PI: Arjun Krishnan

Mentee: Jesus Vazquez

Data: 05/18/2018

**Week 1:**

Cosine Similarity vs. Pearson Correlation

The question is to understand the similarities and differences between Cosine Similarity and Pearson Correlation.

The Cosine similarity has a range of 0 to 1, if both of the vectors are positive but it can take negative values if the one of the vectors has negative numbers. The Cosine similarity will give you the cosine angle between the two vectors. The Pearson correlation will you an index of how closely associated both vectors are to each other. We might often think of a Cosine Similarity as similar to a Pearson Correlation Coefficient. The formulas of both methods are as follows:

|  |  |
| --- | --- |
| Cosine Correlation | Pearson Correlation |
| /Users/Epifanio/Desktop/Screen Shot 2018-05-22 at 11.25.11 AM.png | /Users/Epifanio/Desktop/Screen Shot 2018-05-22 at 11.25.17 AM.png |

The Cosine Similarity coefficient is calculating by dividing the sum of the products of both vectors by product of the square-roots of the sum of both vectors squared. In contrast, the correlation coefficient is calculated by dividing the covariance of both vectors by the square-roots of the sums of their departures from the mean squared. Both of these calculations are fairly similar and will lead to similar results.

ROC Curve vs. PRC Curve

The ROC curve can help us determine how well of a prediction a model has. In our NPL project we have decided to use the ROC curve to better understand how well of a precision our ML model has on predicting the genes that are associated with the condition (Alzheimer). The x-axis in out ROC curve are the false positives (1-Specificity) and the y-axis is the Sensitivity. The specificity is a measure of how often the model says that you have the condition (or the event of interest) given that the condition is present. The sensitivity is a measure on how often the model says that you don’t have condition given that you do not have the condition. In practice, we want both sensitivity and specificity to be high. When the condition or the event is very rare then the ROC is not the best. The precision recall curve is best used for very rare conditions or events. The x-axis for the PR curve is the recall and the y-axis is the precision.

**Given the only a small about of genes/words are associated with the condition we might want to study the PR curve instead of the ROC curve.**

Machine Learning – Annotated Bibliography

Natural Language Processing, Neural Networks, Random Forest, Gene Ontology and more:

**Word and sentence embedding tools to measure semantic similarity of Gene Ontology terms by their definitions -** Dat Duong 1,∗ , Wasi Uddin Ahmad 1 , Eleazar Eskin 1,3, Kai-Wei Chang 1 , Jingyi Jessica Li 2,3,∗