In this mini-project, you will implement a decision-tree algorithm and apply it to molecular biology, a Promoter is a region of DNA that facilitates the transcription of a particular gene. Promoter compilations and analyses have led to computer programs which predict the location of promoter sequences on the basis of homology either to the consensus sequence or to a reference list of promoters. Such programs are of practical significance in searching new sequences.

The task for you is to develop a decision tree algorithm, learn from data, and predict for unseen DNA sequences whether they are promoters or non-promoters.

Our data was provided by UCI Machine Learning Repository and can be found in Molecular Biology (Promoter Gene Sequences) Data Set. It has 106 instances and 57 features. We randomly split the data set into the training (71 instances) and validation (35 instances) sets, which are both well balanced.

1. Download the data set which contains the training data and validation sets. Each DNA sequence is represented by one line, with the first 57 characters (one of 'a', 'g', 'c' and 't') representing the sequence, and the last character (separated by a space from the sequence) indicating the class ('+' for promoter, '-' for non-promoter).
2. Implement the ID3 decision tree learner, as described in Chapter 3 of Mitchell or in your Web link http://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm. You may program in any language that you prefer. Your program should assume input in the above format. For initial debugging, it is recommended that you construct a very simple data set (e.g., based on a boolean formula) and test your program on it.
3. Implement both accuracy (misclassification impurity) and information gain (entropy impurity) for evaluation criterion. Also, implement split stopping using chi-square test. The test-statistic is:

   $$\sum_{i=1}^{v} \frac{(p_i - p'_i)^2}{p'_i} + \frac{(n_i - n'_i)^2}{n'_i}$$

   where p*i* is the number of promoters in the i-th child, p'i is the expected number of promoters in the i-th child, ni is the number of non-promoters in the i-th child and n'i is the expected number of non-promoters in the i-th child. The degree of freedom is v-1
4. Use your algorithm to train a decision tree classifier and report accuracy on validation. Compare accuracies by varying the evaluation criteria and confidence level in determining split stopping. For the latter, use 99%, 95% and 0% (i.e., you always grow the full tree).
5. **Turn in the following:**
   - Your code. Submit your file through UNM Learn. **The due date is 12PM, Feb 13, 2014 (contingent to the late policy stated in the syllabus)**. Your code should contain appropriate comments to facilitate understanding. If appropriate, your code must contain a Makefile or an executable script that receives the paths to the training and testing files
   - A report of at most 4 pages (letter size, 1 inch margins, 12pt font) that describes:
     - A high-level description on how your code works.
     - The accuracies you obtain under various settings.
     - Explain which options work well and why.
     - If all your accuracies are low, tell us what you have tried to improve the accuracies and what you suspect is failing.
6. **Rubric:**
   - If the TA can execute your program 20 pts
   - If the results make sense 10 pts
   - If your code is thoroughly commented 5 pts
   - Your report meets the criteria described above 20 pts
   - Your report is clear, concise and well organized 15 pts
   - TOTAL: 70 pts (7 pts of your final grade)