ML Task2 – NLU

A data set was obtained from KAGGLE. This data set was augmented a little bit and used for the project. Originally it had 15,000 training data and 4500 test data. All of the data were labeled with intents.

Many machine learning, algorithms, work considered, including naïve bayse, logistic regression, support vector machines, trees, random forests, and recurrent neural networks.

All of these algorithms were tried briefly and accuracy were tested. Since naive bayse considerably good accuracy, and it was also a very simple algorithm, it was decided to go ahead with that, and to improve it further.

Naïve bayse algorithm was ideal for the situation, but the biggest issue was with the fact that it did not consider the order of the words taken for the input. For example “good bye” and “bye good” were considered both as farewell by the algorithm.

Hence, if we corrected this, it would have been a very accurate result. This was done by using n-grams. This drastically increased the accuracy

Testing this is very simple to train the data. You have to input the training data set and it’s all given in the comments of the code. The test data can also be inputted using the same way, and it’s all included in the comments of the code.

Naive Bayes

thatware.co

In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

using Bayesian probability terminology, the above equation can be written as

$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$