HW2

Zheyu Ding (zd12)
M.Viginesh Vaibhav (vm24)

1 Gradient and Hessian of NLL(θ) for logistic regression (10 points)

1. $g(z) = \frac{1}{1+e^{-z}} = (1+e^{-z})^{-1}$

   $1 - g(z) = \frac{e^{-z}}{1+e^{-z}}$

   $\frac{\partial g(z)}{\partial z} = (-1)(1+e^{-z})^{-2}(e^{-z}) = \frac{e^{-z}}{(1+e^{-z})^2} = \left(\frac{1}{1+e^{-z}}\right)\left(\frac{e^{-z}}{1+e^{-z}}\right) = g(z)(1-g(z))$

   Hence Proved.

2. We know that, $h_\theta(x) = \frac{1}{1+e^{(-\theta^T x)}}$

   Negative Log Likelihood function,

$$NLL(\theta) = -\sum y^i \log\left(\frac{1}{1+e^{(-\theta^T x)}}\right) + (1-y^i)\log\left(\frac{e^{(-\theta^T x)}}{1+e^{(-\theta^T x)}}\right)$$

   Derivative of NLL (after applying simplification over the terms using the above formula),

$$\frac{\partial NLL(\theta)}{\partial \theta} = -\sum [\, y^i \left(\frac{e^{(-\theta^T x)}}{1+e^{(-\theta^T x)}}\right) - (1-y^i)\left(\frac{1}{1+e^{(-\theta^T x)}}\right)]\, x^i$$

$$= -\sum [y^i \frac{\left(1+e^{(-\theta^T x)}\right)}{1+e^{(-\theta^T x)}} - \frac{1}{1+e^{(-\theta^T x)}}]\, x^i$$

$$= -\sum (y^i - h_\theta(x^i))\, x^i$$

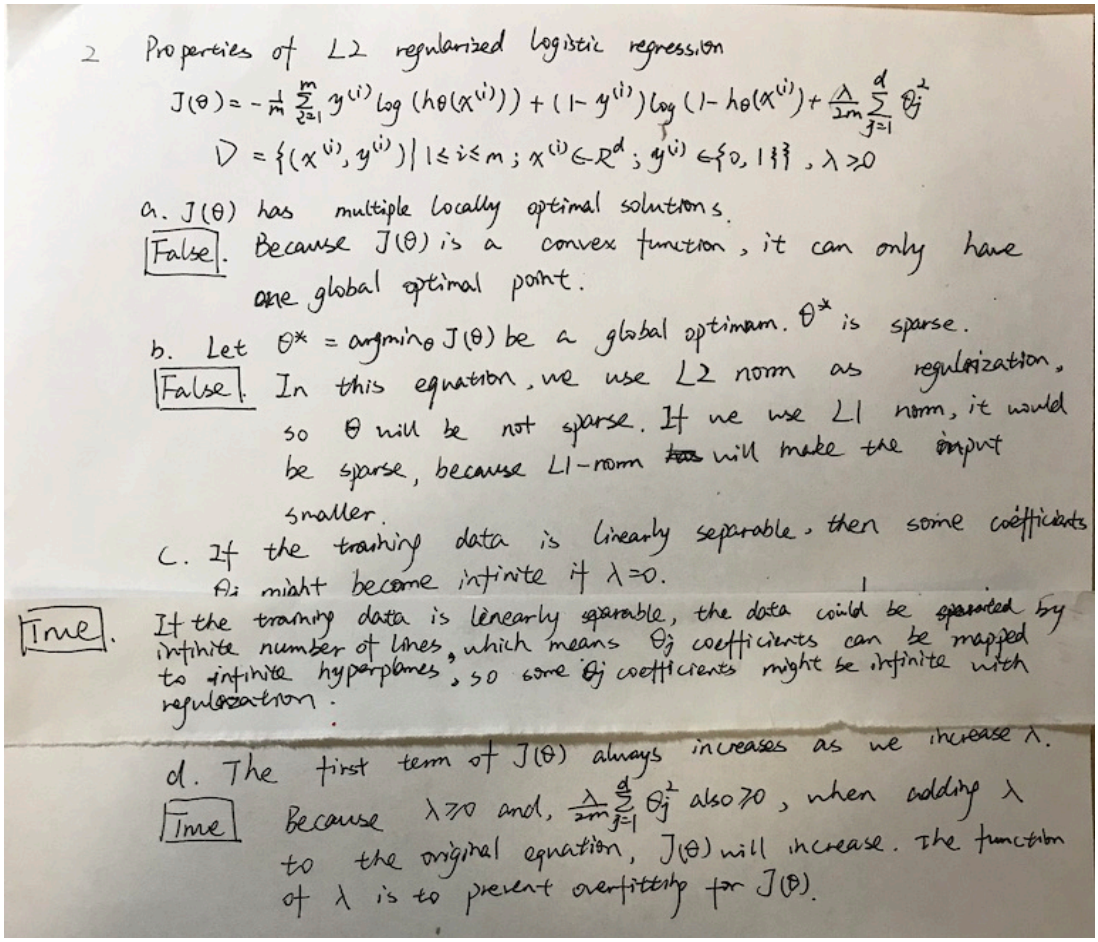
$$= -\sum (h_\theta(x^i) - y^i) x^i$$

   Hence Proved.

