CS57300 Homework — Homework 4 , Problem 1 — (April 10, 2016)
Jialiang Yu ( 0028161618 )

**Q1:**

**A:**

The AIC and BIC are both methods of assessing model fit penalized for the number of estimated parameters.

AIC stands for Akaike Information Criterion and does not assume model is correct.

$$AIC = -\frac{2}{N} lnP[t = f(x_0; \hat{\theta})] + 2\frac{d}{N}$$

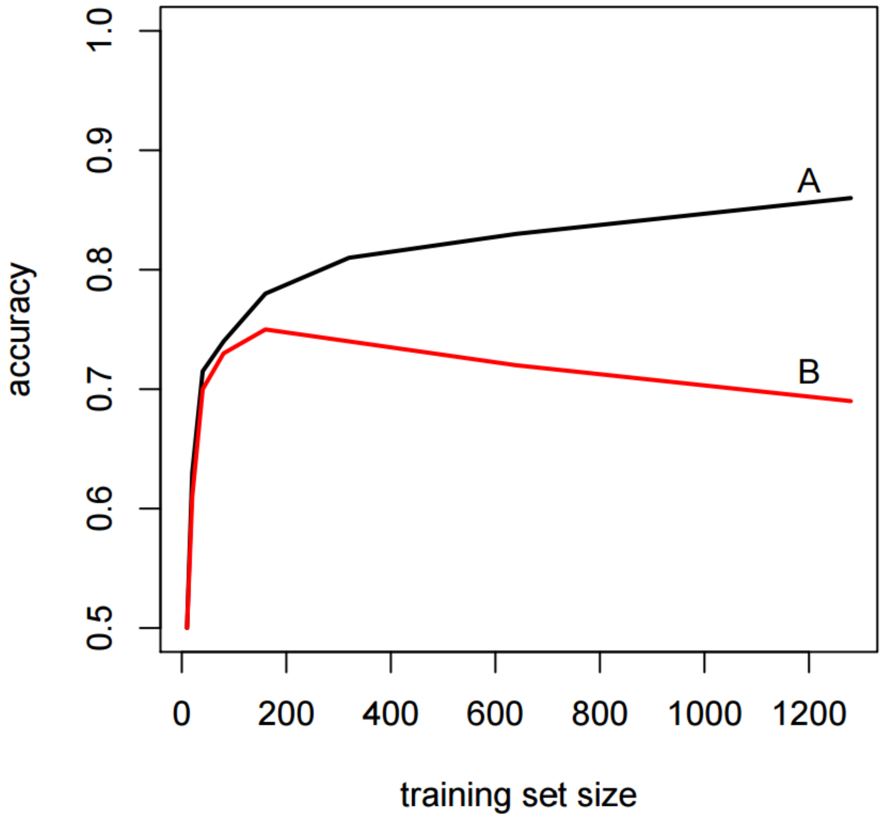
BIC stands for Bayesian Information Criterion and assumes model is correct.

$$BIC = -\frac{2}{N} lnP[t = f(x_0; \hat{\theta})] + \frac{d}{N} logN$$

AIC is basically suitable for a situation where you don't necessarily think there's a model so much as a bunch of effects of different sizes, and you're in a situation you want to get good prediction error. As such, as the sample size expands, the AIC choice of model expands as well, as smaller and smaller effects become relevant.

BIC on the other hand basically assumes the model is in the candidate set and you want to find it. BIC tends to hone in on one model as the number of observations grows. AIC really doesn't. As a result, at large $n$, AIC tends to pick somewhat larger models than BIC.

