Why should one not look at the test set?

Machine learning is about working on unknown situation. Machine should train such a way that it can also encounter new data. If we train our model with train and test data as well, it will be a good model. But it might not show great result on unknown data. It probably biased on certain amount of data. Therefore, we should one not look at the test set.

Why is it useful to look at the confusion matrix?

Confusion Matrix consists of four different parameters like TP, FP, FN, TN that describe the performance of a model. We can directly get the predictive analytics like Precision, Recall, F1-score and Accuracy of a model.

What is the difference between using accuracy and using F-score?

In real world, we encounter a lot of Dataset with imbalanced class. Where a number of certain class of data is higher than other classes. So if we only calculate the accuracy of model it will only consider those outcome which are correct. On the other hand, F1-score consider the error by a model as well.

What is the difference between macro F-score and micro F-score and what different functions and implications do they have?

F1-score is a function of Precision and Recall. Per class F1-score is just the harmonic mean of Precision and Recall of that class. As follows:

F1-score = 2 × (precision × recall)/(precision + recall)

Now if we want to convert these numbers into a single number, we can use several methods like macro, micro, weighted average F1-score.

Macro Average is just an arithmetic mean of per class F1-Score.

Macro-F1 = (Class_1_f1 + Class_1_f2 + Class_1_f3) / 3 = macro-average f1-Score

Whereas, Micro Average

What baselines could you use for the binary sentiment classification and for the six-class topic classification? What would their performance be?

What happens with cross-validation? Did using it influence your results?

What features are used in the classification task you ran?