Intro_to_MultiLabel_Classification

Overview:

MultiLabel Classification originated from the investigation of text categorisation problem, where each document may belong to several predefined topics simultaneously.

Multi-label classification of textual data is an important problem.

This can be thought as predicting properties of a data-point that are not mutually exclusive. For instance, this can be employed to find the genres that a movie belongs to, based on the summary of its plot.

Other examples include multiple label classification for news articles and emails.

In multi-label classification, the training set is composed of instances each associated with a set of labels, and the task is to predict the label sets of unseen instances through analyzing training instances with known label sets.

So how is it different from multiclass classification? Well, MultiClass classification means a classification task with more than two classes; each of them mutually exclusive. The classification makes the assumption that each sample is assigned to one and only one label.

For example, a fruit can be either an apple or a pear but not both at the same time. Whereas, an instance of multi-label classification can be that a text might be about any of religion, politics, finance or education at the same time or none of these.

Problem Statement and Approach

Our task is to predict the tags for stackoverflow posts.

The dataset is provided by natural language processing Coursera course and its resources are available here github link.

Complete project can be implemented using the following IPynb Notebook. (Originally run on Google Colab.)

Dataset

Dataset consists basically of train(100,000), validation(30,000) and test(20,000) posts.

Train data and validation data has 2 primary columns. “Title” of the stackoverflow post and “Tag” corresponding to it. There are 100 different tags available in this dataset.

Test data consists only of “Title” for each post and we need to predict the corresponding “Tag”.

Implementation

The notebook includes implementation of following tasks:

- Text Preprocessing
- Transforming text into a vector using bag-of-words approach and Tfidf-Vectorizer
- Training a multi-label classifier using training dataset by following One-vs-Rest approach where *k* classifiers are trained corresponding to *k* distint labels.
- Basic classifier implemented is a linear model called LogisticRegression.
- Evaluation metrices are defined like Accuracy, F1-score, Average-precision-score for evaluating the trained model on validation data.
- Finally the trained and evaluated model is used to predict tags for test dataset.