3. $H = X^T S X = \sum {x^i}^T \left(h_\theta(x^i)\right)\left(1 - h_\theta(x^i)\right) x^i = \sum (x^i)^2 (h_\theta(x))(1 - (h_\theta(x))$

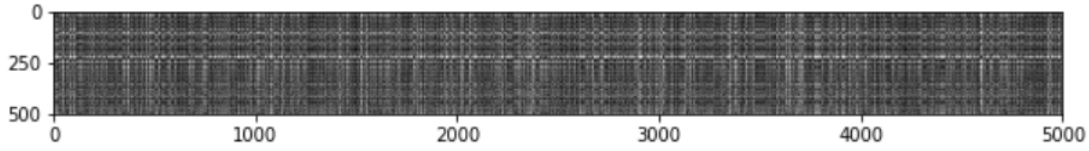
   Here, $(h_\theta(x^i))(1 - (h_\theta(x^i))$ is positive, since $h_\theta(x^i) > 0$, so $\left(1 - h_\theta(x^i)\right) > 0$, therefore their product is positive. Regardless of the value of x^i, $(x^i)^2$ is always positive. Therefore, $H = X^T S X = \sum (x^i)^2 \left(h_\theta(x^i)\right)\left(1 - h_\theta(x^i)\right) > 0$

2 Properties of L2 regularized logistic regression (10 points)



3 Implementing a k-nearest-neighbor classifier (25 points)

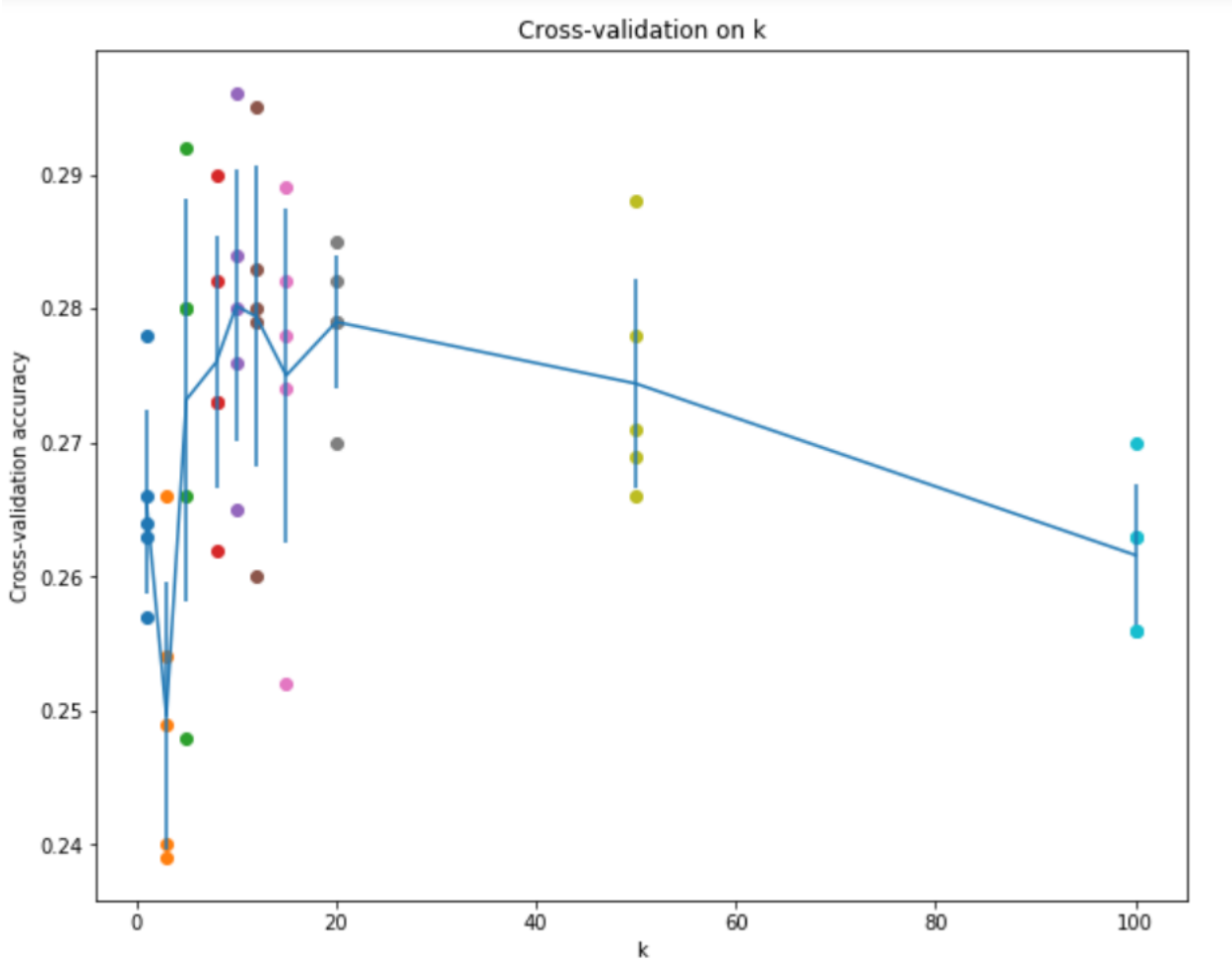
Here are the plots generated in this question:



