Homework 1

DATA 558, Spring 2021

Due April 9, 11:59pm

Instructions

This homework consists of a reading assignment, a mathematical exercise, and a coding exercise. Please submit your solutions via Gradescope. Solutions should consist of three files: a PDF containing your solutions to the non-coding questions, and **both** a Jupyter notebook (`.ipynb`) and the corresponding HTML export with your solution to the coding exercise. **All coding exercises must be completed in Python.** Please be sure to comment the code appropriately. Students are encouraged to discuss homework problems, particularly on Canvas and in the TA hours, but must submit their own solutions.

Reading Assigment

Please read the following sections in *Introduction to Statistical Learning*:

- 2.1: What is Statistical Learning?
- 2.2: Assessing Model Accuracy
- 5.1: Cross-Validation

Problems

Please submit your solutions as a single PDF file under the `Homework 1 - Problems` assignment in Gradescope.

Exercise 1

Compute the derivative of the function $f(\beta)$ with respect to the parameter $\beta \in \mathbb{R}$.

1. $f(\beta) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i\beta)^2$
2. $f(\beta) = \frac{\exp(x\beta)}{\exp(x\beta)+1}$
3. $f(\beta) = \log\left(\sum_{i=1}^{n} \exp(x_i\beta - 1)\right)$
4. $f(\beta) = |y - x\beta|$

Coding

Please implement your solutions to the coding exercise below in a Jupyter notebook. Export the completed notebook to HTML, then submit **both** the notebook (`.ipynb`) file and the HTML file under the `Homework 1 - Coding` assignment in Gradescope. **Please run all cells in your notebook prior to submission, so we can view their output.**

Exercise 2

In this exercise, we return to the topic of validation methods for hyperparameter selection. Here, our goal will be to implement "five-by-two" validation, which we saw in the context of k-nearest neighbor classification in Lab 1, for ridge regression. Please complete all steps below:

1. First, download the data. We will again use the `penguins` dataset. This can be obtained as a pandas dataframe via

   ```
   import pandas as pd
   file = 'https://raw.githubusercontent.com/mwaskom/seaborn-data/master/penguins.csv'
   penguins = pd.read_csv(file, sep=',', header=0)
   ```

2. Consider a regression problem in which we wish to predict a penguin's body mass (i.e. the `body_mass_g` feature) from the features `bill_length_mm`, `bill_depth_mm`, and `flipper_length_mm`. Define the features X and regression target Y accordingly. What are their dimensions?

3. Split X and Y into training and test sets using an 80-20 train / test split.

4. Instantiate and fit scikit-learn's `Ridge` model on the train data, with $\lambda = 1.0$. Note that scikit-learn uses the notation `alpha` for what we call λ.

5. We will use the *mean squared error* (MSE) as a measure of model performance. Given a vector of model predictions $\widehat{Y} \in \mathbb{R}^n$ and the true corresponding target values Y, the MSE is computed as

   $$\text{MSE}(\widehat{Y}, Y) = \frac{1}{n} \left\| \widehat{Y} - Y \right\|_2^2.$$

   Implement this function and compute the MSE of the model from step (3) on the training set.

6. Now use five-by-two validation to select a value for λ from set of values in the following array:

   ```
   lam_vals = np.logspace(-2,4,19)
   ```

   The numpy function `logspace()` produces a sequence of logarithmically-spaced values.

7. Report the best-performing choice of λ^* according to the validation procedure.

8. Fit a ridge regression model to the entire training data using this choice of regularization parameter.

9. Compute and report the MSE of the above model on the test set.