David Liu, Branley Mmasi

The Multi-Label Classification Problem

Throughout the semester, we have discussed various machine learning and classification problems, and numerous possible approaches. However, all our models operated under the assumption that each example had only one label, which is a fine assumption in many real-world scenarios, but often is not the case. In this project, we wish to analyze various multi-label classification models and measure their performance on various types of data sets. We would like to determine which models are better suited to deal with the multi-label classification problem, and what types of modifications are necessary to accommodate that.

For most of our project, we will be following T.-H. Chiang, H.-Y. Lo & S.-D. Lin's 2012 paper 'A Ranking-based KNN Approach for Multi-Label Classification,' which details a KNN model for multi-label classification. In the paper, they discuss the standard multi-label classification models and compare their performance to KNN model they proposed. Their general conclusion is that the KNN model performs as well if not better than the standard models. We would like to implement their algorithm, as well as the standard algorithms, and compare their performance on various datasets, starting with "Multi-label Classification of enzyme substrates" (https://www.kaggle.com/datasets/gopalns/ec-mixed-class), which is a real-world dataset on classifying enzymes into EC class(es).

In terms of our experiment, we plan on training and testing each of our models on our datasets and comparing their results. The paper used 10-fold 5-repeat cross-validation on different hyperparameters including ranking method, so we will likely use a similar method. The paper details four different ways they gathered performance results, Hamming Loss, one error, average precision, and root-mean-square-error. We will likely use one evaluation metric to begin with, and then use the other three if we have time.

We believe that a multi-label model has a great deal of application and impact in many fields. For instance, with the enzyme dataset we are using, scientists working in drug discovery can use different enzymes to achieve different results, based on its EC class. However, it is not so easy to predict the EC class, especially when an enzyme can take on more than one class, so improvements in multi-label classification could improve the medicine and biology industry, leading to more discoveries in key fields. Additionally, sites like Spotify and Youtube give recommendations for songs and videos, and having a multi-label classification model could enable them to give better recommendations since their songs and videos wouldn’t be limited to one type of category.