Homework 9
MATH 166 - Fall 2024
Tufts University, Department of Mathematics
Instructor: James M. Murphy
Due: November 19, 2024

1. Book Questions

Wasserman: Chapter 13: 7 (a); Chapter 20: #1, #2 (you do not need to construct the 95% confidence interval; the data is here: `http://www.stat.cmu.edu/~larry/all-of-statistics/=data/glass.dat`). Note, the link in the book for the car mileage data does not work. The correct link is here: `https://www.stat.cmu.edu/~larry/all-of-statistics/`.

Supplemental Question 1 (Ridge Regression)

Let $X \in \mathbb{R}^{n\times d}$ be a matrix of inputs (each row is a d-dimensional observation). Let $y \in \mathbb{R}^{n\times 1}$ be a vector of outputs. Recall that multivariate linear least squares regression (with no constant term) learns a function $f_{\hat{\beta}} : \mathbb{R}^d \to \mathbb{R}$, $f_{\hat{\beta}}(x) = \sum_{i=1}^d \hat{\beta}_i x_i$, where $\hat{\beta} \in \mathbb{R}^{d\times 1}$ satisfies

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^{d\times 1}}{\arg\min} \|y - X\beta\|_2^2.$$

One can use matrix calculus to show that in this case, the optimal coefficient vector is

$$\hat{\beta} = (X^\top X)^{-1} X^\top y.$$

(a) Explain why when $d > n$, it may be possible for there to exist infinitely many $\beta \in \mathbb{R}^{d\times 1}$ satisfying $\|y - X\beta\|_2^2 = 0$. Why is this a problem for linear regression?

(b) One possible solution to the issue in (a) is to consider the *ridge regression problem*, which adds *Tikhonov regularization* to the optimization problem:

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^{d\times 1}}{\arg\min} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2,$$

where $\lambda > 0$ is a tunable parameter. Use matrix calculus to show that the optimal coefficient vector in this case is

$$\hat{\beta} = (X^\top X + \lambda I)^{-1} X^\top y.$$

(c) Interpret the result in (b).