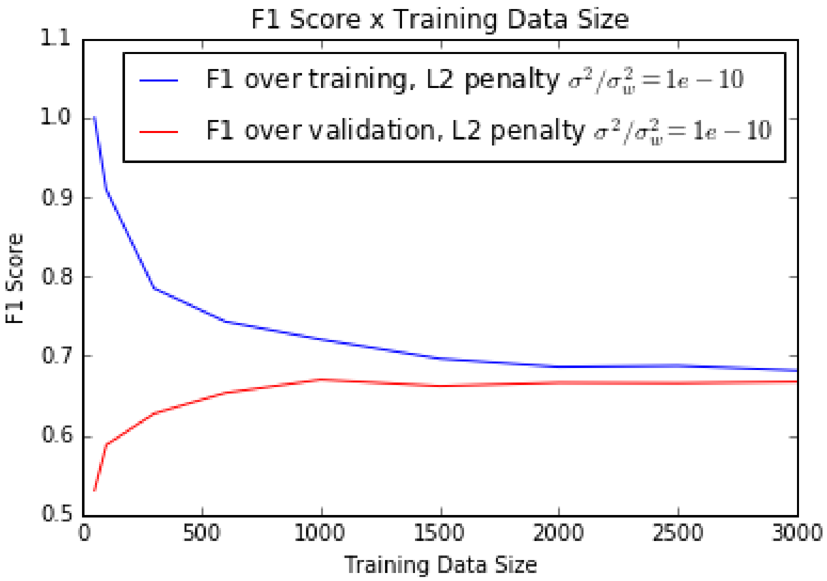
**Q2:**
**A:**



(i) Training set: B Test set: A

(ii) The accuracy of test set will always increase as the size of training set increases. This observation is obtained from the experiments in this assignment. The accuracy of training set could be larger or smaller than that of test set.

**Q3:**
**A:**



(i) Linear regularized least squares $f_s(x) = x^T w$, $\lambda = \frac{\sigma_\epsilon^2}{\sigma_w^2}$

$$\begin{aligned}
\hat{w}_{RLS}(s) &= \arg\min_w \frac{1}{2}\sum_{n=1}^{n}(y_i - Logistic({x_i}^T w))^2 + \frac{\lambda}{2}||w||^2 \\
&= \arg\max_w e^{-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}(y_i - Logistic(x_i^T w))^2} \cdot e^{-\frac{\lambda}{2\sigma_\epsilon^2}||w||^2} \\
&= \arg\max_w e^{-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}(y_i - Logistic(x_i^T w))^2} \cdot e^{-\frac{\lambda}{2\sigma_\epsilon^2}||w||^2} \\
&= P(Y|X,\theta) \cdot P(\theta)
\end{aligned}$$

So

$$Y|X,\theta \sim N(X\theta, \sigma_\epsilon^2)$$
$$\theta \sim N(0, \sigma_w^2)$$

So the mean is 0, and the variance is $\sigma_w^2$.

(ii) The $F_1$ score formula is below:

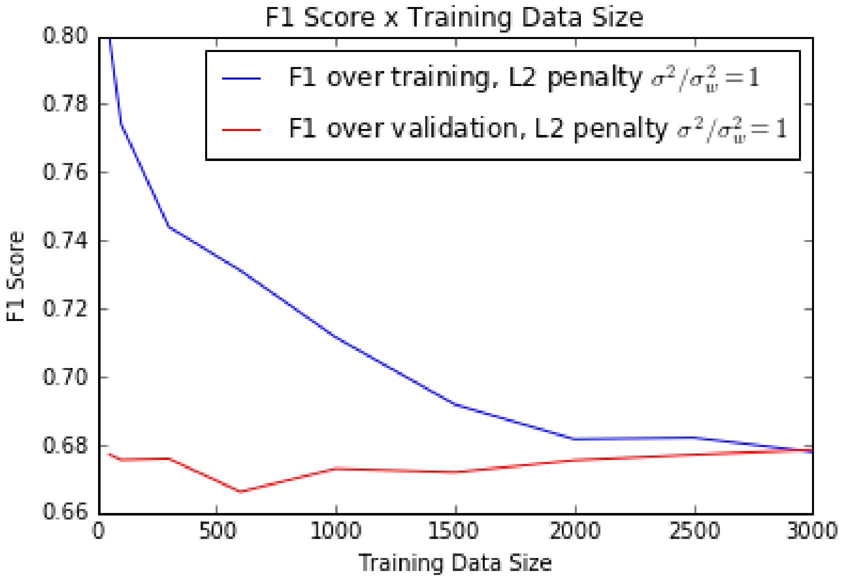
$$F_1 = \frac{2TP}{2TP + FN + FP}$$

The percentage of false prediction, including False Positive and False Negative goes down as the size of training set increases. This scenario happens because of the improvement in training model. When the size of training set is small, the model could overfit to ensure the accuracy of training set is very high which makes the accuracy of test low. When the size of training set increase, the potential of overfitting goes down. So the red curve which represents the test accuracy goes up when the model doesn't perform overfitting as the size of training size increases.

(iii) The blue curve goes down as the size of training set increases because the ratio of True Positive decreases. When the size of training set is small, the model could overfit which leads to the high accuracy of the model. However, as the size of training set increases, the phenomenon of overfitting is emitted which leads to the decrease of accuracy of validation.

