Homework 2

Collaborators:
Name: Junlin Yin
Student ID: 3160104340

Problem 2-1. A Walk Through Linear Models

(a) Perceptron

Answer:

1. Given the size of testing set 1000:
   When the size of training set is 10, training error rate: 0, testing error rate: 10.9%;
   When the size of training set is 100, training error rate: 0.146%, testing error rate 1.38%.
2. Given the size of testing set 1000:
   When the size of training set is 10, average number of iterations: 52;
   When the size of training set is 100, average number of iterations: 2241.
3. Given the size of testing set 1000:
   The Perceptron Machine will iterate infinitely if the maximum number of iterations is not defined.
   The training error cannot reach zero, and the testing error is much higher than that in linearly separatable cases.

(b) Linear Regression

Answer:

1. Given the size of testing set 1000:
   The training error rate: 3.89%, the testing error rate: 4.68%.
2. Given the size of testing set 1000:
   The training error rate: 13.1%, the testing error rate: 14.3%.
3. The training error rate: 49.0%, the testing error rate: 55.0%.
4. The training error rate: 5.00%, the testing error rate: 6.60%.

(c) Logistic Regression

Answer: Here is the breif derivation of the logistic gradient: As the question suggests, the likelihood function can be presented as:

$$P(y|x,\omega) = \hat{y}^y * (1-\hat{y})^{1-y}$$

where $\hat{y} = \frac{1}{1+\exp[-\omega^T x]}$, so the total-log-likelihood function is:

$$\begin{aligned} L &= \sum_{i=1}^{m} l_i \\ &= \sum_{i=1}^{m} y^{(i)} \log \hat{y^{(i)}} + (1 - y^{(i)}) \log (1 - \hat{y^{(i)}}) \end{aligned}$$

where m is the number of samples. Thus we can define the loss function as the minus log-likelihood function, so the gradient is:

$$\begin{aligned} \nabla_\omega J &= -\frac{\partial}{\partial \omega} L \\ &= \sum_{i=1}^{m} \frac{\hat{y^{(i)}} - y^{(i)}}{\hat{y^{(i)}}(1 - \hat{y^{(i)}})} * \frac{\partial \hat{y^{(i)}}}{\partial \omega} \\ &= \sum_{i=1}^{m} x^{(i)} * (\hat{y^{(i)}} - y^{(i)}) \\ &= [\hat{y} - y] X^T \end{aligned}$$

Based on this:

1. Given the size of testing set 1000:
   The training error rate: 0.29%, the testing error rate: 1.09%.
2. Given the size of testing set 10000:
   The training error rate: 11.8%, the testing error rate: 12.8%.

(d) Support Vector Machine

Answer:

1. Given the size of testing set 1000:
   The training error rate: 0, the testing error rate: 3.36%.
2. Given the size of testing set 1000:
   The training error rate: 0, the testing error rate: 1.03%.
3. Average number of support vectors: 3.5.

Problem 2-2. Regularization and Cross-Validation

(a) Implement Ridge Regression and use LOOCV to tune...

Answer:

1. According to LOOCV, the optimal lambda is 10.
2. When $\lambda = 0$, $\sum_{i=0}^{P} \omega_i^2 = 14.6$; when $\lambda = 10$, $\sum_{i=0}^{P} \omega_i^2 = 1.23$.
3. When $\lambda = 0$, the training error is 0 and the testing error is 10.7%; When $\lambda = 10$, the training error is 0 and the testing error is 7.23%.

(b) Implement Logistic Regression and use LOOCV to tune...

Answer: According to LOOCV, the optimal lambda is 0.001.
When $\lambda = 0$, the training error is 0 and the testing error is 6.73%; When $\lambda = 0.001$, the training error is 0 and the testing error is 6.58%.

Problem 2-3. Bias Variance Trade-off

(a) True or False:

Answer:

1. False. Once the model is defined, the bias is defined, too. It doesn't work no matter how many samples are added if the model is over-simple.
2. False. We don't like neither high-bias and high-variance. Although models with high variance can fit the training data perfectly, they don't generalize to testing data very well. However, high-variance problems can be tackled by adding more training samples, which is likely to get further improvement compared with high-bias cases.
3. True. When the parameters get more, there's a greater chance that training data are not enough, thus creating too complicated models.
4. False. Regularization is used to mitigate over-fitting, where testing error will be reduced but training error will not necessarily. On the contrary, the training error will probably increase a little bit because some mistakes may be allowed.
5. False. By using very large λ, the over-fitting penalty gets extremely huge, but the high-bias penalty gets relatively small. Therefore, models with large λ usually suffer from high-bias problem.