State Assignment Using Decision Trees

June 24, 2019

Given a vector x of gene expressions, state assignment determines which of several states s in S is the best label for x. This is a classification problem.

In state assignment, we are given $[X^T S]$, gene expression data paired with its associated state. X has columns indexed by K gene (features) and N rows indexed by data instance. Values in X are discrete. In our case, values are $[-1, 0, 1]$.

We use the notataion $\sigma(x) = s$ to denote that s is the state assigned to the vector x in X^T. The state assigned by algorithm A is $\sigma_A(x)$. Our objective in state assignment is to minimize $Pr(\sigma_A(x) \neq \sigma(x))$ that probability that the state assigned by A differs from the true state. We define $err(A, X) = \sum_{x \in X} 1_{\sigma_A(x) \neq \sigma(x)}$.

We use ν_k to denote t

Technical Approach

1. Data preparation
   (a) Eliminate T_0 since all 0's.
   (b) Combine Normoxia with Resuscitation since insufficient values for Normoxia.
   (c) Combine perfectly correlated features
2. Analysis 1: All data
   (a) Construct DT with all data
   (b) Analyze the effective number of independent features given feature "correlations" (can we use correlation or need a new measure because of nominal values).
   (c) Assess probability of uniquely identifying states under the null hypothesis of random and independent assignment of expression values
3. Analysis 2: Cross validation
   (a) Compare misclassification rates with random for 5 classes

4. Analysis 3: Sensitivity to random perturbations of the data
    (a) Train on data with an error fraction f. Do repeatedly to see stability of features. Relate to correlation blocks.
    (b) Evaluate on new data with an error fraction f

Visualizations

Vectorized Implementation