Questions from the Jupyter Notebook:

- What in the data is the cause behind the distinctly bright rows?
- What causes the columns?

Answer: The rows are bright when the specific test example cannot be classified under any of the available classes in the training data. The columns are bright when none of the samples in the test data fall under the training example's particular class.



The above figure is the plot depicting the cross-validation accuracies of the various values of k. The highest possible accuracy is 28.2%, which is obtained when the value of k is 10.

4: Implementing logistic regression

Problem 4A3: Prediction using a logistic regression model (5 points)

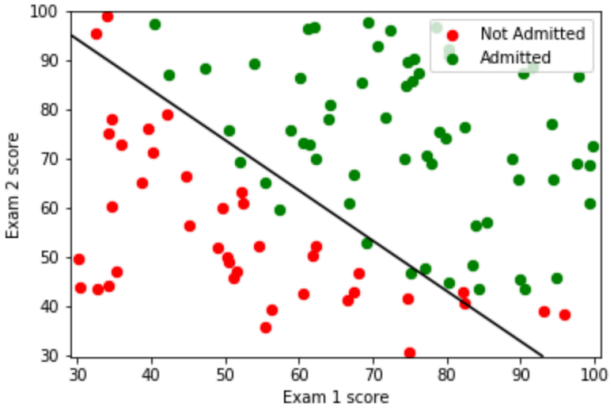


Fig. Own model

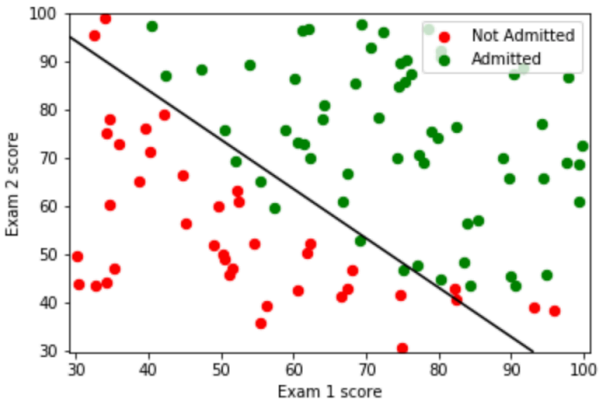


Fig. Sklearn model

Problem 4, Part B: Regularized logistic regression (20 points)

Problem 4B3: Varying λ (3 points)

