2. In the real world, we encounter functions where the relationship between the inputs and their corresponding outputs is not deterministic. Assume $g(x) = x^2$, where x is a real number, meaning $x \in \mathbb{R}$. Now, assume we do not know the function $g(x)$ and aim to approximate it with a linear regression function in the form $h(x) = wx$, where $w \in \mathbb{R}$. The goal is to predict the output corresponding to the same inputs x as closely as possible using the least-squares approach.

   Additionally, for simplicity, assume that x is sampled uniformly from the interval $[-1, 1]$. Show that the input x has a uniform distribution over this interval and that its outputs can be assumed independent of x (for simplicity, you can assume $x \neq 0$).

   Let the noise be $y = g(x) + \nu$, where ν is noise. Given the explanations, answer the following questions:

   (a) Write a brief explanation of why we expect this model to have high bias.

   (b) Calculate the bias of the model $h(z)$ in terms of z (Hint: Start by finding the value of w in the least-squares method).

   (c) Calculate the variance of $h(z)$ in terms of z.

   (d) For a fixed point z, show how $g(z)$ and $R(h, z)$ (the expected risk) are related to the bias and variance. Specifically, calculate the values of bias and variance at $z = 1$ and obtain the risk $R(h, z)$.

   Solution a. Using a linear fit for a quadratic function, which we know is quadratic, introduces more complexity and will result in a large bias.

   b. The value of w using the least squares method is given by $w = X^\dagger y$, where here X is a 1×1 matrix. Thus, we have:

$$X^\dagger = (X^\top X)^{-1} X^\top$$

$$w = X^\dagger y = (X^\top X)^{-1} X^\top y = \frac{x}{x^2} \times (2x^2) = 2x$$

$$h(z) = wz = 2xz$$

$$\text{bias}(h(z)) = \mathbb{E}[h(z)] - g(z) = \mathbb{E}[2xz] - 2z^2 = 2z\mathbb{E}[x] - 2z^2 = 2z \int_{-1}^{1} x \frac{1}{2} dx - 2z^2 = -2z^2$$

   c. The variance of the model $h(z)$ can be obtained using either of the two following methods:

$$\text{var}(h(z)) = \text{var}(2xz) = \mathbb{E}[4x^2 z^2] - \mathbb{E}[2xz]^2$$

$$= \int_{-1}^{1} 4x^2 z^2 \frac{1}{2} dx - 4z^2 \mathbb{E}[x]^2$$

$$= \frac{2}{3} x^3 z^2 |_{-1}^{1} - 0 = \frac{4}{3} z^2$$

$$\text{var}(h(z)) = \text{var}(2xz) = 2z^2\text{var}(x) = 4z^2\mathbb{E}[(x - \mathbb{E}[x])^2] = 4z^2 \int_{-1}^{1} x^2 \frac{1}{2} = \frac{2}{3}z^2x^3|_{-1}^{1} = \frac{4}{3}z^2$$

d. Using the bias-variance decomposition and considering that noise reduction is not feasible[18], we can write the value of $R(h, z)$ as follows:

$$R(h, z) = \text{bias}(h(z))^2 + \text{var}(h(z))$$

Now, for $z = 1$, we have:

$$z = 1 \Rightarrow \begin{cases} \text{bias}(h(z)) = -2 \\ \text{var}(h(z)) = \frac{4}{3} \\ R(h, z) = (-2)^2 + \frac{4}{3} = \frac{16}{3} \end{cases}$$

1. Linear regression models are always linear with respect to the model parameters, denoted as θ, but they are not necessarily linear with respect to the inputs. Suppose n inputs are drawn from some distribution defined on $\mathcal{X} \times \mathbb{R}$, where each pair $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$. We apply a nonlinear transformation $\phi : \mathcal{X} \to \mathbb{R}^d$ that maps inputs to a higher-dimensional space. The prediction for y_i is given by the dot product:

$$\hat{y}_i = \theta^T \phi(x_i),$$

where $\theta \in \mathbb{R}^d$. The training error is defined as:

$$\hat{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta^T \phi(x_i))^2.$$

(a) Express the training error in matrix form, using the design matrix $\Phi = [\phi(x_1), \ldots, \phi(x_n)]^T$.

