DS 100/200: Principles and Techniques of Data Science **Date: July 29, 2019**

Discussion # 10 Solutions

Name:

Cross Validation and Regularization

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

   Solution: There are $4 \times 3 = 12$ pairs of λ, γ and each pair will have 5 validation errors, one for each fold.

2. In the typical setup of k-fold cross validation, we use a different parameter value on each fold, compute the mean squared error of each fold and choose the parameter whose fold has the lowest loss.

   ◯ A. True

   ◯ B. False

   Questions 3, 4, 5, and 6 are all connected.

3. Elastic Net is a regression technique that combines L_1 and L_2 regularization. It is preferred in many situations as it possesses the benefits of both LASSO and Ridge Regression. Minimizing the L2 loss using Elastic Net is as follows, where $\lambda_1, \lambda_2 >= 0$, $\lambda_1 + \lambda_2 = \lambda$, $\lambda > 0$.

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i} (y_i - \theta x)^2 + \lambda_1 \sum_{j=1}^{p} |\theta_j| + \lambda_2 \sum_{j=1}^{p} \theta_j^2$$

   Suppose our goal was to get sparse parameters, i.e. we want as many parameters as possible to be zero. Which of the following choices for λ_1, λ_2 are most consistent with this goal, assuming $\lambda = 1$? **There is only one correct answer.**

   ◯ A. $\lambda_1 = 0, \lambda_2 = 1$

   ◯ B. $\lambda_1 = 0.5, \lambda_2 = 0.5$

   ◯ C. $\lambda_1 = 1, \lambda_2 = 0$

Solution: We know that LASSO encourages sparsity in our optimal weights. Setting λ_1 to 1 means we are using LASSO.

4. What happens to bias and variance as we increase the value of λ? Assume $\lambda_2 = \lambda_1$. **There is only one correct answer in each part.** You will be asked to justify why in the next question.

   (a) Bias:

   ◯ A. Bias goes up

   ◯ B. Bias stays the same

   ◯ C. Bias goes down

   (b) Variance:

   ◯ A. Variance goes up

   ◯ B. Variance stays the same

   ◯ C. Variance goes down

5. Justify why by marking the true statements. **Select all that apply for each part.**

   (a) Bias:

   ☐ A. Bias goes down because increasing λ reduces over fitting.

   ☐ B. Bias goes down because bias is minimized when $\lambda_2 = \lambda_1$.

   ☐ C. Bias goes up because increasing λ penalizes complex models, limiting the set of possible solutions.

   ☐ D. Bias goes up because the loss function becomes non-convex for sufficiently large λ.

   ☐ E. None of the above

   (b) Variance:

   ☐ A. Variance goes down because increasing λ encourages the value of the loss to decrease.

   ☐ B. Variance goes down because increasing λ penalizes large model weights.

   ☐ C. Variance goes up because because increasing λ increases bias.

   ☐ D. Variance goes up because increasing λ increases the magnitude of terms in the loss function.

   ☐ E. None of the above

6. What happens to the model parameters $\hat{\theta}$ as $\lambda \to \infty$, i.e. what is $\lim_{\lambda \to \infty} \hat{\theta}$? **Select all that apply.**

   ☐ A. Converge to 0.

   ☐ B. Diverge to infinity.

   ☐ C. Converge to values that minimize the L2 loss.

   ☐ D. Converge to equal but non-zero values.

☐ E. Converge to a sparse vector.

Solution: The model parameters go to 0.

Logistic Regression

1. You have a classification data set consisting of two (x, y) pairs $(1, 0)$ and $(-1, 1)$. You decide that you want your feature vector $\mathbf{x}$ (the input to a model) for each pair to be a two-element column vector $[1 \quad x]^T$.

You run an algorithm to fit a model for the probability of $Y = 1$ given $\mathbf{x}$:

$$\mathbb{P}\left(Y = 1 \mid \mathbf{x}\right) = \sigma(\theta \cdot \mathbf{x})$$

where $\sigma(t) = \frac{1}{1+\exp(-t)}$. Your algorithm returns $\hat{\theta} = \left[-\frac{1}{2} \quad -\frac{1}{2}\right]$.

(a) Calculate $\hat{\mathbb{P}}\left(Y = 1 \mid \mathbf{x} = [1 \quad 0]^T\right)$

Solution:

$$\begin{aligned}
\hat{\mathbb{P}}\left(Y = 1 \mid \mathbf{x} = [1 \quad 0]^T\right) &= \sigma\left(\left[-\tfrac{1}{2} \quad -\tfrac{1}{2}\right] \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) \\
&= \sigma\left(1 \times -\frac{1}{2} + 0 \times -\frac{1}{2}\right) \\
&= \sigma\left(-\frac{1}{2}\right) \\
&= \frac{1}{1 + \exp(\frac{1}{2})} \\
&\approx 0.38
\end{aligned}$$

(b) The empirical cross-entropy loss (a.k.a. log loss) is given by:

$$L(\boldsymbol{\theta}, \mathbf{X}, \mathbf{y}) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(z_i) + (1 - y_i)\log(1 - z_i)]$$

where $z_i = \sigma(\theta \cdot \mathbf{X}_i)$. Let $\theta = [\theta_0 \quad \theta_1]$. Explicitly write out the empirical loss for the data set $(1, 0)$ and $(-1, 1)$ as a function of θ_0 and θ_1. Note that in this problem, $\mathbf{X}_i$ is the feature vector $\mathbf{x}$ defined in the original problem statement.

Solution:

$$z_i = \sigma(\theta \cdot \mathbf{x}_i) = \sigma\left(\begin{bmatrix}\theta_0 & \theta_1\end{bmatrix}\begin{bmatrix}1 \\ x_i\end{bmatrix}\right) = \sigma(\theta_0 + \theta_1 x_i)$$

For the data point $(1, 0)$, $\mathbf{x}_i = [1 \quad 1]^T$ and $y_i = 0$, so:

$$z_i = \sigma(\theta_0 + \theta_1 \times 1) = \sigma(\theta_0 + \theta_1)$$

For the data point $(-1, 1)$:

$$z_i = \sigma(\theta_0 + \theta_1 \times -1) = \sigma(\theta_0 - \theta_1)$$

We can then write the empirical loss as:

$$\begin{aligned} L(\boldsymbol{\theta}, \mathbf{X}, \mathbf{y}) &= -\frac{1}{2}\left[(1 \times \log(1 - \sigma(\theta_0 + \theta_1))) + (1 \times \log(\sigma(\theta_0 - \theta_1))\right] \\ &= -\frac{1}{2}\left[\log\left(1 - \sigma(\theta_0 + \theta_1)\right) + \log(\sigma\left(\theta_0 - \theta_1\right))\right] \end{aligned}$$

(c) Calculate the empirical loss for $\hat{\theta} = \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}^T$ and the two observations $(1, 0)$ and $(-1, 1)$.

Solution:

$$
\begin{aligned}
L(\hat{\theta}) &= -\frac{1}{2} \left[\log\left(1 - \sigma(\theta_0 + \theta_1)\right) + \log(\sigma\left(\theta_0 - \theta_1\right))\right] \\
&= -\frac{1}{2} \left[\log\left(1 - \sigma(-1)\right) + \log(\sigma\left(0\right))\right] \\
&= -\frac{1}{2} \left[\log\left(1 - \frac{1}{1+e}\right) + \log\left(\frac{1}{2}\right)\right] \\
&= -\frac{1}{2} \left[\log\left(\frac{e}{1+e}\right) + \log\left(\frac{1}{2}\right)\right]
\end{aligned}
$$