CDS Capstone DS-GA 1006

11/16/2017 Status Report - Team Beyond Google

Daniel Amaranto, Brenton Arnaboldi, Akash Kadel

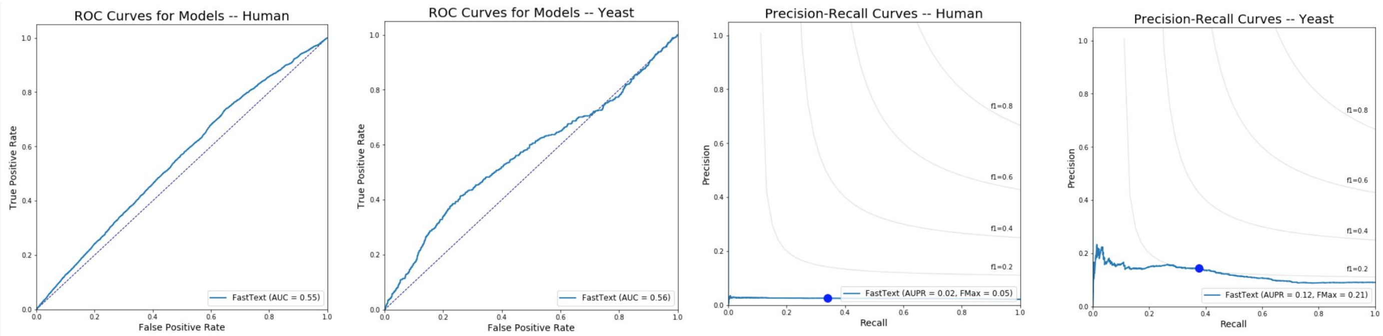
Evaluation/Results

As discussed in last week's status report, we are evaluating our models based on the following metrics:

- Area under the ROC Curve (AUC)
- Area under the Precision-Recall Curve (AUPR)
- F-Score (max)

For each model, the output matrix and the label matrix are 2D-arrays with dimensions *N x D*, where *N* = number of test set proteins and *D* = number of Go Terms. We are calculating average precision and average recall based on all the 147 (number of GO terms) values we estimated.

The results from our **FastText** model are illustrated below. The AUC scores for FastText range around **0.55-0.56** for both human and yeast proteins. The AUPR curve for Human proteins is a bit discouraging – we need to investigate this result. The blue dot in the AUPR curves indicates the point at which the F-score is maximized. So far our results are improving from the baseline. The FastText F-score was **0.05**, LSTM model achieves **0.07** for humans. Our **GRU** model achieves a better value of **0.09**.



Summary: Our major concerns from the Discussion section of the Research Memo persist. The relatively small dataset size, the sparsity of positive labels, and the multilabel prediction are resulting in very low F-score. The imbalance of classes (there are very few positive predictions of each label relative to the number of sequences) gives very high accuracy when making no predictions. To improve our results we are working on the following:

1. Incorporating k-mers (The reason we have not included it in our progress report is due to the computational requirements to use Deep Learning in such sparse data.)
2. Extensive Hyperparameter tuning (The reason of not including it here is same as above).
3. Training multiple deep learning models to learn each GO term separately.