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

Homework Assignment #4

Assigned: Mar 28, 2024

Due: Tuesday, Apr 16, 2024 by 11:59pm

Two exercises: 70 points in total (Exercise 1: 40 points; Exercise 2: 30 points)

Instructions:

Please submit this assignment in two parts: well-commented code and document. For the code, please submit a single package containing Jupyter notebooks and corresponding datasets. Your submission package should be compressed and named firstname_lastname_hw4.zip (e.g., jose_lugo-martinez_hw4.zip). **In your package there should be everything necessary to successfully execute your code.** For this homework, you should submit **two** Jupyter notebooks with prefix exercise-1 and exercise-2 corresponding to Exercise 1 and Exercise 2, 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_hw4.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 (40 points): Batch Selection

a. (10 points) Starting from your code base from Homework 1, Exercise 2 and the selected classification task, implement a greedy batch-wise selection criteria using **uncertainty sampling** for your chosen base learner. 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 **five (5) random** observations. After initialization, you should select a batch of size **three (3)**, that is, you should select the three most uncertain instances for observation. Continue your simulation until 50% of instances have been observed. **In a single plot**, show the average performance (and standard deviation) of predictions on the unobserved set as a function of the number of instances

observed for the selected method against passive learning and uncertainty sampling (Exercise 1 of Homework 2) across 10 simulations each. Make sure to also include comparison against passive learning for a batch size of **three (3).** Comment on the differences in performance between each method.

b. (10 points) Starting from your code base from Homework 1, Exercise 2 and the selected classification task, implement a batch-wise diversity-based sampling method for your chosen base learner. 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 **five (5) random** observations. After initialization, you should select a batch of size **three (3)** instances using diversity sampling for observation. In no more than a few sentences, describe **how you selected the query selection method and provide corresponding details for that method**. Continue your simulation until 50% of instances have been observed. **In the same plot as part a**, show the average performance (and standard deviation) of predictions on the unobserved set as a function of the number of instances observed for the batch-wise diversity-based sampling method across 10 simulations each.

c. (15 points) Following the same guidelines as part b, implement one of the following batch-wise sampling methods: (i) Hoi *et al*. (ICML, 2006), (ii) Guo & Schuurmans (NeurIPS, 2007), (iii) Ravi & Larochelle (ICLR, 2018), or (iv) Bailey et al. (eLife, 2024). **In the same plot as part a**, show the average performance (and standard deviation) of predictions on the unobserved set as a function of the number of instances observed for the selected sampling method across 10 simulations each.

d. (5 points) Compare the results between part 1a, 1b and 1c. Discuss whether the results matched your expectations and explain your reasoning.

Exercise 2 (30 points): Proactive Learning Modify your classification code base to consider costs using the file "ex2_data.csv" as your dataset. Implement a cost-based selection method similar to Proactive Learning: **Scenario 3** (Lecture 15). Your selection method is allowed to see the costs while making selections so there is no need to try to estimate costs. **Note that you only need to implement the variable cost oracle.** The cost of an instance is determined based on what is in the cost column of that instance in the file provided. For comparisons, generate two additional selection methods: (i) a method that will choose the least costly instance for observation (flip a coin to break ties) and (ii) a method that will randomly select an instance. The random selection method will ignore cost but be assessed based on the costs it incurred through random selection. Run 10 simulations with a **max budget of 500** (~25% of full budget) for each of the three selection methods. All the simulations should start out with a random selection of **five (5)** instances. Terminate your simulations when they have used up the budget. **In a single plot**, show the resulting classification accuracy as a function of budget expenditures. **Note:** You should consider smoothing techniques. Comment on the differences between the performance of the selection strategies. Discuss whether the results matched your expectations and explain your reasoning.