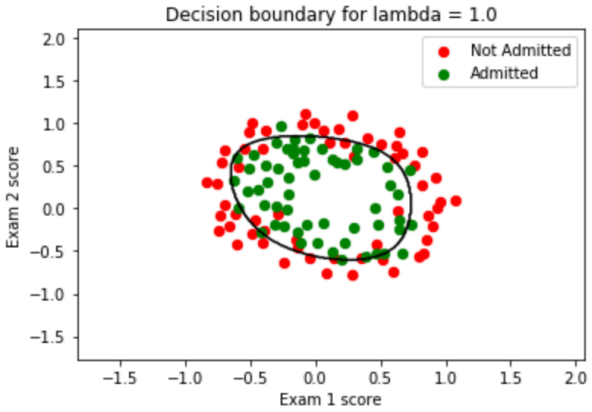


Fig. Training data lambda = 1

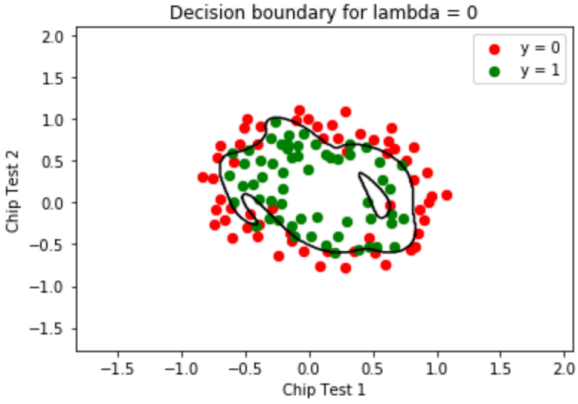


Fig. Training data lambda = 0

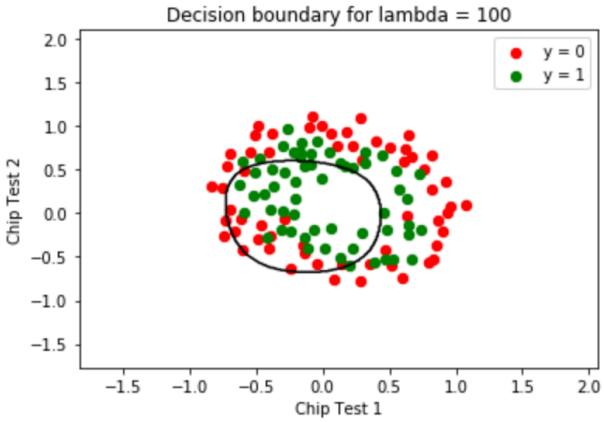


Fig. Training data lambda = 100

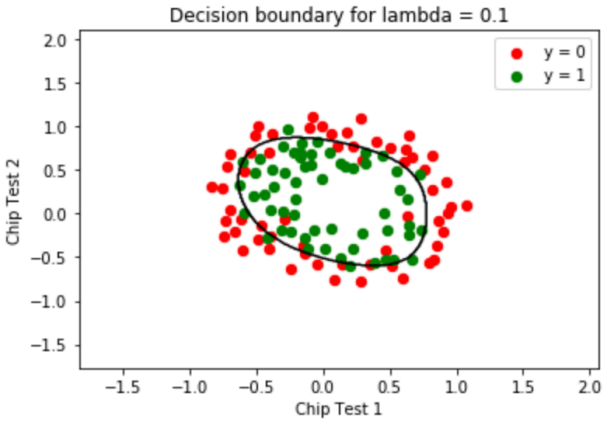


Fig. Training data lambda = 0.1

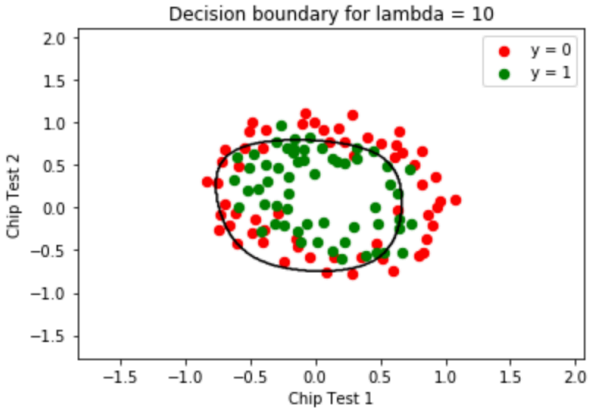


Fig. Training data lambda = 10

Judging from the plots above, when lambda is too large like 10 and 100, the boundary will be underfitting. When lambda is too small like 0, the boundary will be overfitting.

Problem 4B4: Exploring L1 and L2 penalized logistic regression (5 points)

lambda	L1 Loss	L2 Loss	L1 non-zeros	L2 non-zeros
0.001	0.2679	0.2960	24	28
0.01	0.2911	0.3167	19	28
1.0	0.4381	0.4678	7	28
3.0	0.6137	0.5496	3	28
10.0	0.6931	0.6126	0	28

Chart. Compare L1 and L2

According to the chart, there are several differences between L1 and L2:

1) Compared to L2, L1 penalty will make the model sparser especially when the value of regularization increases, which means L1 is more appropriate for sparse problem.
2) With the value of regularization increasing, both regularization methods will lead to the loss larger.
3) L1 model's loss increases faster than the L2 model's loss with the same regularization, which means L1 is more robust than L2.
4) The absolute value of each theta in L1 is tend to larger than the theta in L2.

Problem 4 Part C: Logistic regression for spam classification (10 points)

Best lambda	Regularization type	Preprocess	Accuracy	Number of non-zeros
0.100	L2	Standardization	0.9297	58
0.600	L2	Log	0.9434	58
1.100	L2	Binary	0.9277	58
4.600	L1	Standardization	0.9219	52
1.600	L1	Log	0.9440	43
3.600	L1	Binary	0.9258	39

Chart. Compare L1 and L2

According to the chart, the best lambda in L2 model will be larger than L1 model, which means L1 regularization will be more robust. And L1 model will be sparser than L2 model. In this task, I will recommend L1 model with Log preprocess. Not only is the accuracy the highest, but also some of the parameters will be zero, which means the model will become less complexity and work faster. Therefore, I will recommend L1 model with Log preprocess.