(b) Solve the ordinary least squares (OLS) problem for this setting, deriving the optimal θ.

(c) Under what condition does the OLS solution have a unique result? Provide a rigorous proof using linear algebra.

Solution:

(a)

$$\Phi = \begin{bmatrix} \phi(x_1)^T \\ \phi(x_2)^T \\ \vdots \\ \phi(x_n)^T \end{bmatrix} \in \mathbb{R}^{n \times d}.$$

Let $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T \in \mathbb{R}^n$. Then, the training error can be rewritten as:

$$\hat{R}(\theta) = \frac{1}{n} \|\mathbf{y} - \Phi\theta\|_2^2.$$

(b) The OLS objective function is:

$$\min_{\theta} \|\mathbf{y} - \Phi\theta\|_2^2.$$

The optimal θ is obtained by setting the gradient to zero:

$$\frac{\partial}{\partial \theta} \left(\|\mathbf{y} - \Phi\theta\|_2^2 \right) = -2\Phi^T(\mathbf{y} - \Phi\theta) = 0 \to \Phi^T\Phi\theta = \Phi^T\mathbf{y}.$$

Solving for θ:

$$\theta^* = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}.$$

(c) For θ^* to be uniquely defined, the matrix $\Phi^T\Phi \in \mathbb{R}^{d\times d}$ must be invertible. This requires:

$$\text{rank}(\Phi^T\Phi) = d.$$

We prove that $\Phi^T\Phi$ is invertible if and only if Φ has full column rank.

- Suppose Φ has full column rank, meaning its columns are linearly independent. Then, for any nonzero $\theta \in \mathbb{R}^d$, $\theta^T\Phi^T\Phi\theta = \|\Phi\theta\|_2^2 > 0$. This is because the only member of the null space of Φ is the zero vector. Since this quadratic form is always positive for nonzero θ, $\Phi^T\Phi$ is positive definite and thus invertible.
- If $\Phi^T\Phi$ is invertible, then for any $\theta \neq 0$, we must have $\Phi^T\Phi\theta = 0 \Rightarrow \theta = 0$. This implies that the columns of Φ are linearly independent, as if they were not, there would exist some θ in the null space of Φ, causing $\Phi\theta = 0$. This means that $\Phi(X)$ has full column rank.

Thus, $\Phi^T\Phi$ is invertible if and only if Φ has full column rank, i.e., $\text{rank}(\Phi) = d$.

۲. (۱۰ نمره) **رگرسیون خطی -** در رگرسیون لاسو، بردار وزن اپتیمال به صورت زیر بدست می‌آید:

$$\omega^* = argmin J_\lambda(\omega)$$

به طوری که:

$$J_\lambda = \frac{۱}{۲}||y - X\omega||_۲^۲ + \lambda||\omega||_۱$$

که در آن $X \in R^{n\times d}$ و روی داده‌ها whitening انجام داده باشیم. (یعنی $X^T X = I$)

آ) نشان دهید که عمل whitening روی داده‌ها باعث می‌شود که ویژگی‌ها از هم مستقل شوند به طوری که ω_i^* به تنهایی از i امین ویژگی نتیجه شود. برای اثبات این، ابتدا نشان دهید که J_λ می‌تواند به صورت زیر نوشته شود:

$$J_\lambda(\omega) = g(y) + \sum_{i=۱}^{d} f(X_{:,i}; y; \omega_i; \lambda)$$

که $X_{:,i}$ نشان دهنده i امین ستون ماتریس X است.

ب) اگر $\omega_i \geq ۰$ باشد، ω_i را پیدا کنید.

پ) اگر $\omega_i < 0$ باشد، ω_i را پیدا کنید.

