Predictive modeling

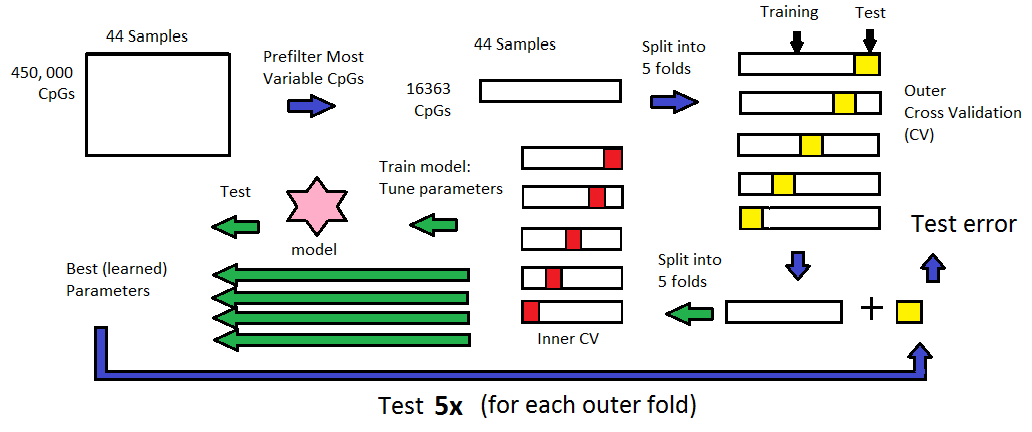
Support Vector Machines (SVM) is a powerful supervised learning method used to recognize patterns in data. SVM has been used in many bioinformatics applications for classification, regression, and outlier detection because it avoids overfitting, it can account for nonlinear relationships, and is robust to noise [1]. SVM is a statistical model based where the high dimensional feature space contains the input data that are mapped by a ‘kernel function’ (nonlinear).

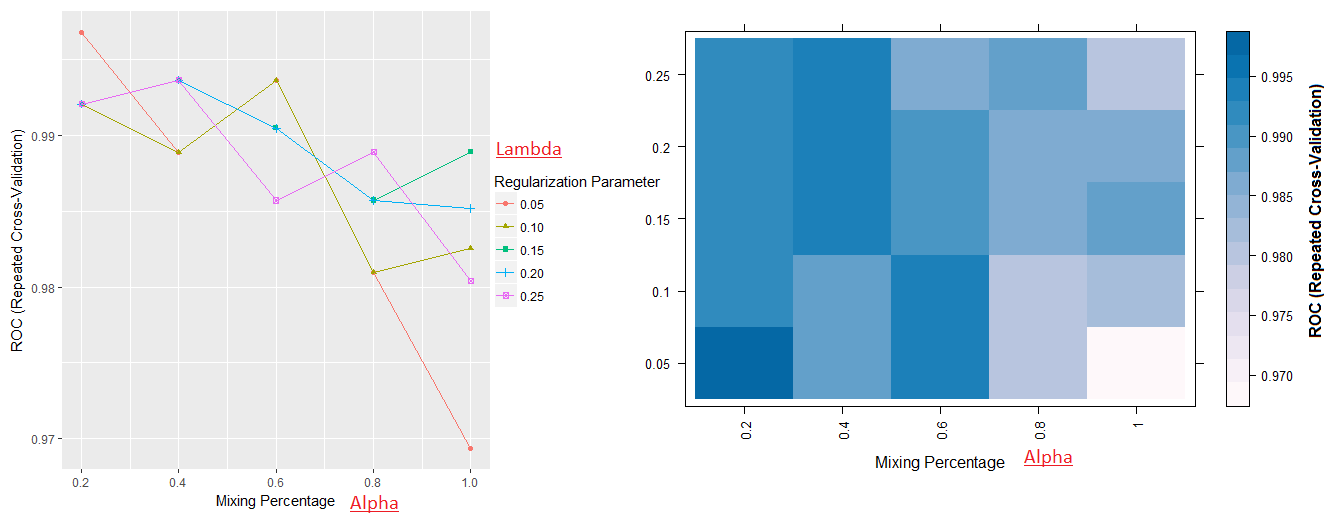
‘Elastic net’ is a regularized regression method that combines the L1 and L2 penalties from LASSO and ridge regularization methods. In this way, Elastic net overcomes some of the limitations of LASSO and ridge. Elastic net regularized regression (GLMnet) has been used to create a multi-tissue age predictor based on DNA methylation before [2], indicating that this might be well-suited for our dataset.

We propose to compare SVM and Elastic net regression methods in generating a classifier that will predict ethnicity based on DNA methylation features (CpGs).

To facilitate computation efficiency, only CpGs with variability were chosen to train the models on. We chose to use an arbitrary threshold to determine which CpGs to retain. CpGs with a standard deviation (SD) greater than 0.10, leaving 16363 CpGs for model building. We reason that only variable CpGs are likely to be able to be used to distinguish ethnicity, however we are uncertain with the viability of this strategy in setting a threshold.

In order to estimate the optimal parameters of these models (Elastic net: α, λ; SVM: penalty factor C) and to estimate prediction accuracy, data was randomly divided into 5 sets for cross validation (CV), using a testing and training set. The samples in the training set was used to train the model with a grid of tuning parameters. Five values of α, and λ (25 different combinations) were used for Elastic net, while 5 values of C for SVM were tested. The parameters resulting in the model with the highest ROC on the testing sets were chosen. An outer layer of cross validation (nested) was used to estimate test error (figure 1).



We used the Caret package [3] to train the models and apply cross validation because of its many available useful functions. For example, using the trainControl() function allows the specification a range of values for the model to be built on (figure 2). The default function is to select the parameter values that results in the highest performance (ROC, accuracy, Kappa.. etc.). However, we chose to maximize alpha (L1 norm) to obtain a smaller panel of ethnicity markers. The final tuning parameters were α = 0.75, λ = 0.077 for our GLMnet -generated model.

[1] - Vapnik VN. Statistical Learning Theory. Wiley, New York (1998)

[2] - Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14:R115.

[3] - Kuhn, M. (2008). Caret package. Journal of Statistical Software, 28(5)