Tashfeen, Ahmad
CSCI 6213: Data Science Fundamentals

Homework 5

1. Recapitulation

We start by reminding ourselves of the variables,

$$x_i = \begin{bmatrix} t_1 \\ \vdots \\ t_p \end{bmatrix}, \quad X_{(n,p)} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}, \quad \beta_{(p,1)} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad y_{(n,1)} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

In the previous assignment, we differentiated the error function and minimised the derivative to get β,

$$\beta = (X^T X)^{-1} X^T y \quad \text{when} \quad \frac{\partial (X\beta - y)^T (X\beta - y)}{\partial \beta} = 0.$$

In this assignment, we'll explore β from a more Bayesian point of view.

2. Maximum a Posteriori Probability Technique Review

In Bayesian statistics, a *maximum a posteriori probability* (MAP) estimate is an estimate of an unknown quantity that equals the mode of the posterior distribution. Let's say that this distribution has a distribution parameter ζ and (like the normal distribution) the mode, mean and the maximum for this distribution are equal. Fix a new sample point x from the sample's distribution f.

$$x \sim f(\zeta)$$

We may now observe the posterior probability $\mathbf{P}(\zeta \mid x)$. The idea is that we should reconsider our beliefs of the distribution parameter in the light of new evidence. The evidence being the new sample point x. Suppose that we had only the two choices of $\zeta \in \{\zeta_i, \zeta_j\}$ and post new sample point x we also know that,

$$\mathbf{P}(\zeta_j \mid x) > \mathbf{P}(\zeta_i \mid x)$$

Then we should let the new $\zeta = \zeta_j$. More generally,

$$\zeta = \max_i \mathbf{P}(\zeta_i \mid x)$$

To calculate $\mathbf{P}(\zeta_j \mid x)$, we use the Bayes' Theorem,

$$\mathbf{P}(\zeta_i \mid x) = \frac{\mathbf{P}(x \mid \zeta_i)\mathbf{P}(\zeta_i)}{\mathbf{P}(x)} \quad \text{therefore} \quad \zeta = \max_i \mathbf{P}(\zeta_i \mid x) = \max_i \frac{\mathbf{P}(x \mid \zeta_i)\mathbf{P}(\zeta_i)}{\mathbf{P}(x)} = \max_i \mathbf{P}(x \mid \zeta_i)\mathbf{P}(\zeta_i)$$

The above also allows/requires us to incorporate a prior distribution of ζ since we need it to calculate $\mathbf{P}(\zeta_i)$. Note that x and ζ are not required to be from the same type of distribution. If they are not then $\mathbf{P}(x \mid \zeta_i)$ and $\mathbf{P}(\zeta_i)$ maybe calculated differently, e. g., have different probability density functions.

3. MAP Estimate to Solve Ridge Regression

We can drive the normal equation with a regularisation term also known as the solution to the Ridge regression using an MAP estimate of the observed data. We proceed with a linear model having Gaussian noise/error since the data came from the real world.

$$y = x^T \beta + \varepsilon$$

Most of the time, noise ε is modelled as a Gaussian random variable.

$$\varepsilon \sim \mathcal{N}(0, c^2), \quad \text{for} \quad c \in \mathbb{R}$$

This is justified since noise maybe caused by numerous little factors, all adding up to a Gaussian due to the central limit theorem. We further assume $\beta \in \mathbb{R}^p$ to have a prior Gaussian distribution centred around 0 with a variance of $\tau^2 \mathbb{I}_p : \beta \sim \mathcal{N}(0, \tau^2 \mathbb{I}_p)$. Due to the above modelling of noise and β, we know that y are also Gaussian and with expected values as a linear function of x. We state our assumptions on the data:

Normal Distribution: All y_i are normally distributed: $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$.
Variance: We assume that all y_i have the same variance σ^2.
Linearity: We assumed that there is indeed a linear relationship between $x_i^T \in X_{(n,p)}$ and $y_i \in \mathbb{R}$. This is mathematically expressed as: the expected or mean value of y_i is a linear function of x_i,

$$\mathbb{E}[y_i|x_i] = \mu_i = x_i^T \beta.$$

Independent & Identically Distributed (IID): We assume that all y_i are independent and are from the same type of distribution.

