Automation of Biological Research: 02-450/02-750

Carnegie Mellon University

Homework 5

Version: 1.0; updated 10/29/2015

Due: November 20 (Friday) by 11:59pm

Hand-in: Email your responses to: ABR-instructors@googlegroups.com

Notice: Please create an archive file (either tar or zip) of your code and report. The name of the archive should be your Andrew ID. Your report should be in PDF format and your name should appear at the top of the document.

Overview

This homework has 2 questions. Both questions require some programming. They are intended to test your understanding of the structure-based active learning algorithm "DH" covered in lecture 13, and your ability to implement an active learning algorithm for a regression problem.

Question 1: "DH" (60 points)

Scenario: Cancer is a complex disease involving the uncontrolled growth of genetically heterogeneous masses of cells called tumors. People informally refer to a given cancer by its location (e.g., 'lung cancer', 'brain cancer', 'pancreatic cancer', etc), but the reality is that there may be several cancer subtypes associated with each location which, in turn, may require different treatments. Therefore, it is important to determine the cancer's subtype prior to selecting a treatment. The most accurate way to do this is to perform a biopsy which is an expensive and invasive procedure. Over the past 10 to 15 years, significant research has been devoted to identifying cancer subtype 'biomarkers' from less expensive procedures, including measuring the quantities of various proteins in serum or urine.

In this question you will use the DH algorithm to label the data from 1,000 patients. Each sample is a 1 by 25 vector containing the concentrations of 25 proteins from one of two sources (serum or urine). There are two cancer subtypes (type 1 and type 2). The goal is to learn a model capable of distinguishing the two types from a protein panel. You will also determine whether it is better to use serum data or urine data when making a prediction. An oracle is available to get the true label for any instance by performing a biopsy.

Provided Files: The files *computeLoss.m, chooseBestPruningAndLabeling.m, getSerumData.m , and getUrineData.m* are provided to you as subroutines. You do not need to change these files, but you should look at them so that you understand what they take as input and what they return as output. The function ABRHW5_Q1 is provided to run your code for parts D, E, F, and G. It will run the necessary experiments and plot the results.

You will have to complete the implementations of the functions in the files *getLeaves.m, assignLabels.m, DH_SelectCase1.m* and *DH_ SelectCase2.m.* Please first read *DH_SelectCase1.m* to understand the relationships among these files.

Tasks:

A. (10 points) Complete the implementation in the file *getLeaves.m*. You should read the comments in the file *getLeaves.m* so that you understand the inputs and outputs. Also, you may need to study the output produced by the built-in *linkage* function in MATLAB: http://www.mathworks.com/help/stats/linkage.html#zmw57dd0e352239
B. (5 points) Complete the implementation in the file *assignLabels.m*. You should read the comments in the file *assignLabels.m* so that you understand the inputs and outputs.
C. (10 points) Refer to Algorithm 1 in *Hierarchical Sampling for Active Learning (Dasgupta & Hsu)* and complete the implementation in the file *DH_ SelectCase1.m*. Here you only need to select nodes from the pruning **proportional to the size of subtree rooted at each node** (see Case 1 in section named "*The select procedure"* in the DH paper). You should read the comments in the file*DH_SelectCase1.m* so that you understand the inputs and outputs. You should also call the function *chooseBestPruningAndLabeling.m* in order to compute the steps labeled "update admissible A and compute scores" and "find best pruning and labeling" in the pseudocode in the paper.
D. (10 points) Run the function **ABRHW5_Q1('d')**. It will produce a plot charting the loss as a function of the number of queries averaged over 5 separate runs of your DH code using the serum data. Include the plot as part of your report. Approximately how many queries are needed before the generalization error converges? Note that it may take a few minutes for this routine to run, depending on how fast your machine is.
E. (10 points) Run the function **ABRHW5_Q1('e')**. It will produce a plot charting the loss as a function of the number of queries averaged over 5 separate runs of your DH code using the urine data. Include the plot as part of your report. Describe any differences between the curves in parts D and E. What might account for these differences (hint, you should look at and/or plot the data returned by the functions getSerumData and getUrineData).
F. (10 points) Complete the implementation of DH in *DH_ SelectCase2.m.* The only difference between this and the version in part C is that Selection Case 2 uses a

confidence-adjusted selection probability (see Case 2 in section named "*The select procedure"* in the DH paper). Run the function **ABRHW5_Q1('f')**. It will produce a plot charting the loss as a function of the number of queries averaged over 5 separate runs of your *DH_ SelectCase2* code using the serum data. Compare the results with those from part D. Which selection strategy is more accurate?

G. (5 points) Run the function **ABRHW5_Q1('g')**. It will produce a plot charting the loss as a function of the number of queries averaged over 5 separate runs of your *DH_ SelectCase2* code using the urine data. Compare the results with those from part E. Which selection strategy is more accurate?

What to hand in: your code, your plots, and your explanation of the plots.

Question 2: Paired Sampling (40 points)

Scenario: Pancreatic cancer is the fourth most common form of Cancer in the United States. The disease has a very poor prognosis because it is usually diagnosed at a late stage. Most patients die within 12 months of diagnosis, but a small percentage survives for five or more years.

In this question you will use active learning to build a **regression model** capable of predicting the expected survival time, post-diagnosis. Like the previous question, the data consist of a proteomics panel. The labels are obtained by a surgical procedure which conclusively determines the cancer's stage which, in turn, is highly predictive of prognosis. Surgery is very risky because is sometim al (especially in older patients).

You will implement an active learning algorithm that **requests a pair of labels** (i.e., data from two patients) each round. The labels are the number of months that the patient is expected to survive. The design of the algorithm is up to you. However, your algorithm must have two batch selection modes. The first mode (the default mode) selects the top two points (i.e., as ranked by your chosen utility score). The second mode (the diverse mode) selects two points that have high utility, but *also* have very different predictive labels. Specifically, one of the points should have a predicted label that is less than or equal to the 33^{rd} percentile (out of all the predictions), and the other point have a predicted label that is greater than or equal to the 66^{th} percentile (out of all the predictions). The intuition behind the diverse selection mode is that it forces the algorithm to select points that provide information multiple areas of the domain.

Provided Files: The files *ABRHW5_Q2.m, RegressionModel.m,* and *getSerumDataRegression.m* are provided to you as subroutines. The function ABRHW5_Q2s is provided to run your code for part B. It will run the necessary experiments and plot the results.

Tasks:

A. (35 points) Implement your regression model in the file *RegressionModel*.m The function takes three arguments, the data, the labels, and a flag named 'mode' which switches between the two batch modes (1=default mode; 2 = diverse mode). Your algorithm has a budget of 200 calls (i.e., 100 pairs) to the oracle. Select the first 20 patients (i.e., 10 pairs) at random. Your algorithm should select the remaining 90 pairs using the appropriate batch selection strategy. Your program should return a 1 by 90 vector returning the generalization error (defined as the mean absolute error over all 1000 patients).
B. (5 points) Run the function **ABRHW5_Q2**. It will produce a plot charting the loss as a function of the number of rounds averaged over 5 separate runs. Note that it may take a few minutes for this routine to run, depending on how fast your machine is.

What to hand in: your code and your plot from part (B).