Data 100, Spring 2023

Discussion #8

Ridge and LASSO Regression

1. Earlier, we posed the linear regression problem as follows: Find the θ value that minimizes the average squared loss. In other words, our goal is to find $\hat{\theta}$ that satisfies the equation below:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, L(\theta) = \underset{\theta}{\operatorname{argmin}}\, \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

Here, $\mathbb{X}$ is a $n \times (p+1)$ matrix, θ is a $(p+1) \times 1$ vector and $\mathbb{Y}$ is a $n \times 1$ vector. Recall that the extra 1 in $(p+1)$ comes from the intercept term. As we saw in lecture, the optimal $\hat{\theta}$ is given by the closed-form expression $\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$.

To prevent overfitting, we saw that we can instead minimize the sum of the average squared loss plus a regularization term $\lambda g(\theta)$. The optimization problem for such a *regularized* loss function then becomes:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, L(\theta) = \underset{\theta}{\operatorname{argmin}} \left[\frac{1}{n}\|\mathbb{Y} - \mathbb{X}\theta\|_2^2 + \lambda g(\theta)\right]$$

- If use the function $g(\theta) = \sum_{j=1}^{p} \theta_j^2 = ||\theta||_2^2$, we have "ridge regression". Recall that g is the ℓ_2 norm of θ, so this is also referred to as "ℓ_2/L_2 regularization".
- If we use the function $g(\theta) = \sum_{j=1}^{p} |\theta_j| = ||\theta||_1$, we have "LASSO regression". Recall that g is the ℓ_1 norm of θ, so this is also referred to as "ℓ_1/L_1 regularization".

In this question, we intentionally choose to regularize also on the intercept term to simplify the mathematical formulation of the ridge and LASSO regression. In practice, we would not actually want to regularize the intercept term (and you should always assume that there should not be a regularization on the intercept term).

For example, if we choose $g(\theta) = ||\theta||_2^2$, our goal is to find $\hat{\theta}$ that satisfies the equation below:

$$\begin{aligned}\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, L_2(\theta) &= \underset{\theta}{\operatorname{argmin}} \left[\frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2 + \lambda||\theta||_2^2\right] \\ &= \underset{\theta}{\operatorname{argmin}} \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbb{X}_{i,\cdot}^T\theta)^2 + \lambda\sum_{j=0}^{d}\theta_j^2\right]\end{aligned}$$

Recall that λ is a hyperparameter that determines the impact of the regularization term. Like ordinary least squares, we can also find a closed-form solution to ridge regression: $\hat{\theta} = (\mathbb{X}^T\mathbb{X} + n\lambda\mathbf{I})^{-1}\mathbb{X}^T\mathbb{Y}$. For LASSO regression, there is no such closed-form expression.

(a) Suppose we are dealing with the OLS case (i.e., don't worry about regularization yet). As model complexity increases, what happens to testing error of the model trained using OLS? What about the training error?

(b) Now suppose we choose one of the above regularization methods, for some regularization parameter $\lambda > 0$, and then we solve for our optimum. In terms of variance, how does a regularized model compare to ordinary least squares regression (assuming the same features between both models)?

(c) In ridge regression, what happens if we set $\lambda = 0$? What happens as λ approaches ∞?

(d) If we have a large number of features (10,000+) and we suspect that only a handful of features are useful, which type of regression (Lasso vs Ridge) would be more helpful in interpreting useful features?

(e) What are the two benefits of using ridge regression over OLS?

Cross Validation

2. After running 5-fold cross validation, we get the following mean squared errors for each fold and value of λ when using Ridge regularization:

Fold Num	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	Row Avg
1	80.2	70.2	91.2	91.8	83.4
2	76.8	66.8	88.8	98.8	82.8
3	81.5	71.5	86.5	88.5	82.0
4	79.4	68.4	92.3	92.4	83.1
5	77.3	67.3	93.4	94.3	83.0
Col Avg	79.0	68.8	90.4	93.2	

Suppose we wish to use the results of this 5-fold cross validation to choose our hyperparameter λ, among the following four choices in the table. Using the information in the table, which λ would you choose? Why?

3. You build a model with two hyperparameters, the coefficient for the regularization term (λ) and our learning rate (α). You have 4 good candidate values for λ and 3 possible values for α, and you are wondering which λ, α pair will be the best choice. If you were to perform five-fold cross-validation, how many validation errors would you need to calculate?

4. Explain how you would use leave-one-out cross validation to choose λ as in part 2.

.

Guessing at Random

5. A multiple choice test has 100 questions, each with five answer choices. Assume for each question that there is only one correct choice. The grading scheme is as follows:

   - 4 points are awarded for each right answer.
   - For each other answer (wrong, missing, etc), one point is taken off; that is, -1 point is awarded.

   A student hasn't studied at all and therefore selects each question's answer uniformly at random, independently of all the other questions.

   Define the following random variables:

   - R: The number of answers the student gets right.
   - W: The number of answers the student does not get right.
   - S: The student's score on the test.

   (a) What is the distribution of R? Provide the name and parameters of the appropriate distribution. Explain your answer.

   (b) Find $\mathbb{E}(R)$.

   (c) True or False: SD$(R) =$ SD(W).

   (d) Find $\mathbb{E}(S)$, the student's expected score on the test.

   (e) Find SD(S).