Assignment #1: Model Complexity, KNN, & Properties of Least Squares
STA9890
Statistical Learning for Data Mining

Assignment Parameters:

Date Assigned: 2024-02-05
Date Due: 2024-02-20 @ ~~5:45~~ 10:00pm

Submission Mechanism(s):

- Blackboard (strongly preferred)
- Email to instructor: michael.weylandt@baruch.cuny.edu
  Email submissions must be titled *exactly* as `STA9890-S2024-HW1-LASTNAME,FIRSTNAME.pdf`

Question 0: Formatting and Presentation (15 points)

Upload your submission to this assignment as a single `PDF` file on Blackboard. Ten points will be assigned based on formatting and presentation of your submission. For the best presentation, I recommend the use of LaTeX or similar software (*e.g.* Markdown + MathJax), but other software is allowed.

All code used to produce figures in your submission should be included at the end of the PDF document.

Question 1: Properties of OLS (12 points)

Prove the following properties of OLS:

(a) Suppose data is generated as $y = \boldsymbol{\beta}_*^\top \boldsymbol{x} + \epsilon$ for some mean-zero ϵ noise. Show $\hat{\boldsymbol{\beta}}$ is unbiased, *i.e.*, $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}_*$. (3 points)

(b) Suppose data is generated as $y = \alpha + \boldsymbol{\beta}_*^\top \boldsymbol{x} + \epsilon$ for some mean-zero ϵ noise and $\boldsymbol{x}$ that is mean zero. Show that α is equal to the average value of y. (3 points)

(c) Given training data of the form $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, show that the OLS in-sample prediction error (residuals) are given by $\hat{\boldsymbol{y}} - \boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{y}$. (3 points)

(d) Suppose we fit OLS with an intercept term. Show that the mean (unsquared) error (residual) must be zero. (3 points)

(e) Prove that $\boldsymbol{X}^\top \boldsymbol{X}$ is strictly positive-definite if the data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ has rank n ($n \geq p$). (3 points)

(f) Given n observations and p features, when does OLS achieve 0 training error? (You may assume the data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is full-rank) (3 points)

Question 2: K-Nearest Neighbors (33 points)

(a) Write a K-nearest neighbor classifier (15 points) from scratch.

Your function should perform the following

(a) Given a test point $\tilde{\boldsymbol{x}}$, compute the distances to each row of the training data matrix $\boldsymbol{X}$ using Euclidean (L_2 distance)

(b) Identify the K-nearest neighbors of $\tilde{\boldsymbol{x}}$

(c) Take the majority vote of the nearest neighbors

(b) Download the `zip` code data from the data page at `https://hastie.su.domains/ElemStatLearn/`. Extract the training and test data for the `3`s and `8`s.

(c) Apply your K-NN classifier with $K = 5$ to the training data and use it compute test error on the test data (misclassfication rate). (3 points)

(d) Repeat this process for multiple values of K and find which one minimizes i) training error; and ii) test error. Interpret your results (8 points)

(e) Standardize your data so that each column of $\boldsymbol{X}$ has the same variance (*e.g.*, using the `scale` function in `R`) and repeat the above process. Do your results change? (7 points)

Question 3: The Bias-Variance Trade-Off in OLS (20 points)

In this problem, you will examine the bias-viarance trade-off of OLS (least squares) regression.

(a) In class, we briefly derived the Bias-Variance-Irreducible decomposition of mean squared error. Give a formal proof of the following result

$$\mathbb{E}[\text{MSE}] = \text{Bias}^2 + \text{Variance} + \text{Irreducible Noise}$$

including definitions of all four terms. (4 points)

(b) Generate $n = 25$ training samples from the following linear model:

$$y = 3x_1 - 2x_2 + \epsilon$$

where x_1, x_2 are independently $\mathcal{N}(0, 5^2)$ and ϵ is drawn from a standard normal distribution. Calculate and report the following error measures: (2 points)

- In-sample (training) MSE
- In-sample Bayes MSE (or the irreducible error): *i.e.*, the prediction error you would get from the *true* solution
- Out-of-sample (test) MSE on 25 new test points
- Out-of-sample Bayes MSE on the same 25 test points.

(c) Repeat the above process a large number of times to get the expected values of the four errors. What are they for this problem and how do they change as you vary n? (3 points)

(d) Repeat this process a large number of times and show that OLS is indeed unbiased for this linear model (2 points)

(e) Using the MSE decomposition above and the fact that OLS is unbiased for linear models, what is the variance and what is the irreducible noise for this problem? (2 points)

(f) Repeat the above analysis for the non-linear function:

$$y = \text{sign}(x_1) * x_1^2 + \cos(x_2) + \epsilon$$

What is the bias, variance, and irreducible noise for this problem? How do they change with n? (4 points)

(g) Repeat the analysis again but now with $x_1, x_2 \overset{\text{IID}}{\sim} \mathcal{N}(25, 5^2)$: how do the bias, variance, and irreducible noise change? Why? (3 points)

Question 4: Fitting Least Squares Models (20 points)

In class, we derived the closed form expression for OLS coefficients: $\hat{\boldsymbol{\beta}} = \boldsymbol{X}^{\dagger}\boldsymbol{y} = (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{y}$. In this question, we will explore alternate approaches to fitting linear models.

(a) Generate 100 samples from the following data generating process (2 points):

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}_5, \boldsymbol{I}_{5\times 5})$$
$$y \sim \mathcal{N}\left(x_1 + \sqrt{x_2^2 + 5} + 0.1x_3^3 + \cos|x_4| + \frac{1}{|x_5| + 3}, 0.25\right)$$

(b) Fit a linear model to this data using the closed-form solution (not the built-in `lm` function!) and plot $\hat{y}$ against y. (3 points)

(c) *Gradient Descent* methods fit models by taking small steps in the direction of the gradient of the loss function. For classic OLS, this gradient is given by:

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$
$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = -\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

and the gradient update is given by:

$$\boldsymbol{\beta}^{(k+1)} \leftarrow \boldsymbol{\beta}^{(k)} - c * \left.\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}}$$

That is, at step-k, use your current estimate of $\boldsymbol{\beta}$, $\boldsymbol{\beta}^{(k)}$ to compute the gradient, multiply it by a small constant c,[1] and subtract it to get your new estimate $\boldsymbol{\beta}^{(k+1)}$. Repeat this process until $\boldsymbol{\beta}^{(k+1)}$ stops changing ($\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k+1)}\| < 1 \times 10^{-6}$).

Implement gradient descent for OLS and show that you get nearly the same value as the closed form solution. (5 points)

(d) Generate 100 new test data points. Repeat gradient descent, but now using the original (training) and new (test) data points to compute the training and test loss at each step. Plot the evolution of these quantities over GD iterates. Does it make sense to run GD all the way to (numerical) convergence? (5 points)

(e) Modern machine learning methods are often fit with *weight decay*. Weight decay updates the gradient descent update to:

$$\boldsymbol{\beta}^{(k+1)} \leftarrow \boldsymbol{\beta}^{(k)} - c * \left.\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}} - \omega\boldsymbol{\beta}^{(k)}$$

i.e., we subtract off a little bit extra in each update step.

Implement OLS with weight decay and show that it corresponds to the equivalent ridge regression solution, $(\boldsymbol{X}^\top\boldsymbol{X} + \omega\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$. (2 points)

(f) Use 5-fold cross validation to identify the optimal value of the weight decay parameter. (3 points)

[1] For this problem, $c < 1/200$ should suffice.