(iv) The training curve and the validation curve converge as the decrease of accuracy in training and increase of accuracy in test because of the alleviation of overfitting. Besides, the training set accuracy will always be higher than the test set.

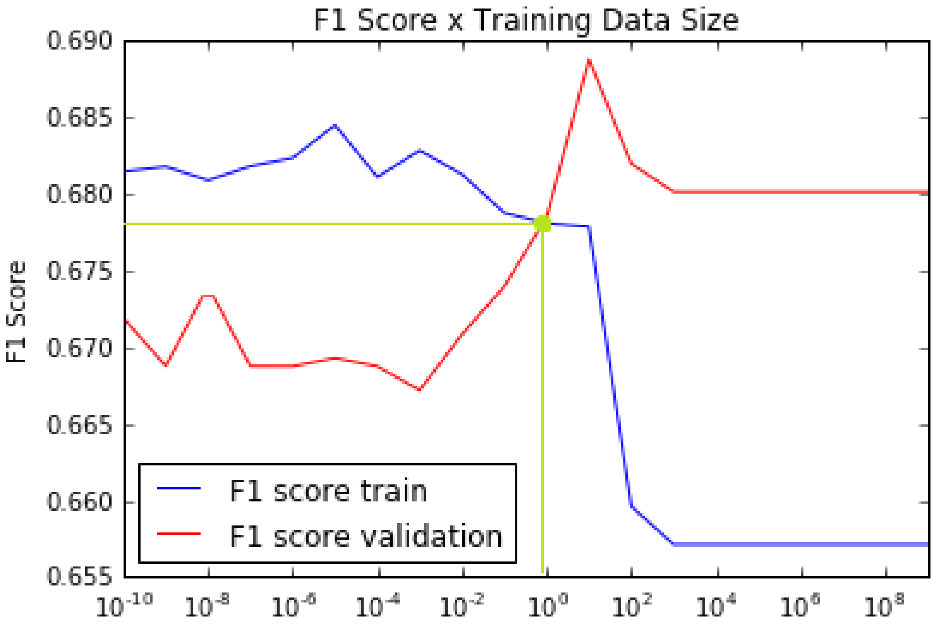
(v)



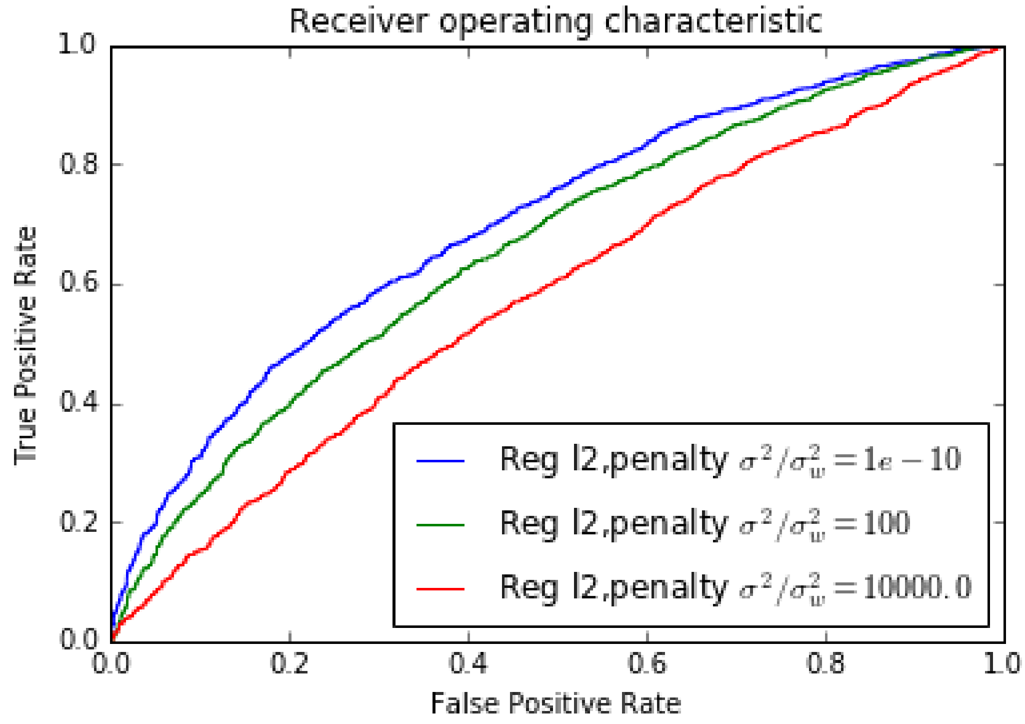
By replace $sigma_{sq}$ with 1, the regularization is stronger. The model could be underfitting. Since $\frac{\sigma}{\sigma_w^2}$ increases, $\sigma_w^2$ will decrease dramatically. We know $w \sim N(0, \sigma_w)$, so the parameters will be restricted in a small range of value around zero, which will probably result in underfitting.

