HW 1

Karl Hiner

September 3, 2023

1 Linear Regression

1.1 a

$$\hat{w} = (X^T X)^{-1} X^T Y \tag{1}$$

$$Y^i = w^T X^i + \varepsilon^i, \tag{2}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $w \in \mathbb{R}^d$, and $\{X^i, Y^i\}$ is the i-th data point, with $1 \leq i \leq m$.

Using the normal equation (Eqn. 1), and the model (Eqn. 2), derive the expectation $\mathbb{E}[\hat{w}]$. Note that here X is fixed, and only Y is random.

$$\begin{aligned}
\mathbb{E}[\hat{w}] &= \mathbb{E}\big[(X^T X)^{-1} X^T Y\big] && \text{Eqn. 1} \\
&= \mathbb{E}\big[(X^T X)^{-1} X^T (Xw + \varepsilon)\big] && \text{Substitute } Y \\
&= \mathbb{E}\big[(X^T X)^{-1} X^T X w + (X^T X)^{-1} X^T \varepsilon\big] && \text{Distribute} \\
&= \mathbb{E}\big[w + (X^T X)^{-1} X^T \varepsilon\big] && \text{Simplify} \\
&= w + (X^T X)^{-1} X^T \mathbb{E}[\varepsilon] && \text{Linearity of expectation} \\
&= w && \text{Since } \mathbb{E}[\varepsilon] = 0
\end{aligned}$$

1.2 b

Similarly, derive the variance $\mathrm{Var}[\hat{w}]$.

$$\begin{aligned}
\mathrm{Var}[\hat{w}] &= \mathrm{Var}\big[(X^T X)^{-1} X^T (Xw + \varepsilon)\big] && \text{Eqn. 1, Substitute } Y \\
&= \mathrm{Var}\big[(X^T X)^{-1} X^T X w + (X^T X)^{-1} X^T \varepsilon\big] && \text{Distribute} \\
&= \mathrm{Var}\big[w + (X^T X)^{-1} X^T \varepsilon\big] && \text{Simplify} \\
&= \mathrm{Var}\big[(X^T X)^{-1} X^T \varepsilon\big] && \text{Since } w \text{ is constant} \\
&= (X^T X)^{-1} X^T \mathrm{Var}[\varepsilon] X (X^T X)^{-1} && \mathrm{Var}[\boldsymbol{b}^T \boldsymbol{X}] = \boldsymbol{b}^T \mathrm{Var}[\boldsymbol{X}] \boldsymbol{b} \\
&= (X^T X)^{-1} X^T (\sigma^2 \boldsymbol{I}_M) X (X^T X)^{-1} && \mathrm{Var}[\varepsilon] \triangleq \sigma^2 \\
&= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} && \text{Commute } \sigma^2 \\
&= \sigma^2 (X^T X)^{-1} && \text{Simplify}
\end{aligned}$$

1.3 c

Under the white noise assumption above, does $\hat{w}$ follow a Gaussian distribution with mean and variance in (a) and (b), respectively? Why or why not?

Answer: Yes, $\hat{w}$ follows a Gaussian distribution with mean and variance in (a) and (b), respectively. This is because $\hat{w}$ is a linear combination of the random variables ε_i, which are Gaussian by assumption. Since $\hat{w}$ is a linear combination of Gaussian random variables, it is itself Gaussian, with $\hat{w} \sim \mathcal{N}(w, \sigma^2(X^T X)^{-1})$.

1.4 d: Weighted linear regression

Suppose we keep the independence assumption but remove the same variance assumption. In other words, data points would be still sampled independently, but now they may have different variance σ_i. Thus, the variance (the covariance matrix) of ε would still be diagonal, but with different values:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_m^2 \end{bmatrix}$$

Derive the estimator $\hat{w}$ (similar to the normal equations) for this problem using matrix-vector notations with Σ.

