# CS1571 HW 4 Report

Tianjian Meng

12-06-2017

1. Scores

| | False positive | False negative | Overall |
|---|---|---|---|
| Fold_1 | 0.034782608695652174 | 0.06413043478260876 | 0.09891304347826087 |
| Fold_2 | 0.03369565217391304 | 0.05326086956521739 | 0.08695652173913043 |
| Fold_3 | 0.03260869565217391 | 0.05543478260869565 | 0.08804347826086957 |
| Fold_4 | 0.03152173913043478 | 0.07717391304347826 | 0.10869565217391304 |
| Fold_5 | 0.033659066232356136 | 0.0738327904451683 | 0.10749185667752444 |
| Avg | 0.03325364051293197 | 0.06476852858074332 | 0.09802216909367528 |

2. Statistical analysis

| | Positive | Negative |
|---|---|---|
| Fold_1_train | 0.6058136375984787 | 0.39418636240152133 |
| Fold_1_dev | 0.6065217391304348 | 0.3934782608695652 |
| Fold_2_train | 0.6058136375984787 | 0.39418636240152133 |
| Fold_2_dev | 0.6065217391304348 | 0.3934782608695652 |
| Fold_3_train | 0.6060853029068188 | 0.3939146970931812 |
| Fold_3_dev | 0.6054347826086957 | 0.39456521739130435 |
| Fold_4_train | 0.6060853029068188 | 0.3939146970931812 |
| Fold_4_dev | 0.6054347826086957 | 0.39456521739130435 |
| Fold_5_train | 0.6059782608695652 | 0.39402173913043476 |
| Fold_5_dev | 0.6058631921824105 | 0.3941368078175896 |

Because all the negative instances are located in the front of the dataset and all the positive instances are located in the back of the dataset, nearly all folds have similar distribution. Since positive ratio is higher than negative ratio in all dev data folds, the risk that classifier would predict the data into positive would be lower. So the fact that false positive ratio is low than false negative ratio makes sense.

3. Compare with just choosing majority class

| | False positive | False negative | Overall |
|---|---|---|---|
| Fold_1 | 0.3934782608695652 | 0.0 | 0.3934782608695652 |
| Fold_2 | 0.3934782608695652 | 0.0 | 0.3934782608695652 |
| Fold_3 | 0.39456521739130435 | 0.0 | 0.39456521739130435 |
| Fold_4 | 0.39456521739130435 | 0.0 | 0.39456521739130435 |
| Fold_5 | 0.3941368078175896 | 0.0 | 0.3941368078175896 |
| Avg | 0.39404477287546186 | 0.0 | 0.39404477287546186 |

Basically speaking, just choosing the majority class would result in the situation that false positive ratio equals to the ratio of negative samples and 0 false negative ratio. In this dataset, the class distribution is not very imbalanced, so just choosing the majority class would get a much higher error rate than using naive bayes. But if it is a very imbalanced dataset (like positive ratio or negative ratio is higher than 95%), just choosing the majority class may have a better result than naive bayes does.