ECE M148 | Homework 3
Introduction to Data Science | Due: May 03, 12:00 PM
Instructor: Lara Dolecek | TAs: Harish GV, Jayanth Shreekumar

Please upload your homework to Gradescope by May 03, 12:00 PM.
You can access Gradescope directly or using the link provided on BruinLearn.
You may type your homework or scan your handwritten version. Make sure all the work is discernible.

1. Assume you have a dataset $\mathcal{D}$ with n samples. You want to create bootstrapped datasets of size k using sampling with replacement.

   (a) Assume you create one bootstrapped dataset of size k. Additionally, assume that we fix a data point $x \in \mathcal{D}$. What is the probability that x does not appear in the bootstrapped dataset?

   (b) Now, assume that $k = n$. What does the probability converge to as n goes to infinity? What does this limit imply about the percentage of the original dataset that will not be sampled at n gets large?

   (c) Assume that you create r bootstrapped datasets of size k each. Additionally, assume that we fix a data point $x \in \mathcal{D}$. What is the probability that x does not appear in any bootstrapped dataset?

2. In this question, let us consider the difference between lasso and ridge regularization. Recall that the lasso regularization of a vector β is $\lambda \sum_{i=1}^{k} |\beta_i|$ and that the ridge regularization is $\lambda \sum_{i=1}^{k} \beta_i^2$. Consider two vectors $x_1 = [4, 5]$ and $x_2 = [-2, 2]$. Additionally, set $\lambda = 1$.

   (a) What is the lasso regularization of x_1 and x_2? What is the change in the lasso regularization when going from x_1 to x_2?

   (b) What is the ridge regularization of x_1 and x_2? What is the change in the ridge regularization when going from x_1 to x_2?

   (c) In your own words, explain the effects of ridge vs lasso regularization.

3. **Coding Question** - Plot the Voronoi regions for $k = 1, 2, 3, 4$ using the k-nearest neighbours classifier on the points: [[1, 1], [4, 1], [2, 3], [3, 3], [3, 4], [5, 4], [6, 5], [4, 5]]. The first 4 points are in class 0 and the rest are in class 1. A .ipynb file has been provided with starter code to get you started. Did you find anything curious about the plots? How do you explain them?

4. **Coding Question** - Plot the logistic function $\frac{1}{1+e^{-(\beta_0+\beta_1 \times x)}}$ for $x \in [-10, 10]$ and the following parameter values:

   (a) $\beta_0 = 2$ and $\beta_1 = 1$

   (b) $\beta_0 = 10$ and $\beta_1 = 2$

   (c) $\beta_0 = 1$ and $\beta_1 = 10$

   (d) $\beta_0 = 1$ and $\beta_1 = 5$

   For what choices of β_0, β_1 does the function become steeper?

5. Recall the problem of ridge linear regression with n points and k features:

$$L_{Ridge}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{\beta}^T\boldsymbol{x}_i)^2 + \lambda\sum_{j=1}^{k}\beta_j^2$$

where λ is a hyper-parameter. The goal is to minimize $L_{Ridge}(\boldsymbol{\beta})$ in terms of $\boldsymbol{\beta}$ for a fixed training dataset $(y_i, \boldsymbol{x}_i)$ and parameter λ.

(a) In your own words, explain the purpose of using ridge regression over standard linear regression.

(b) As λ gets larger, how will this affect $\boldsymbol{\beta}$? What value do we expect $\boldsymbol{\beta}$ to converge on?

(c) Consider parameters $\boldsymbol{\beta}_\lambda$ that were trained using ridge linear regression with a specific lambda. Let us consider the test MSE using $\boldsymbol{\beta}_\lambda$. Note that the test MSE is the following formula for the test data

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{\beta}_\lambda^T\boldsymbol{x}_i)^2$$

and does not include regularization.

Sketch a plot of how you expect the Test MSE to change as a function of λ. Your sketch should be a smooth curve that shows how the test MSE changes as λ goes from 0 to ∞. Provide justification for your plot. Assume that the right most edge of the graph is where λ is at ∞. Additionally, assume that the linear regression without regularization is overfitting.

6. True of False questions. For each statement, decide whether the statement is True or False and provide justification (full credit for the correct justification).

   (a) In L_2 regularization of linear regression, many coefficients will generally be zero.

   (b) In the leave one out cross validation over the data set of size N, we create and train $N/2$ models.

   (c) 95% confidence interval refers to the interval where 95% of the training data lies.

   (d) If K out of J features have already been selected in Stepwise Variable Selection, then we will train $J - K$ new models to select the next feature to add.

   (e) $P(A|B) = P(B|A)$ if $P(A) = P(B)$ and $P(A)$ is not zero.