BE700 Spring 2019

Homework 2

Please read the required reading on microarrays (ONLY PAGES 1-10 ARE REQUIRED FOR HW2. The rest will be covered later in the class)
It will also be posted on our reading material.
https://tinyurl.com/y9w47kru
AI magazine. 2004 Mar 15; 25 (1). doi: 10.1609/aimag.v25i1.1745
Using Machine Learning to Design and Interpret Gene-Expression Microarrays.
Molla M, Waddell M, Page D, Shavlik J.

This paper and data can be found in the Readings and Homework section on Blackboard, respectively.

Names of all coauthors should appear on the HW. Each section should clarify

a) Both students did the work and the results agree.
b) One of the students did a specific part and the other did other parts (please attribute each part accurately).
c) Each written paragraph or section should have an author (both students wrote this section WILL NOT BE ACCEPTED). If one person wrote and the other edited please specify and attach in supplement the text before editing.

Part 1: Due Thursday Feb 7th BEFORE 9 am

1. Download the data supplied.

   a. Run Weka K-NN nearest neighbor algorithms on this data. The K-NN algorithm is found in lazy ML algorithms under IBK. Report your relative accuracies (overall error) on this data using 1,3,5,9 nearest neighbor algorithms. Briefly discuss your findings (are there differences in accuracy and more). THIS IS EASY!

2. Identifying prognostic biomarkers in matlab, python or R.

   a. For each of the 30 probes, perform a T-test comparing the distribution of poor prognosis and good prognosis patients.
   b. Classify these prognostic genes into groups of up/down in THE POOR CLASS.
   c. Select the top 5 probes in terms of their p-values and report the p-values of all probes in a table. Please attach a supplement with this information in XLS format.
   d. Plot the distribution (HISTOGRAM) of gene expression values for good and poor prognosis for each of the top 5 probes. Attach the figures with SHORT legends to your write-up.
   e. Map these top 5 probes to genes using the supplied file and USE their GENE names.

Part 2: Due Thursday Feb 14th BEFORE 9 am.

   f. This question is due as part of part 2. Write a short description of 2-4 selected genes describing their possible relationship to cancer in general or lung cancer specifically. In particular:
      a. Are those genes documented in the literature as oncogenes or tumor suppressors?
      b. Why might they be associated with poor prognosis?
      c. What hallmarks are they potentially related to?
   g. This question is also due as part of part 2. Write a short discussion of your results and experience. In particular:
      a. Please contrast/compare the different accuracies obtained in different parts of your study. Speculate what might explain the difference in accuracy.

b. Discuss the specific accuracies in the context of the confusion matrix (i.e precision, recall, etc).

c. Using your top 5 and top 3 probes from question 2 part c, perform problem 1a again and discuss your findings as compared to 1a.

d. EXTRA CREDIT: UNDER SECTION HEADING NEW IDEAS, propose ideas to improve the accuracies of your K-NN classifiers with any creative “tricks” or techniques.

Email your write-ups to maurerj@bu.edu with the subject line “BE700 2019 HW 2”.