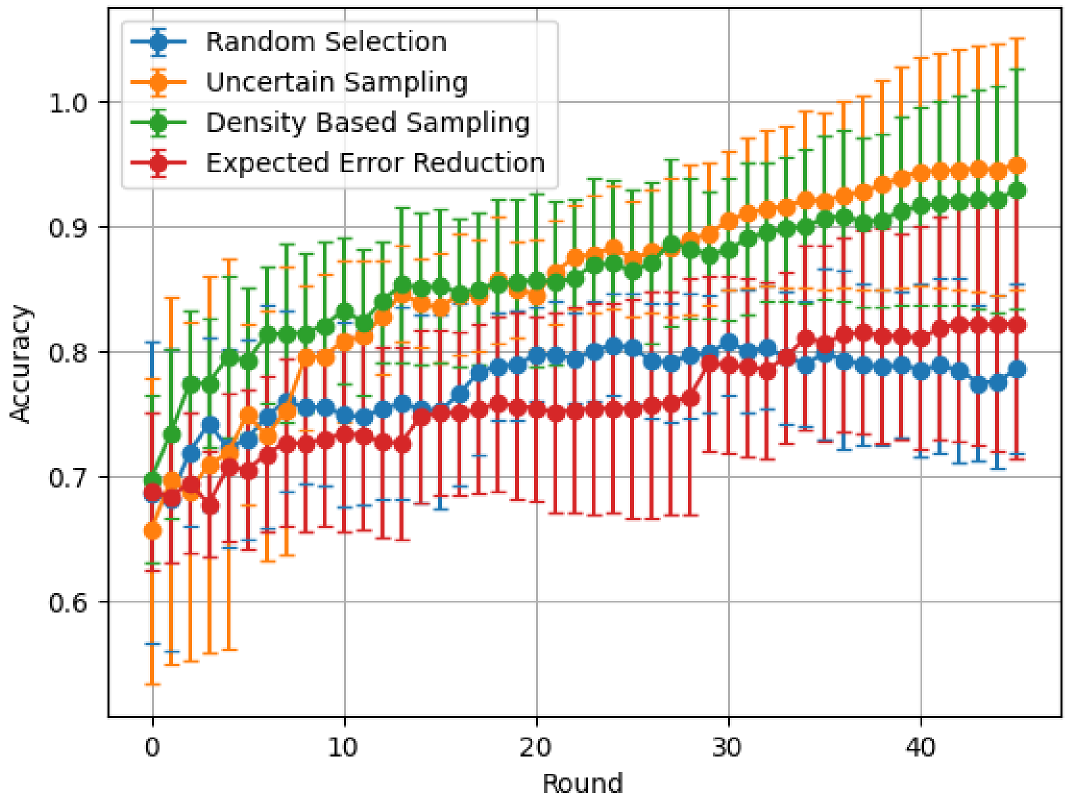

Figure 1: Prediction accuracy against rounds of training with base learner GaussianNB() of random selection, uncertain sampling, density based sampling, and expected error reduction.

2:

(a) I convert the uncertain sampling method into a mellow version of uncertain sampling, where instead of finding the choice with least confidence, I randomly pick one from the choices with a confidence in the lower half.

(b)

