State Assignment Using Decision Trees

June 28, 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

Data preparation

1. Transform data to trinary values
   (a) Normalize for size of RNA library and gene length
   (b) log_2 values
   (c) Convert x, the values normalized as above, to trinary values. $v \in \{-1, 0, 1\}$. $v = -1$ if $log_2 x \leq -1$, $v = 1$ if $log_2 x \geq 1$. Otherwise, $v = 0$.
2. Eliminate T_0 since all 0's.
3. Combine Normoxia with Resuscitation since insufficient values for Normoxia.
4. Combine perfectly correlated features

Trinary Values

1. The trinary encoding is ordinal w.r.t. the underlying values.

2. Note that intuitions about variances are preserved. For example, the sample variance of $v_{0,1} = (0, 1, \cdots, 1)$ is the same as the sample variance of $v_{1,0} = (1, 0, \cdots, 0)$. That is, for vectors of length n, $E(v_{0,1}) = \frac{n-1}{n}$ and so $Var(v_{0,1}) = \left(\frac{n-1}{n}\right)^2 + (n-1)\left(1 - \frac{n-1}{n}\right)^2 = \left(1 - \frac{1}{n}\right)^2 + (n-1)\left(\frac{1}{n}\right)^2 = Var(v_{1,0})$. A similar argument applies to $v_{-1,0} = (-1, 0, \cdots, 0)$ and $v_{0,-1}$.

Single Tree Analysis

1. 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
2. Analysis 2: Cross validation
   (a) Compare misclassification rates with random for 5 classes
3. 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

Random Forests

1. Create random forest with all features
2. Construct the feature evaluation vector $E = \{e_{ij}\}$ for features j and tree i such that $e_{ij} = score_i$ if feature j is in tree i.
3. Cluster the features creating $C_1, \cdots, C_M$.
4. Find the clusters with the largest $\frac{1}{|C_m|} \sum_{j \in C_m} |e_{*j}|$
5. For each
   (a) Construct a decision tree with just its features
   (b) Calculate the score
6. Choose features in clusters with trees with the largest scores.
7. Do random forest on the subset selected and which features co-occur in trees and which do not. Those that do not are likely equivalent. Those that do co-occur are in some ways complementary.

Visualizations

Vectorized Implementation