Answer:

We want to minimize

$$\arg\min_w \sum_{i=1}^{m} \frac{1}{\sigma_i^2}(y_i - w^T x_i)^2.$$

In matrix-vector notation, this is equivalent to

$$\arg\min_w (Y - Xw)^T \Sigma^{-1} (Y - Xw).$$

Taking the derivative with respect to w and setting it to zero:

$$\begin{aligned} \frac{\partial}{\partial w}\big((Y - Xw)^T \Sigma^{-1}(Y - Xw)\big) &= 0 \\ \frac{\partial}{\partial w}\big(w^T X^T \Sigma^{-1} X w - 2w^T X^T \Sigma^{-1} Y + Y^T \Sigma^{-1} Y\big) &= 0 \\ -2X^T \Sigma^{-1} Y + 2X^T \Sigma^{-1} X w &= 0 \\ (X^T \Sigma^{-1} X) w &= X^T \Sigma^{-1} Y \end{aligned}$$

Thus, the weighted least squares estimator $\hat{w}$ is:

$$\hat{w} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$$

2 Ridge Regression

For linear regression, it is often assumed that $y_i = w^T x_i + \varepsilon$, where $w, x \in \mathbb{R}^d$ by absorbing the constant term (bias) in an affine hypothesis into w, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian random variable. Given m i.i.d. samples $z_i = (x_i, y_i), 1 \leq i \leq m$, we define $Y = (y_1, \cdots, y_m)^T$ and

$X = (x_1, \cdots, x_m)^T$. Thus, we have $Y \sim \mathcal{N}(Xw, \sigma^2 I_m)$. Show that the ridge regression estimate is the mean of the posterior distribution under a Gaussian prior $w \sim \mathcal{N}(0, \tau^2 I)$. Find the explicit relation between the regularization parameter λ in the ridge regression estimate of the parameter w, and the variances σ^2 and τ^2.

Answer:

The ridge regression estimate is defined as

$$\begin{aligned}\hat{w}^{\text{Ridge}} &\triangleq \arg\min_w \sum_{i=1}^{m} (w^T x_i - y_i)^2 + \lambda \|w\|^2 \\ &= \arg\min_w \|Xw - Y\|^2 + \lambda \|w\|^2.\end{aligned}$$

Taking the derivative with respect to w and setting it to zero results in the following expression for $\hat{w}^{\text{Ridge}}$ (as derived in class):

$$\hat{w}^{\text{Ridge}} = (X^T X + \lambda I)^{-1} X^T Y$$

Now, we'll show that this estimator is also the mean of the posterior distribution of w when we assume a Gaussian prior $w \sim \mathcal{N}(0, \tau^2 I)$.

The posterior distribution of w is proportional to the product of the likelihood and the prior:

$$\begin{aligned}p(w|Y) &\propto p(Y|w)p(w) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\|Y - Xw\|^2\right) \exp\left(-\frac{1}{2\tau^2}\|w - 0\|^2\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\|Xw - Y\|^2\right) \exp\left(-\frac{1}{2\tau^2}\|w\|^2\right)\end{aligned}$$

(We neglect the normalization constant $P(Y)$ since it does not depend on w. We also neglect the Gaussian constant normalizing factors since they will not change the location of the mode.)

We want to find the mean of this posterior distribution. Since multiplying two Gaussian PDFs results in another Gaussian PDF, and since the mode of a Gaussian PDF is the mean, we can find the mean of this posterior by taking its negative and minimizing it with respect to w:

$$\begin{aligned}-log(p(w|Y)) &= -log\left(\exp\left(-\frac{1}{2\sigma^2}\|Xw - Y\|^2\right) \exp\left(-\frac{1}{2\tau^2}\|w\|^2\right)\right) \\ &= \frac{1}{2\sigma^2}\|Xw - Y\|^2 + \frac{1}{2\tau^2}\|w\|^2 \\ 0 &= \frac{\partial}{\partial w}\left(\frac{1}{2\sigma^2}\|Xw - Y\|^2 + \frac{1}{2\tau^2}\|w\|^2\right) \\ 0 &= \frac{1}{\sigma^2}X^T(Xw - Y) + \frac{1}{\tau^2}w \\ 0 &= \frac{1}{\sigma^2}X^T Xw - \frac{1}{\sigma^2}X^T Y + \frac{1}{\tau^2}w \\ \hat{w}^{\text{Mean}} &= \left(\frac{1}{\sigma^2}X^T X + \frac{1}{\tau^2}I\right)^{-1} \frac{1}{\sigma^2}X^T Y\end{aligned}$$

Now, we can compare $\hat{w}^{\text{Mean}}$ to $\hat{w}^{\text{Ridge}}$:

$$\hat{w}^{\text{Mean}} \stackrel{?}{=} \hat{w}^{\text{Ridge}}$$

$$\left(\frac{1}{\sigma^2}X^TX + \frac{1}{\tau^2}I\right)^{-1}\frac{1}{\sigma^2}X^TY \stackrel{?}{=} (X^TX + \lambda I)^{-1}X^TY$$

These two expressions are equal if we set $\lambda = \frac{\sigma^2}{\tau^2}$.