ECE1508, Machine Learning and Genomic Medicine, Fall 2015

Assignment 1
Due October 22, 2015

Download the ClinVar database, which contains variants identified in clinical labs along with labels of pathogenic, likely pathogenic, benign, likely benign and unknown significance. Also, read the paper describing the ACMG variant classification guidelines, which is posted on the course web site.

1. Download the minor allele frequencies for the ClinVar variants, either from the ClinVar website (if its there) or from the UCSC genome browser web site. For each ClinVar category, plot the distribution of allele frequencies. Comment on how this relates to the category.

2. Download the conservation track from the UCSC genome browser web site and for each of the 5 ClinVar variant sets, plot the distribution of conservation scores. Comment on differences between these distributions and how they relate to the category labels.

3. Set aside the variants of unknown significance and discard the variants labeled likely benign or likely pathogenic. Randomly divide the remaining variants into training and test sets, doing so by randomly picking genes. It is important that training and test cases not come from the same gene – explain why. Train a logistic regression classifier that has a single input, conservation, and outputs a distribution over the two categories. Apply it to the training set and the test set. Then plot the training and test ROC curves and compare them. Is there overfitting? Compute the area under the ROC curve and also the TPR at an FPR of 0.1% and 1%. Repeat the random procedure repeatedly to get the average values and confidence intervals for these three metrics; report the results and comment.

4. Repeat Q3, but use both conservation and the minor allele frequency as inputs to the classifier. Compare the performance to that obtained using just the conservation as input.

5. Download other kinds of information from the UCSC genome browser and see if you can improve the classifier.

6. Access variant scores from the ANNOVAR database and try using these scores as additional inputs to your classifier. Compare with the other methods that you already trained. Note that if the ANNOVAR score has a missing value for a variant, you'll need to carefully choose what value to use to represent a missing value. To do this, plot the distributions of the ANNOVAR scores for the benign and pathogenic variants (the ones that don't have missing values of course), and this should give you an idea for what value a missing score should be set to.

7. Try out other things, such as multilayer nonlinear neural networks, random forests, and other kinds of data that can be used to classify the variants.