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

Homework Assignment #1

Assigned: Jan 30, 2024

Due: Thursday, Feb 08, 2024 by 11:59pm

Three exercises: 100 points in total (Exercise 1: 20 points; Exercise 2: 40 points; Exercise 3: 40 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_hw1.zip (e.g., jose_lugo-martinez_hw1.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 (20 points): Offline vs Online Learning 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 appropiate learning algorithm to learn a model for your simulation in the offline setting as well as the online setting.

a. (15 points) For the online setting, your simulation should start with **fifty percent (50%) random** observations (no need to ensure that you have a balance of labels in your initial set). Then, you should select **ten percent (10%) random** observations to add to your training set. Continue your simulation until you have no observations remaining. Finally, run the simulator **five (5)** times with different seeds. Finally, generate a plot showing the average and standard deviation of the

regret. Note: Make sure to use an appropriate loss function depending on the learning task.

b. (5 points) Describe the results as well as discuss whether the results matched your expectations and explain your reasoning.

Exercise 2 (40 points): Passive vs Active Learning (Classification) Visit the UCI Machine Learning Repository and pick a classification dataset**.** You will use the selected dataset for a classification-based simulation. 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. (15 points) In each cycle of your active learning simulation, you should select **one (1) random** observation to add to your training set. Continue your simulation until you have 50% observations and 50% unobserved remaining. Run the classification 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 and the average and standard deviation of the classification **accuracy** on the unobserved set as a function of the round number. Note that with random selections, these two measurements of accuracy are expected to converge toward the end of your simulations, however with different query selection strategies and different arrangements of the experimental space, you may see large differences in these measurements.

b. (20 points) In each cycle of your active learning simulation, you should select **one (1)** observation using **uncertainty sampling** to add to your training set. **Make sure to concisely describe (i.e., no more than a few sentences), how you estimated uncertainty for that learner.** Continue your simulation until you have 50% observations and 50% unobserved remaining. Run the classification simulator **five (5)** times with different seeds. In the same plot as part 2a, add plots showing the average and standard deviation of the 5-fold cross-validation accuracy on the training set and the average and standard deviation of the classification accuracy on the unobserved set as a function of the round number.

c. (5 points) Compare the results between part 2a and 2b. Discuss whether the results matched your expectations and explain your reasoning.

Exercise 3 (40 points): Passive vs Active Learning (Regression) Visit the UCI Machine Learning Repository and pick a regression dataset**.** You will use the selected dataset for a regression-based simulation. 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.

a. (15 points) In each cycle of your active learning simulation, you should select **one (1) random** observation to add to your training set. Continue your simulation until you have 50% observations and 50% unobserved remaining. Run the regression simulator **five (5)** times with different seeds. Generate a plot showing the average and standard deviation of the 5-fold cross-validation **error** on the training set and the average and standard deviation of the classification **error** on the unobserved set as a function of the round number.

b. (20 points) In each cycle of your active learning simulation, you should select **one (1)** observation using **uncertainty sampling** to add to your training set. **Make sure to concisely describe (i.e., no more than a few sentences), how you estimated uncertainty for that learner.** Continue your simulation until you have 50% observations and 50% unobserved remaining. Run the regression simulator **five (5)** times with different seeds. In the same plot as part 3a, add plots showing the average and standard deviation of the 5-fold cross-validation error on the training set and the average and standard deviation of the regression error on the unobserved set as a function of the round number.

c. (5 points) Compare the results between part 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.

Recommended approach for determining accuracy on the unobserved set for one round of one simulation: Using all your training data for that round, train a model. Assess that model on all the unobserved instances in your simulation at that round. This will give you an estimate of the performance on unobserved experiments.

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.