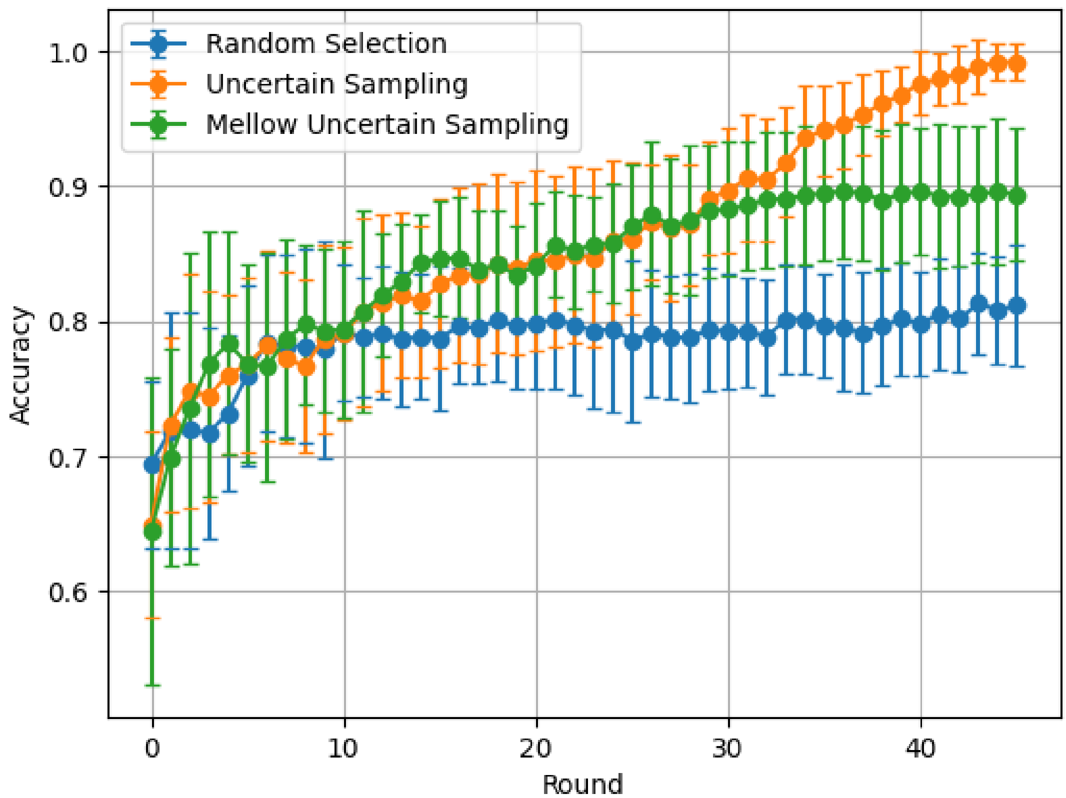

Figure 2: Prediction accuracy against rounds of training with base learner GaussianNB() of random selection, uncertain sampling, and mellow uncertain sampling.

3:

(a) The way I generate the data is to randomly generate a uniformly distributed x components in a range of [0,10], and assign labels by splitting them with $x[0] < x[1]$. In this way, the density should have a minimal impact as the data are evenly distributed. As Figure 3 shows, uncertain sampling in this way will produce a model with a slightly higher prediction accuracy in a more efficient way.