(vi) Test the ratio incremented by 10 each time. That is let $\frac{\sigma^2}{\sigma_w^2}$ be $e^{-10}, e^{-9}, e^{-8}, e^{-7}, e^{-6}, e^{-5}, e^{-4}, e^{-3}, e^{-2}, e^{-1}, e^{1}, e^{2}, e^{3}$.

The best result is when $\frac{\sigma^2}{\sigma_w^2} = 1$


F1 Score x Training Data Size
F1 Score
0.690
0.685
0.680
0.675
0.670
0.665
0.660
0.655
$10^{-10}$
$10^{-8}$
$10^{-6}$
$10^{-4}$
$10^{-2}$
$10^{0}$
$10^{2}$
$10^{4}$
$10^{6}$
$10^{8}$
F1 score train
F1 score validation

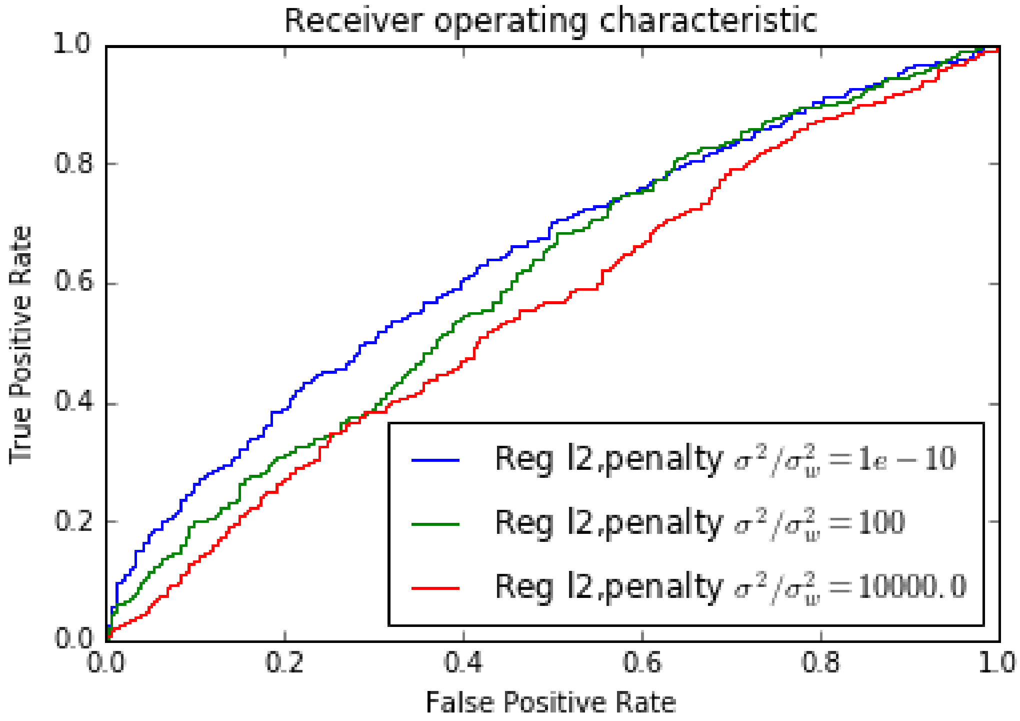
**Q4:**
**A:**



(i) The training set and the validation set are the same. So when $\lambda = \frac{\sigma}{\sigma_w^2}$ gets smaller, the model will be getting overfitting. The $sigma_{sq}$ are different: $1e-10, 100, 1e4$, and the accuracy is measured by the area under each curve. Since the training set and the validation set are the same, a smaller $\lambda$ will provide a higher accuracy.

(ii) Since the lower $\frac{\sigma^2}{\sigma_w^2}$ is, the higher accuracy that the ROC curve will achieve. So we will know that $\frac{\sigma}{\sigma_w^2} = 1000$ will sit in the middle of the other curves.

(iii)



The curve for $\frac{\sigma^2}{\sigma_w^2} = 100$ is not consistently worth than the one for $\frac{\sigma^2}{\sigma_w^2} = 10^{-10}$ because the accuracy of the model decreases as the validation set changes.