Project 2 Data Analysis: Finding Explanations with Decision Trees

Due: April 17, 2013

Resources
- data files: /home/scratch/kemper/DataForCS626
- data table as a result of HW3
- R
- KNIME: http://www.knime.org
- Bertholt et al, Ch 8.1 and Ch 8.5

Overall Challenge: Develop a classifier for Saha’s data set that accurately determines if a gene is expressed and which gene that is.

Given the amount of work, it is strongly recommended to develop a script in R (or a model in KNIME) to perform the calculation of a decision tree and the

Task 1: One Decision Tree derived from full information for all classes

a) Construct a decision tree for the following scenario:
- Attributes: all attributes that describe the distribution and sequence of values (as obtained in Homework 3).
- Class: (gene,stage,duration of experiment) which is \{a1A,a1B,a1C,a1H,GAD,NBT,VGlut\} x \{14,18,22\} x \{2,12\} such that the class has 7 x 3 x 2 = 42 values.

Store and provide the description of the decision tree in a format that is readable by R or KNIME.

b) Describe the decision tree. How complicated is it (#nodes, depth, width)? Which attributes are used?

c) How well does the tree work? What is the misclassification rate for the given training data? What is the outcome of a k=10 fold cross-validation (mean,variance of error)?

d) Your conclusion: Does it make sense to use a single decision tree to do the overall job of classifying the data?

Task 2: Individual decision trees from full information for individual classes

a) Construct one decision tree for each of the 42 classes (gene,stage,duration of experiment) that tells if the given record belongs to that class, yes or no:
- Attributes: all attributes that describe the distribution and sequence of values (as obtained in Homework 3).
- Class: binary, 1 = true, 0 = false.

Store and provide the 42 different decision trees in a format that is readable by R or KNIME.

b) Produce a table that gives an overview over all 42 trees:
- columns to describe the class: gene, stage, duration of experiment,
- columns to describe the tree: #nodes, depth, width
- columns to describe the quality of classifier: misclassification rate, mean and variance of error from a k=10 fold cross validation
- columns to describe selected attributes: for each attribute provide a binary (1=yes, 0=no) value if that particular attribute is used in the tree

c) Interpret the results:
- Which genes, in which stage can be recognized well?
- Which attributes are particularly useful to recognize a gene?
- Are the same set of attributes considered for one gene across development stages?

Task 3: Experiment with particular attributes and derived attributes to improve results of Task 1 or Task 2. Try to find the best possible decision.

a) Select subsets of attributes to give more guidance to the greedy algorithm on what to pick. Pick attributes that describe the distribution and the sequence of values.
   a) Focus on attributes based on moments: mean, variance, skewness, kurtosis
   b) Focus on attributes based on quantiles: median, interquartile range, range (max-min)
   c) Focus on attributes based on correlation: autocorrelation lag 1-k
   d) Focus on attributes based on batches:
      a) slope of a linear regression through the batch means
      b) slope of a linear regression through the batch variances

b) Can you prune the decision trees improve the results?

You need not look at all trees of Task 1 and Task 2. Pick 2-3 trees for Task 3 that seem most promising / challenging to you.

Prepare yourself to present your results in class on April 19!
You need to present and discuss your findings. The plan is to invite Professor Saha to this presentation.

OPTIONAL BONUS TASK: Explore Pearson’s correlation across the original time series data. Can you see any significant correlation between cells that express a particular gene?