Assignment 2, COMP 540

Chen Zeng(cz39), Zhihui Xie(zx18)

January 31, 2018

1 Gradient and Hessian of $NLL(\theta)$ for logistic regression

Answer1 Because of:

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

We can get:

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

Take the derivation on both side of the z, then we can get:

$$g'(z)\left(1+e^{-z}\right) - g(z)\,e^{-z} = 0$$

Finally, after some basic transformation, we can get:

$$\frac{\partial g(z)}{\partial z} = g(z)\left(1-g(z)\right)$$

Therefore, the conclusion is proved.

Answer2 As we know:

$$NLL(\theta) = \frac{1}{2}\sum_{i=1}^{m}\left(y^{(i)} - \theta^T x^{(i)}\right)^2$$

Take the derivation on it of the parameter θ_j, we can get:

$$\frac{\partial NLL(\theta)}{\partial \theta_j} = x_j^{(i)}\sum_{i=1}^{m}\left(\sum_{j=1}^{d}\left(x_j^{(i)}\theta_j\right) - y^{(i)}\right)$$

Therefore, take the derivation on $NLL(\theta)$ of vector θ, we can get:

$$\frac{\partial NLL(\theta)}{\partial \theta} = \sum_{i=1}^{m} \left(\theta^T x^{(i)} - y^{(i)}\right) x^{(i)}$$

It's the same as another format:

$$\frac{\partial NLL(\theta)}{\partial \theta} = \sum_{i=1}^{m} \left(h_\theta \left(x^{(i)}\right) - y^{(i)}\right) x^{(i)}$$

Therefore, the conclusion is proved.

Answer3 Assume z is a nonezero vector with the size n × 1, X is a m × n full rand matrix, $S_{i,i}$ is the ith element in the diagonal of S and it is strictly positive.

$$\begin{aligned} z^T H z &= z^T X^T S X z \\ &= (Xz)^T H (Xz) \\ &= \sum_{i=1}^{m} S_{i,i} (\sum_{j=1}^{k} x_j z_j)^2 \\ &> 0 \end{aligned}$$

So H is positive definite.

2 Properties of L2 regularized logistic regression

Answer1 False. J(θ) is a convex function and has a global minimum, so it does not have multiple locally optimal solutions.

Answer2 False. Ridge regularization drives θ components to zero but not to exactly zero while Lasso regularization drives many θ components to exactly zero especially when the regularization parameter λ is large.

Answer3 True. For example, if the data points of two classes are linear separable in 2-dimension plane and a single line L crossing the origin can separate them into two classes, the line L can be represented as ax + by = 0 and the coefficients a and b can be infinite.

Answer4 False. As λ increases from zero, the model will experience from overfitting, fitting, to unfitting. It means the model will becomes better at first and then becomes worse. This is the abstract changing of the model as λ increases and some of the detailed changing is not predictable. The first term of J is NLL, which is used to judge whether the model is good or not. Therefore, as λ increases, the first term of J will not always increses.

3 Implementing a k-nearest-neighbor classifier

Please help to refer to the knn folder.

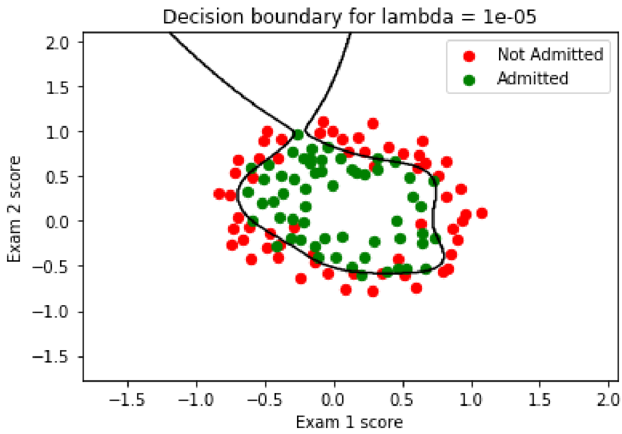

Figure 4.1: Overfitting

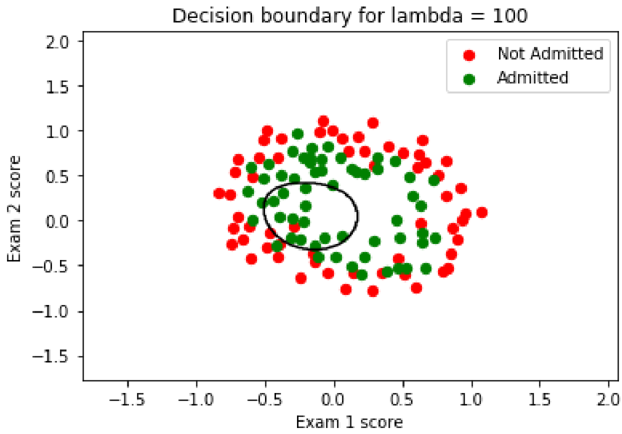

Figure 4.2: under-fitting

4 Implementing logistic regression

Problem 3, Part B: Varying λ

- When = 0.00001, the modle is overfitting on this data set. Figure 4.1 shows the decision boundary for overfitting.
- When = 100, the modle is under-fitting on this data set. Figure 4.2 shows the decision boundary for under-fitting.

Problem 3 Part C: Fitting regularized logistic regression models (L2 and L1)

- We'd like to recommend to use L1 regularization. The accuracy with these two regularization are similar, while it equals to 0.943359375 when we use L2 regularization and it equals to 0.944010416667 when we use L1 regularization. However, L1 regularization leads more θ to be exactly zero which make the parameters to be sparser.