Automation of Scientific Research
Comp Bio 02-450/02-750
Spring 2024

Homework Assignment #2

Assigned: Feb 13, 2024

Due: Thursday, Feb 22, 2024 by 11:59pm

Three exercises: 100 points in total (Exercise 1: 60 points; Exercise 2: 10 points; Exercise 3: 30 points)

Instructions:

Please submit this assignment in two parts: well-commented code and report. For the code, please submit a single package containing Jupyter notebooks and corresponding datasets. Your submission package should be compressed and named firstname_lastname_hw2.zip (e.g., jose_lugo-martinez_hw2.zip). In your package there should be everything necessary to successfully execute your code. For this homework, you should submit three Jupyter notebooks with prefix exercise-1, exercise-2, and exercise-3 corresponding to Exercise 1, Exercise 2, and Exercise 3, respectively. Each program should solve the assigned exercise. Make sure to add comments to each program, including your name. In the case of the report, you should submit a single PDF file named the firstname_lastname_hw1.pdf reporting **all answers, all figures along with description, and all relevant results and discussion.** This report must be **typed** and make sure that you type your name and CMU username on top of the first page of PDF file. **Finally, you may use whatever ML packages you find helpful; however, the implementation of the query selection algorithms should be your own.**

Academic Integrity: All assignments are individual, except when collaboration is explicitly allowed. All the sources used for problem solving must be acknowledged (e.g., web sites, books, research papers, personal communication with people). Academic integrity is taken seriously! For detailed information refer to the syllabus section on Academic Integrity.

Exercise 1 (60 points): Systematic Evaluation of Heuristic Query Selection Methods
Visit the UCI Machine Learning Repository and pick one dataset. **Make sure that you select either a classification dataset or a regression data set.** Depending on the learning task: classification or regression, choose an appropriate learning algorithm as the base learner for your simulation. Initially, all the training data should be hidden from the base learner. As the rounds progress, you will reveal the observations to the base learner. Your simulation should start with **twenty percent (20%)** **random** observations (no need to ensure that you have a balance of labels in your initial set).

a. (5 points) Describe the selected data set, base learner, and loss function.

b. (15 points) In each cycle of your active learning simulation, you should select **one (1)** observation using **query by committee** to add to your training set. **Make sure to concisely describe (i.e., no more than a few sentences), how you constructed and maintained your committee as well as quantified committee disagreement for your learner.** Continue your simulation until you have no observations remaining. Run the classification simulator **ten (10)** times with different seeds. Generate a plot showing the average and standard deviation of the 5-fold cross-validation performance (e.g., accuracy) as a function of the number of instances in the training data set.

c. (15 points) In each cycle of your active learning simulation, you should select **one (1)** observation using **expected model change or minimization of expected risk (PICK ONE)** to add to your training set. **Make sure to concisely describe (i.e., no more than a few sentences), how you the selected query selection method and provide corresponding details for that learner.** Continue your simulation until you have no observations remaining. Run the classification simulator **ten (10)** times with different seeds. In the same plot as part 1b, generate a plot showing the average and standard deviation of the 5-fold cross-validation performance (e.g., accuracy) as a function of the number of instances in the training data set.

d. (15 points) In each cycle of your active learning simulation, you should select **one (1)** observation using **density-based sampling** to add to your training set. **Make sure to concisely describe (i.e., no more than a few sentences), how you computed pairwise similarity between instances as well as the query selection strategy (i.e., function ϕ described in class).** Continue your simulation until you have no observations remaining. Run the classification simulator **ten (10)** times with different seeds. In the same plot as parts 1b and 1c, generate a plot showing the average and standard deviation of the 5-fold cross-validation performance (e.g., accuracy) as a function of the number of instances in the training data set.

e. (5 points) Run the corresponding code from Homework 1 (i.e., classification or regression) for both passive learning and active learning using uncertainty sampling **but continue your simulation until you have no observations remaining.** Run the classification simulator **ten (10)** times with different seeds. In the same plot as parts 1b, 1c, and 1d, generate a plot showing the average and standard deviation of the 5-fold cross-validation performance (e.g., accuracy) as a function of the number of instances in the training data set for passive learning and active learning.

f. (5 points) Compare the results between parts 1b-1e. Discuss whether the results matched your expectations and **explain your reasoning**.

Notes (PLEASE READ):

- To ensure we are only comparing selection method behavior, be sure that all simulations above are using the same base learner. This eliminates any potential bias from a change in the base learner.

- Your simulations should have the same random initialization sets. You can accomplish this by picking 10 random seed values and using those same seeds to initialize each of your simulations for different selection methods. This eliminates any potential bias in your results caused by different initialization sets. **To check this in your results, look for all your simulations to have the same average and standard deviation of accuracy in the first round.**

Exercise 2 (10 points): Aggressive vs Mellow Convert **ONE** of the query selection methods in Exercise 1b-d from an aggressive modality to a mellow modality. **Make sure to concisely describe your approach to convert it to a mellow modality.** Run the classification simulator **ten (10)** times with different seeds. Generate a plot showing the average and standard deviation of the 5-fold cross-validation performance (e.g., accuracy) as a function of the number of instances in the training data set. Compare it to the original version of the query selection method.

Exercise 3 (30 points): IWAL Algorithm Visit the UCI Machine Learning Repository and pick a binary classification dataset**.** You will use the selected dataset for implement the Importance Weighted Active Learning (IWAL) algorithm as described in Beygelzimer *et al.* (2009). Choose an appropriate learning algorithm as the base learner and loss function for your IWAL simulation.

a. (20 points) Initially, all the training data should be hidden from the base learner. As the rounds progress, you will reveal the observations. In each cycle of your simulation, you should select **one (1)** observation as described in Algorithm 1 IWAL (subroutine *rejection-sampling*). Implement the *bootstrapping rejection sampling* subroutine described in Section 7.2 of the paper, where parameter *b* is set to **twenty percent (20%)** of the training set size. You can set *k* to be the same as the committee size for Exercise 1b, but it is up to you. Run the IWAL simulator **five (5)** times with different seeds. Generate a plot showing the average and standard deviation of the 5-fold cross-validation **accuracy** on the training set as a function of the number of instances in the training data set.

b. (5 points) Run the corresponding code from Exercise 2 of Homework 1 for both passive learning and active learning using uncertainty sampling **but continue your simulation until you have no observations remaining.** Run the classification simulator **five (5)** times with different seeds. In the same plot as part 3a, generate a plot showing the average and standard deviation of the 5-fold cross-validation performance (e.g., accuracy) as a function of the number of instances in the training data set for passive learning and active learning.

c. (5 points) Compare the results between parts 3a and 3b. Discuss whether the results matched your expectations and **explain your reasoning**.

Recommended approach for 5-fold cross validation accuracy calculations for one round of one simulation: Split your training data at that round into 5 different sets of equal size (or as equal as they can be). Train a model using 4 of the 5 sets. Assess that model on the remaining set. Continue this process until each set has been used for assessment once. Add up the errors from all 5 folds and divide by the total number of observed instances. This will yield an average error that will not be severely biased by imbalances in the fold sizes. This will give you a good estimate of the model performance given the data you have available to you at that round. This will not necessarily give you an estimate of generalization performance.

Final remarks: Throughout the rest of the semester, we will modify various aspects of these simulations (input data, base predictive model, active learning selection method, stopping criteria, etc.). Therefore, please keep this in mind as you are developing your code for flexibility and modularity.