Relaxed LASSO in GO space

Luis Pedro Coelho

Instituto de Medicina Molecular

**Summary Sentence**: Relaxed LASSO after grouping gene expression values by GO term.

**Background/Introduction**

This problem can be posed as a simple regression problem. However, the number of dimensions greatly exceeds the number of examples. Therefore, we focused on reducing the number of dimensions in order to then be able to apply simple regularized regression to the problem.

For simplicity, we worked only with the gene expression data (microarray and RNA-seq). The fact that two different data sources were available for the same underlying phenomenon also led us to prefer this data. We grouped genes by their GO terms in order to obtain a smaller number of dimensions.

After projecting into GO terms, we used relaxed LASSO. Standard L₁ penalization achieves two goals: (1) sparsity in that many (even most) coefficients are exactly zero; and (2) regularization in that non-zero values are smaller (in absolute value) than in unregularized regression. Due to the very large number of variables, a very large penalization was needed for sparsity, which led to over-regularlization. Relaxed LASSO, by separating, the effects of sparsity and regularization achieves a better balance between these two.

**Methods**

Only gene expression data was used, with RNA-seq & microarray data being combined to form a average value.

1. Use the seq calls provided from the RNA-seq data to determine which genes are changing. Only these “active” genes will be used.
2. Preprocess the RNA-seq data with a log-transform,where is the new value and is the old one, followed by normalization to z-scores.
3. Combine the RNA-seq and the microarray data into a single prediction by averaging the z-scored values.
4. Any gene without both an RNA-seq and microarray measurement was discarded.

Now, we obtain a matrix of gene expression which contains only the active genes. We do not use this matrix directly. Instead we look up GO terms for all genes (ignoring the Cellular Component vocabulary). All genes which map to the same GO term are combined. We combine the genes by reducing with the following function

maxabs(a, b) = if |a| > |b| then a else b

That is, we keep the maximum in absolute terms, while preserving its sign. For learning, we further process the data by

* + 1. re-normalising to z-scores (the maxabs transform does not preserve this property).
    2. selecting features which correlate with at least 20 drug signatures.

Finally, we use relaxed lasso for the optimisation. A first lasso pass with λ=2⁻¹²·¹² is used for feature selection, and a second pass with λ' = λ/10 is used for the final learning. An initial attempt to use cross-validation to learn λ led to massive over-fitting and a value in the middle of the range was chosen.

The optimisation uses coordinate descent for optimisation and ignores the regression error in the missing entries.

Where the weight of example ,, , is zero if the entry is missing.

**Conclusion/Discussion**

The main driving force behind these ideas were the pressure for feature selection and dimensionality reduction.

One major problem with this approach is the need for setting parameters (the penalization factors λ). Cross-validation was a possible solution, but due to the small number of datapoints, the variation between different folds was enormous and the final result was very unstable, a value for the regularizer was then hard-coded. A more robust solution would have been desirable. Similarly, other choices in the methodology were evaluated by cross-validation, but it would have been preferable to be able to rely on internal metrics.

**References**

**Authors Statement**

LPC developed the methodology, implemented it, and wrote the report.