Summary of Miniproject 2 Analysis

Tuesday, March 8, 2022 11:05 AM

We have two Dataset [TwentyNewsGroup, Sentiment140]

TwentyNewsGroup: we only considered 4 out of 20 classes because computation was heavy and PC computing power wass too much

Sentiment140: This has 1.6million training samples so we only considered 6% of the training samples due to resource constraints and takes +30 mins to train one model

HyperParameter Tuning

① Feature Selection → Due to time constraints, we're sticking with CountVectorizer only
Due to resource constraints, parameter {min_df, max_df} we choose for each datasets are:

TwentyNewsGroup: {0.01, 0.7}
↳ 1% ↳ 70%

Sentiment140 : {0.005, 0.8}
↳ 0.5% ↳ 80%

features are selected within these min, max threshholds

② Hyperparameter for norm chosen as cost for SoftmaxRegression
↳ We use 5-fold validation on both L1 and L2-norm and extract accuracy on validation and choose the best one.

L2 → Best Norm 66.33% > 66.31%
↳ L1

③ Hyperparameter Tuning:

Naive Bayes $\alpha = \{0.001, 0.005, 0.010, 0.015, 0.020\}$

Softmax Regression $\lambda = \{0.01, 0.05, 0.10, 0.50, 1.00\}$

→ We apply 5-fold validation and apply all hyperparameter for both model on BOTH Dataset and select best model and selected best hyperparameter for both models

Twenty News Group → α for NB = 0.010
→ λ for SR = 0.50

Sentiment140 → α for NB = 0.001
→ λ for SR = 1.00

③ We applied best hyperparameter for both models to test datasets and selected the best models out of two

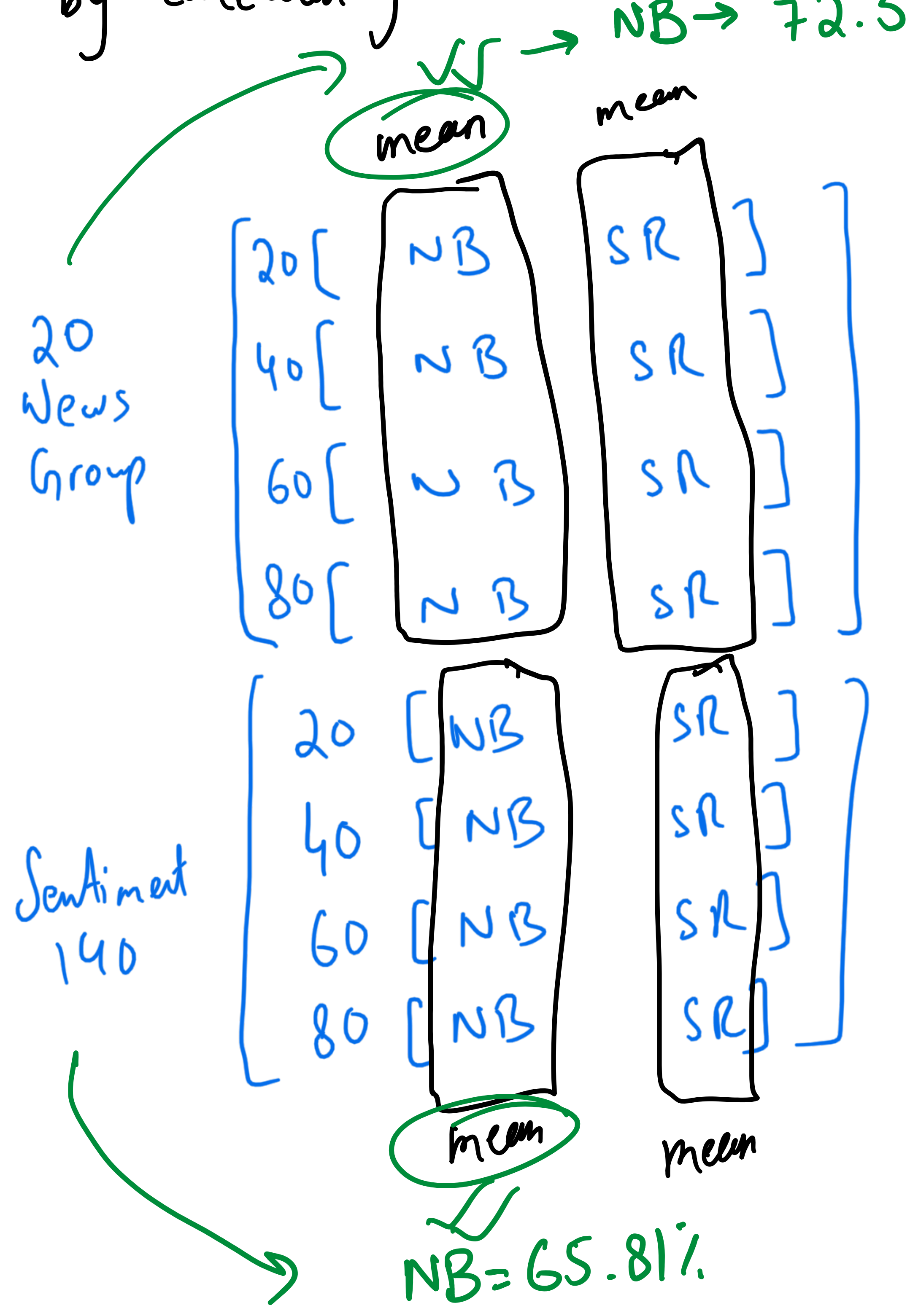
TwentyNewsGroup → NB = 73.85%

Sentiment140 → SR = 65.46%

④ Draw Confusion Matrix to get a sense of True Positive, False Positive [Need more work on this]

⑤ Using best hyperparameters and varying the traing dataset by {20%, 40%, 60%, 80%} and then finding the optimum model by calculating mean of all accuracies.

→ NB → 72.52%


mean
mean
20 News Group
20
NB
SR
40
NB
SR
60
NB
SR
80
NB
SR
Sentiment 140
20
NB
SR
40
NB
SR
60
NB
SR
80
NB
SR
mean
mean
NB = 65.81%