ت) با توجه به قسمت‌های قبل، در چه شرایطی ω_i صفر می‌شود؟ این شرایط چگونه قابل اعمال است؟

ث) همانطور که می‌دانید، در رگرسیون ریج، عبارت نرمال‌سازی در تابع هزینه به صورت $\frac{1}{2}\lambda||\omega||_2^2$ ظاهر می‌شود. در این حالت، چه زمانی ω_i صفر می‌شود؟ تفاوت این حالت و حالت قبلی چیست؟

② الف)

$$J_\lambda(w) = \frac{1}{2}\|y - Xw\|_2^2 + \lambda\|w\|_1 = \frac{1}{2}(y - Xw)^T(y - Xw) + \lambda\|w\|_1$$

$$= \frac{1}{2}y^Ty - y^TXw + \frac{1}{2}w^T\cancel{X^TX}^{\,I}w + \lambda\|w\|_1 = \frac{1}{2}\|y\|_2^2 - y^TXw + \frac{1}{2}\|w\|_2^2 + \lambda\|w\|_1$$

که اگر قرار دهیم $g(y) = \frac{1}{2}\|y\|_2^2$ و نیز $f(X_{:,i}; y; w_i; \lambda) = -y^TX_{:,i}w_i + \frac{1}{2}w_i^2 + \lambda|w_i|$

آنگاه داریم که: $J_\lambda(w) = g(y) + \sum_{i=1}^{d} f(X_{:,i}; y; w_i; \lambda)$

ب) هدف ما کمینه کردن $J_\lambda(w)$ است. بنابراین داریم که جایی این مقدار کمینه می‌شود که در آن صفر شود (چون تابع محدب است). با توجه به $w_i > 0$ داریم:

$$\frac{\partial J_\lambda}{\partial w_i} = -y^TX_{:,i} + w_i + \lambda = 0 \Rightarrow w_i^* = y^TX_{:,i} - \lambda$$

و چون $w_i > 0$ بود، بنابراین باید داشته باشیم که $y^TX_{:,i} \geq \lambda$ شود.

پ) مشابه قسمت قبل برای $w_i < 0$ داریم که $\lambda|w_i| = -\lambda w_i$ است و داریم:

$$\frac{\partial J_\lambda}{\partial w_i} = -y^TX_{:,i} + w_i - \lambda = 0 \Rightarrow w_i^* = y^TX_{:,i} + \lambda$$

اما چون $w_i < 0$ بود، پس باید $y^TX_{:,i} < -\lambda$ باشد.

ت) با توجه به نتایج دو قسمت قبل نتیجه می‌گیریم که برای مقادیر λ که بین $-y^TX_{:,i}$ و $y^TX_{:,i}$ قرار گیرد، آنگاه مقدار w_i در هیچ‌یک از دو حالت قبل صادق نیست و بنابراین $w_i = 0$ خواهد بود. بنابراین مقادیری که $|y^TX_{:,i}| \leq \lambda$ است، آنگاه $w_i = 0$ خواهد بود. بنابراین اگر بخواهیم که وزنی صفر شود کافی است λ را طوری تعیین کنیم که $|y^TX_{:,i}| \leq \lambda$ برقرار شود.

ث) قید مسئله را برای Ridge بازنویسی می‌کنیم:

$$J_\lambda^r(w) = \frac{1}{2}y^Ty - y^TXw + \frac{1}{2}w^Tw + \frac{\lambda}{2}w^Tw$$

$$f(X_{:,i}; y; w_i; \lambda) = -y^TX_{:,i}w_i + \left(\frac{\lambda+1}{2}\right)w_i^2$$

بنابراین داریم:

$$\frac{\partial J_\lambda^r}{\partial w_i} = -y^TX_{:,i} + (\lambda+1)w_i = 0 \Rightarrow w_i^* = \frac{y^TX_{:,i}}{\lambda+1} \Rightarrow \boxed{w_i = 0 \text{ if } y^T \perp X_{:,i}}$$