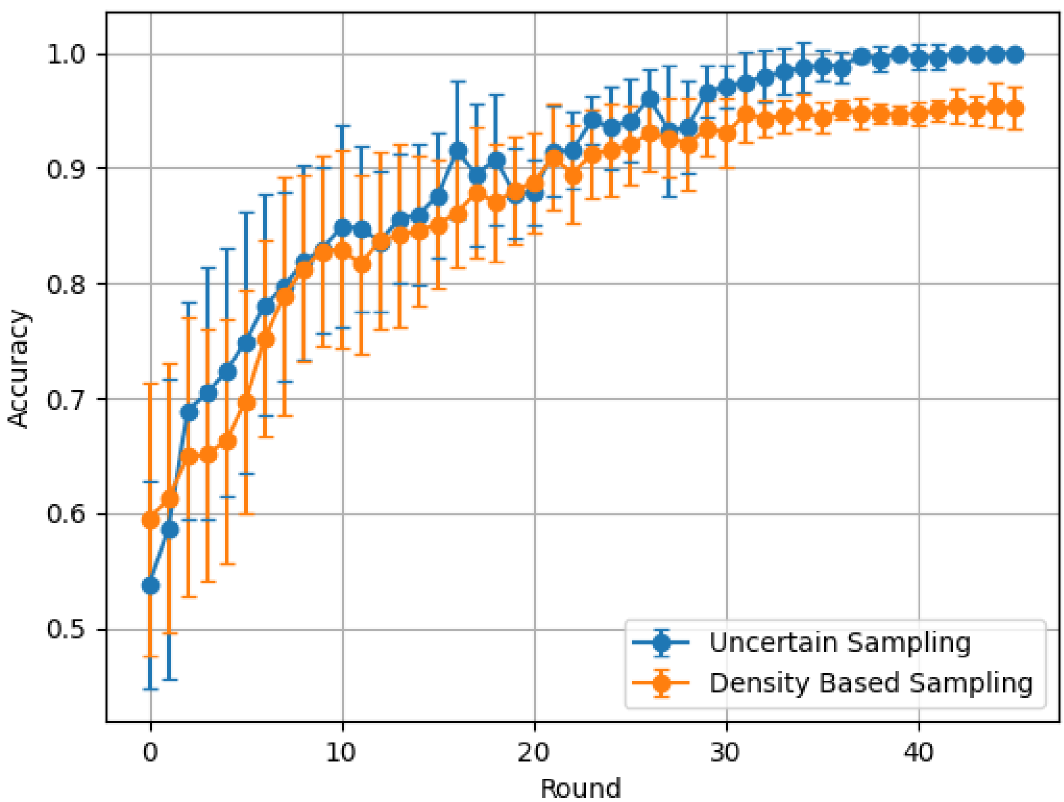

Figure 3: Learning curves of uncertain sampling and density based sampling with uniformly generated instances.

Katrina Liu
xiaol3@andrew.cmu.edu

(b) The way I generate the data that would prefer density based sampling is to first generate the x data with two clusters centered at (2.5,7.5) and (7.5,2.5), and then continue to assign the labels by splitting them with $x[0] < x[1]$. In this way, the density based sampling would prefer the points in the cluster and therefore increase accuracy of the model with better efficiency.

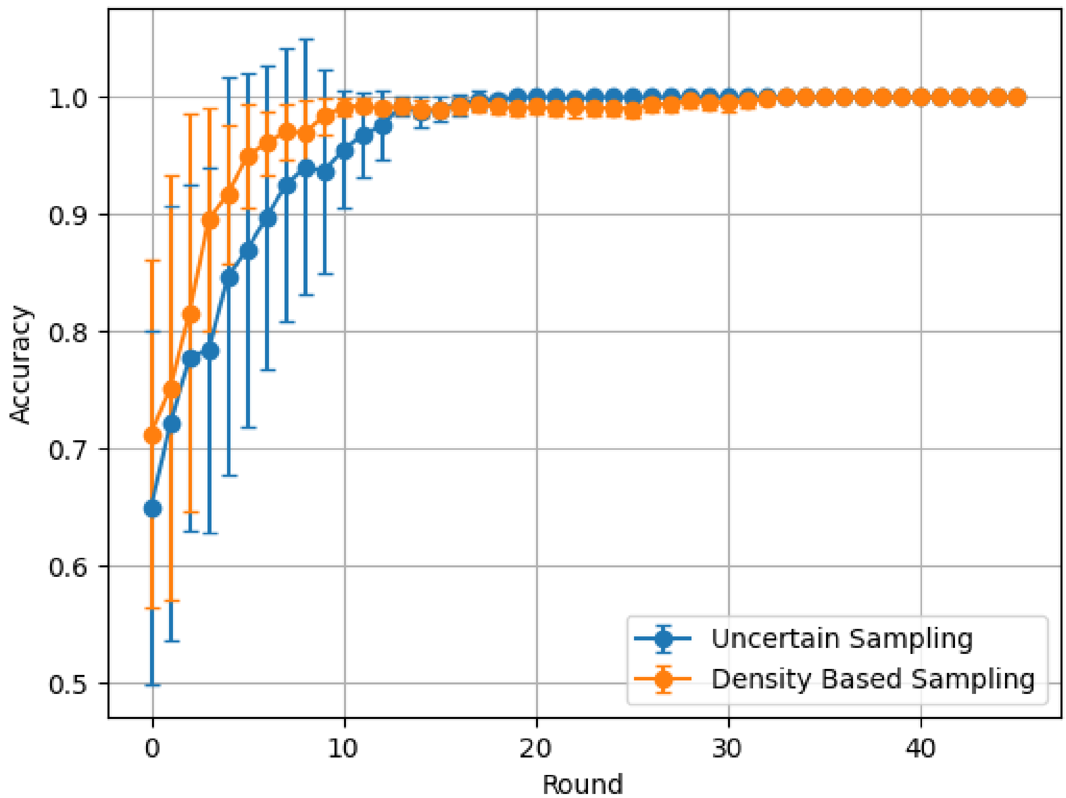

Figure 4: Learning curves of uncertain sampling and density based sampling with cluster generated instances.