NLP Task by
Mohamed Adel gomma Mohamed Elghannam

Report

=Which techniques you have used while cleaning the data if you have cleaned it?

- Firstly I have removed the duplicate values so it does not affect the train of my model
- there was no null values

=Why have you chosen this classifier? (E.g. I used Multinomial Naive Bayes because it is easy to interpret with text data and there are more than two outcomes).

- I have chosen support vector classifier because it is very effective in multi class classification and I can also control the data points that can enter the margin (C argument) .

- SVM is more effective in high dimensional spaces(especially when using one hot encoding which have increased my dimensional spaces)

-SVM is effective in cases where the number of dimensions is greater than the number of samples.

-I tried other classifiers but SVC got the best accuracy results (98.77%).

=How do you deal with (Imbalance learning)?

I used SMOTE for oversampling the imbalanced labels .

SMOTE works by utilizing a k-nearest neighbor algorithm to create synthetic data. SMOTE first start by choosing random data from the minority class, then k-nearest neighbors from the data are set. Synthetic data would then be made between the random data and the randomly selected k-nearest neighbor. Let me show you the example below.

The procedure is repeated enough times until the minority class has the same proportion as the majority class.

= How can you extend the model to have better performance?

- I used linear kernel instead of radial based function and poly which got me the best result
- I have increased the training data and shuffled it in the split process so the model can see all the job names and be able to predict the industry

= How do you evaluate your model? (i.e. accuracy, F1 score, Recall)

I have used Accuracy classification score to evaluate my model , because in multiple classification this function computes subset accuracy : the set of labels predicted (0,1,2,3 that i have made) must exactly match the corresponding set of labels in y_true.

So I want to get all the correct predictions even for 0 or 1 or 2 or 3

$$Accuracy = \frac{Correct\ Predictions}{All\ Predictions} = \frac{TP + TN}{TP + TN + FP + FN}$$

=What are the limitations of your methodology or Where does your approach failed? (e.g. your predictions are biased because you do not have enough data for a certain class)

One of the most important limitations is that model training must be made by all the data, not 80% of data because if we trained with 80% only the other jobs will be unseen and the model will not be able to predict all the industries . So I have increased the training dataset and shuffled it before splitting .