Homework 4 on Gene Function Prediction

CS 3824

T. M. Murali

Assigned on October 28, 2011
Due on November 10, 2011

1. (20 points) In class, we proved that the Hopfield network algorithm converges in a finite number of steps when every edge in the graph has a weight of 1. The proof worked by asking how the energy changed when the algorithm changed the state of a node v by applying the update rule. We showed that the only edges that contributed to the change in energy were incident on v. We used this fact to show that the change in energy must be negative. Since all edge weights are 1, the energy must go down by a value of at least 1. We established a lower bound on the energy to conclude that the number of steps is finite.

   Generalise the proof when every edge in the graph has a positive edge weight. You may make some reasonable assumptions about the precision with which we represent real numbers. (Hint: Consider the case when every edge weight is a positive integer.)

2. (10 points) Consider the precision-recall curve for an algorithm such as SinkSource. During cross validation, SinkSource computes a score between 0 and 1 for every positive and every negative example. To compute the precision-recall curve, we vary a threshold t monotonically from 1 to 0. At every value of t, we compute the precision and recall. Do either or both of precision and recall change monotonically with t? Explain your answer.

3. (70 points) This question involves some programming. You are given functional annotation data for an organism, i.e., for every GO term, you know which genes in that organism are annotated by that GO term. You have to create a plot of the number of genes for which there are different types of evidence codes, much like the plots you saw for Arabidopsis in the class. Here are the specific steps:

   (i) Pair yourself up with another student. If there are an odd number of students in the class, one group may have three students. *You must do your work individually.* The purpose of forming a group is solely for comparing the results you get with the other member(s) in your group, thereby reducing the number of bugs in your code.

   (ii) The organisms of interest are *Saccharomyces cerevisiae* (baker's yeast), *Homo sapiens* (human), *Arabidopsis thaliana* (the model plant), *Drosophila melanogaster* (the fruitfly), and *Caenorhabditis elegans* (a nematode/worm). Pick one organism for your group. Coordinate with the other groups so that no two groups pick the same organism. In this manner, we will cover as many of these important organisms as possible.

   (iii) Visit `http://bioinformatics.cs.vt.edu/~murali/teaching/2011-fall-cs3824/` and download the file corresponding to the organism you have selected:

   ***Saccharomyces cerevisiae*:** gene_association_true_path.sgd

***Homo sapiens*:** gene_association_true_path.goa_human

***Arabidopsis thaliana*:** gene_association_true_path.tair

***Drosophila melanogaster*:** gene_association_true_path.fb

***Caenorhabditis elegans*:** gene_association_true_path.wb

(iv) Each line of this file is tab-delimited. Each line contains information on a gene-GO term pair. The columns mean the following:

orf: an identifier for the gene.

goid: an identifier for the GO term.

goname: the name of the GO term.

hierarchy: the GO category that the term belongs to. The values in this column are 'c' (for cellular component), 'f' (for molecular function), and 'p' (for biological process).

evidencecode: the evidence code for the annotation. Note that the same gene-GO term pair may appear in different lines, with different evidence codes. If the evidence code is 'ND', you can ignore this gene-GO term pair.

annotation type: the usual value in this column should be 1. If the value is 0, it indicates that the status of the annotation of that gene with respect to that GO term is unknown. If the value is -1, it indicates that there is evidence suggesting that the gene should not be annotated to that term.

(v) Group evidence codes as follows (look at `http://geneontology.org/GO.evidence.tree.shtml` for guidance and ignore the code ND):

(i) group EXP: evidence codes IMP, IGI, IPI, IDA, and IEP.

(ii) group COMP: evidence codes ISS, IGC, and ICA.

(iii) group AUTH: evidence codes TAS and NAS.

(iv) group IC: evidence code IC.

(v) group IEA: evidence code IEA.

(vi) Write a programme to parse the file. Your programme should produce separate plots for three categories in GO: biological process, molecular function, and cellular component.

(vii) The plots should report the fraction and number of genes that have:

(a) at least one annotation in group EXP,

(b) at least one annotation in group COMP but have no annotation in group EXP,

(c) at least one annotation in group AUTH but have no annotation in group COMP or in group EXP,

(d) at least one annotation in group IC but have no annotation in groups COMP, EXP, or AUTH,

(e) at least one annotation in group IEA but have no annotation in groups COMP, EXP, AUTH, or IC.

(viii) To compute these fractions, you will need to compute the number of genes that have at least one annotation (with any evidence code other than 'ND'). Of course, you need to compute this number thrice, once for each GO category.

(ix) Repeat these calculations and plots after restricting your analysis only to the GO terms that annotate at most $k\%$ of the total number of genes with at least one annotation, where $k = 1, 5$, and 10.