از طرفی توجه داریم که $y^TX_{:,i}$ همان جواب بدون اضافه کردن قید هموارسازی است. بنابراین می‌توان این دو روش را بدین صورت تعبیر نمود که در لاسو همان جواب ساده است که یک شیفت داده شده است، و در بازه‌ی خاصی نیز مقادیر را صفر می‌کند؛ بنابراین لاسو برای مسائلی که نیاز به صفر کردن برخی پارامترها داریم قابل استفاده است. از طرفی Ridge عملاً پارامترها را اسکیل می‌کند و آن‌ها را بدین صورت کوچک و نرمال می‌کند و در این حالت وزن‌ها را با اسکیل کردن کوچک و نرمال می‌شوند و ضرایب را دقیقاً صفر نمی‌کند و نهایتاً می‌تواند آنها را به صفر نزدیک کند.

۱) یک سکه داریم که احتمال شیر آمدن آن، p، نامشخص است. برای تخمین p سکه را n بار پرتاب کرده‌ایم که در نتیجه r بار شیر آمده است. برآوردگر بیش‌ترین درست‌نمایی (MLE) برای تخمین p را به دست آورید و اریبی و واریانس آن را محاسبه کنید.

۵) در مساله‌ی رگرسیون خطی با n داده مشاهده شده و p متغیر کمکی (covariate)، فرض کنید $\mathbf{X}_{n\times p}$ ماتریس متغیرهای کمکی مشاهده شده، $\mathbf{y}_{n\times 1}$ بردار متغیرهای پاسخ مشاهده شده و $\beta_{p\times 1}$ بردار ضرایب رگرسیون باشد. طبق مدل احتمالاتی رگرسیون داریم $\mathbf{y} = \mathbf{X}\beta + \epsilon$ که $\epsilon \sim \mathcal{N}(0,\sigma^2 I)$. اگر $n > p$، ماتریس $\mathbf{X}$ رتبه‌ی کامل داشته باشد و $\hat{\beta}$ تخمین‌گر بیش‌ترین درست‌نمایی باشد و قرار دهیم $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$،

الف) نشان دهید $\hat{\mathbf{y}} - \mathbf{y} = [\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T - I]\epsilon$.

ب) نشان دهید $\mathbb{E}[\|\hat{\mathbf{y}} - \mathbf{y}\|^2]/n = \sigma^2(n-p)/n$.

پ) از دو قسمت قبل درباره‌ی خطای آموزش و بیش‌برازش چه می‌فهمیم؟

۸) از تعدادی مهندس خواسته شده که محیط یک بیضی را اندازه بگیرند و تخمین‌های $x_1, \cdots, x_n$ به دست آمده است. تخمین‌ها مستقل از هم و نااریب هستند. فرض کنید مقدار واقعی محیط برابر μ باشد و $x_i \sim \mathcal{N}(\mu, \sigma_i^2)$. چون مهندس‌ها از ابزارهای مختلفی استفاده می‌کنند σ_iها لزوما با هم برابر نیستند و مقدارشان هم نامشخص است. هدف به دست آوردن تخمین تا جای ممکن بهتری برای μ است.

الف) اگر σ_iها را بدانیم چه تخمین‌گری برای μ پیشنهاد می‌کنید؟ چرا؟

ب) تابع درست‌نمایی را بنویسید و برای به دست آوردن برآوردگر بیش‌ترین درست‌نمایی تلاش کنید.

پ) به روش بیزی عمل کنید: λ_i را معکوس σ_i^2 بگیرید. در توزیع پیشین فرض کنید μ و λ_iها مستقل‌اند و $\lambda_1, \cdots, \lambda_n \sim \Gamma(a, b)$. (تابع چگالی توزیع گاما به شکل $g(\lambda) = c\lambda^{a-1}e^{-b\lambda}$ است) هم‌چنین توزیع پیشین μ را گاوسی با میانگین صفر و واریانس بی‌نهایت (واریانسی که به بی‌نهایت میل می‌کند) بگیرید. توزیع پسین را محاسبه کنید و برآوردگر بیشترین احتمال پسین را به دست آورید.