State the problem of finding the best β as finding the most probable *posterior* $\mathbf{P}(\beta \mid y_i)$,

$$\beta = \max_\beta \mathbf{P}(\beta \mid y_i) \quad \text{and due to the Bayes' Theorem,} \quad \mathbf{P}(\beta \mid y_i) = \frac{\mathbf{P}(y_i \mid \beta)\mathbf{P}(\beta)}{\mathbf{P}(y_i)}$$

By the IID assumption,

$$\begin{aligned}
\mathbf{P}(y \mid \beta) &= \prod_{i=1}^{n} \mathbf{P}(y_i|\beta) \\
&= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x_i^T\beta - y_i)^2\right) && \text{PDF of } \mathcal{N}(\mu_i, \sigma^2) \\
\max_\beta \mathbf{P}(\beta \mid y) &= \max_\beta \frac{\mathbf{P}(y \mid \beta)\mathbf{P}(\beta)}{\mathbf{P}(y)} \\
&= \max_\beta \frac{\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x_i^T\beta - y_i)^2\right)\mathbf{P}(\beta)}{\mathbf{P}(y)} \\
&= \max_\beta \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x_i^T\beta - y_i)^2\right)\mathbf{P}(\beta) \\
&= \max_\beta \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x_i^T\beta - y_i)^2\right) \frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{1}{2\tau^2}(0 - \beta)^2\right) \\
&= \max_\beta \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x_i^T\beta - y_i)^2\right) \frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{1}{2\tau^2}\beta^2\right) \\
&= \max_\beta \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(x_i^T\beta - y_i)^2\right) \exp\left(-\frac{1}{2\tau^2}\beta^2\right) \\
&= \max_\beta \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(x_i^T\beta - y_i)^2 - \frac{1}{2\tau^2}\beta^2\right) \\
&= \min_\beta \prod_{i=1}^{n} \exp\left(\frac{1}{2\sigma^2}(x_i^T\beta - y_i)^2 + \frac{1}{2\tau^2}\beta^2\right) \\
&= \min_\beta \prod_{i=1}^{n} \exp\left((x_i^T\beta - y_i)^2 + \frac{2\sigma^2}{2\tau^2}\beta^2\right) \\
&= \min_\beta \prod_{i=1}^{n} \exp\left((x_i^T\beta - y_i)^2 + \lambda\beta^2\right) && \lambda = \frac{\sigma^2}{\tau^2} \\
&= \min_\beta \sum_{i=1}^{n} \ln\exp\left((x_i^T\beta - y_i)^2 + \lambda\beta^2\right) \\
&= \min_\beta \sum_{i=1}^{n} (x_i^T\beta - y_i)^2 + \lambda\beta^2 \\
&= \min_\beta (X\beta - y)^T(X\beta - y) + \lambda\beta^T\beta
\end{aligned}$$

At this point, in order to minimise the above, we differentiate $(X\beta - y)^T(X\beta - y) + \lambda\beta^T\beta$ and set it equal to zero. From the previous assignment, we know that,

$$\frac{\partial (X\beta - y)^T(X\beta - y)}{\partial \beta} = 2X^TX\beta - 2X^Ty \quad \text{and} \quad \frac{\partial \lambda\beta^T\beta}{\partial \beta} = 2\lambda\beta$$

Therefore, the derivative equal to zero is $2X^TX\beta - 2X^Ty + 2\lambda\beta = 0$,

Question 1. Let $\mathbb{I}$ be a p by p identity matrix. Solve $2X^TX\beta - 2X^Ty + 2\lambda\beta = 0$ for β as the new linear model. Show your work.

Question 2. Using the same two columns you picked in the last assignment from Boston house–price data, draw a (appropriately labelled) plot using the equation you got in question 1. The plot should graph integers $-10 < \lambda < 20$ on the x–axis and the model's root mean squared error on the y–axis. Use a file called `ridge.py` for this question.

Question 3. From your graph, what is the best value of the hyper-parameter λ that corresponds to the smallest root mean squared error? Give both the best $-10 < \lambda < 20$ and the corresponding error.

Submission Instructions

1) Submit a PDF that answers the questions and contains all the plots that the assignment asks for.
2) Submit your `ridge.py`.

Computer Science, Petree College of Arts & Sciences, Oklahoma City University