Homework-4

April 30, 2020

Exercise 1: Linear SVM

Data Processing (see code in homework4.py)

Gradient Derivation

Given our objective function, our gradient function is:

$$\nabla F(\beta) = 2\lambda\beta \; + \frac{\rho}{n} \sum_{i=1;\, y=+1}^{n} \nabla \ell_{hh}(y_i, x_i^T\beta) \; + \; \frac{1-\rho}{n} \sum_{i=1;\, y=-1}^{n} \nabla \ell_{hh}(y_i, x_i^T\beta) \tag{1}$$

Where $\nabla \ell_{hh}(y_i, x_i^T\beta)$ is as follows. ***Note:*** to differentiate with respect to β, it is necessary to represent the t term in $\nabla \ell_{hh}(y,t)$ as $x_i^T\beta$.

$$\nabla \ell_{hh}(y_i, x_i^T\beta) := \begin{cases} 0 & \text{if } y_i x_i^T\beta > 1+h \\ \frac{1}{2h}(1+h-y_i x_i^T\beta)(-y_i x_i) & \text{if } |1-y_i x_i^T\beta| \le h \\ -y_i x_i & \text{if } y_i x_i^T\beta < 1-h \end{cases} \tag{2}$$

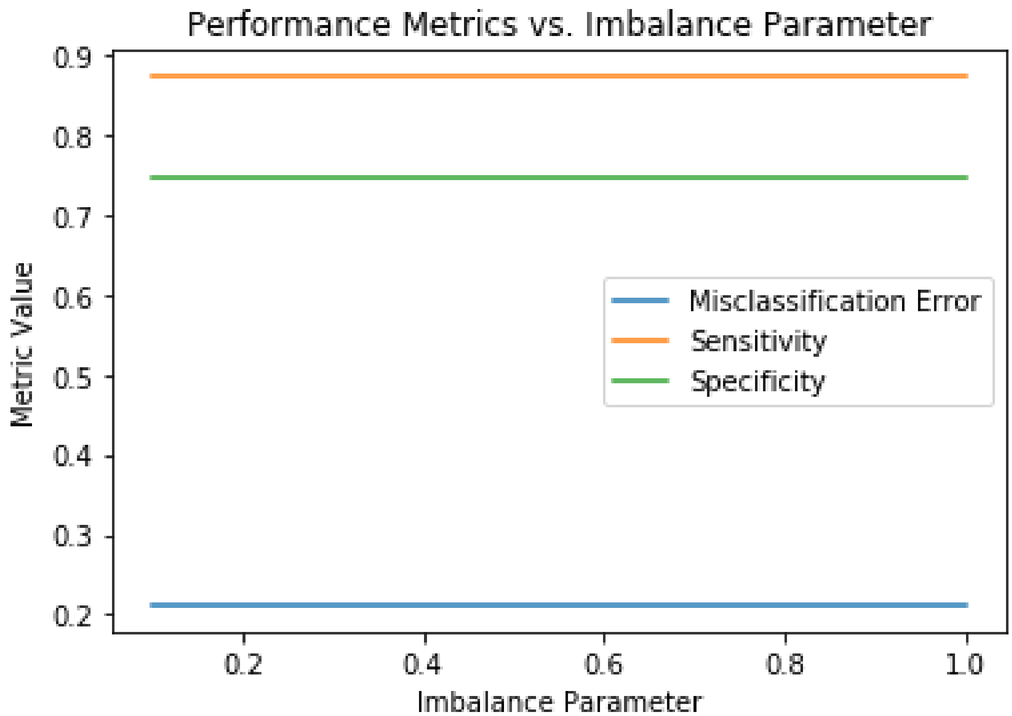
Function Definitions (see code in homework4.py)

Training Performance ($\lambda = 1, \rho = 1$*)*

```
General misclassification error is: 21.16%.
Achieved sensitivity is: 87.30%.
Achieved specificity is: 74.60%.
```

Training Performance ($\lambda = 1, \rho = 0.1, 0.2, ..., 0.9, 1.0$*)*

A recurring problem emerges (and is found in others' work, as well). The variation in ρ seems to have a negligible effect on β, and hence a negligible effect on all performance metrics. Outside of derivation and coding issues, it is believed that qualities inherent to the data may be causing this behavior. Hence, the plotted lines are nearly flat (they increase slightly for the much larger values of ρ).



Varying λ & ρ

We tune the algorithm along two dimensions: $\lambda = 0.1, 0.2, ..., 0.9, 1.0$ and $\rho = 0.1, 0.2, ..., 0.9, 1.0$. Our grid of values for λ and ρ is hence of size $10\ x\ 10$. To calculate area-under-the-curve, where the ROC curve is defined as $\frac{sensitivity}{1-specificity}$, we use ***sklearn.metrics.roc_auc_score***. The pair of λ and ρ values that optimize area-under-the-curve for the ***validation data***, and the optimal AUC value itself, are as follow. Lastly computed is the sensitivity and specificity for this pair of values, on the ***testing set***.

```
For the validation set:

The optimal value of lambda is: 0.1.
The optimal value of rho is: 0.1.
The corresponding, optimized area-under-the-ROC-curve is: 0.98.

For the testing set (using above optimal parameters):

The achieved sensitivity is: 38.46%
The achieved specificity is: 85.71%
```

Exercise 3

Linear Regression Insomnia

For Bobbie, the time taken to train the model will play a significant part in the decision regarding which optimization algorithm he should use. That means if he is training a linear regression model (say, using gradient descent), I would recommend he use the ***Fast Gradient*** algorithm as discussed in class. This includes the use of *backtracking line search* for an adaptive step-size, and has the general advantage of converging to desirable solution faster than standard gradient descent algorithm. Note that if possible, I also recommend that Bobbie incorporate some form of *penalized/regularized* linear regression, since he is also tasked with validating the model on a dataset. This way, his model is discouraged from overfitting, without adding much time to the process.

The Next Big Thing

I would not invest anything into Joey's company. The inclusion of the test set during training will naturally lead to a poorly-generalizable model. The model (never tested) will follow the noise of availabe samples too closely, or overfit. The combining of the data means that while Joey may predict his testing samples perfectly, the variance of his model will be high. That is, performance on other datasets will be undoubtedly inconsistent. I will urgently relay to Joey that, clearly, this means his model is not ready for real-world use. I will then re-evaluate why Joey has stayed my colleague for so long.