Intro to Big Data Science: Assignment 2 Reference Answer

TAN 12212523@mail.sustech.edu.cn

April 16, 2025

Exercise1 (Maximum Likelihood Estimate)

1. The likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\right).$$

Taking the natural logarithm, we obtain the log-likelihood

$$l(\theta) = \ln(L(\theta)) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}.$$

To get the MLE estimators for (μ, σ^2), we need to differentiate $l(\theta)$ with respect to μ and σ^2 and let them equal to zeros, which means

$$\frac{\partial l(\theta)}{\partial \mu} = \frac{\sum_{i=1}^n (x_i-\mu)}{\sigma^2} = 0, \quad \frac{\partial l(\theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^4} = 0.$$

It follows that

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2.$$

2. The expectation of $\hat{\mu}$ is unbiased that

$$E(\hat{\mu}) = E\left(\frac{1}{n}\sum_{i=1}^n x_i\right) = \frac{1}{n}\sum_{i=1}^n E(x_i) = \frac{1}{n}\sum_{i=1}^n \mu = \mu.$$

Let $\delta_i \triangleq \mu - x_i$. Expanding $\hat{\sigma}^2$ gives:

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^n \left(x_i - \frac{1}{n}\sum_{j=1}^n x_j\right)^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n x_i x_j, \\
&= \frac{1}{n}\sum_{i=1}^n (\mu-\delta_i)^2 - \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n (\mu-\delta_i)(\mu-\delta_j), \\
&= \left[\mu^2 - \frac{2\mu}{n}\sum_{i=1}^n \delta_i + \frac{1}{n}\sum_{i=1}^n \delta_i^2\right] - \left[\mu^2 - \frac{\mu}{n^2}\sum_{i=1}^n\sum_{j=1}^n (\delta_i+\delta_j) + \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \delta_i\delta_j\right].
\end{aligned}$$

Using the fact that $E(\delta_i) = 0$, $E(\delta_i^2) = \sigma^2$, and $E(\delta_i \delta_j) = 0$ for $i \neq j$ (independence), we get

$$
\begin{aligned}
& E(\hat{\sigma}^2) \\
=& \left[\mu^2 - \frac{2\mu}{n}\sum_{i=1}^{n} 0 + \frac{1}{n}\sum_{i=1}^{n}\sigma^2\right] - \left[\mu^2 - \frac{\mu}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(0+0) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} 0 + \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2\right] \\
=& \left[\mu^2 + \frac{n\sigma^2}{n}\right] - \left[\mu^2 + \frac{\sigma^2}{n}\right] \\
=& \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2,
\end{aligned}
$$

which implies that

$$E\left(\frac{n}{n-1}\hat{\sigma}^2\right) = \frac{n}{n-1}E(\hat{\sigma}^2) = \frac{n}{n-1} \cdot \frac{n-1}{n}\sigma^2 = \sigma^2.$$

Therefore, $\hat{\mu}$ is an unbiased estimator of μ, but $\hat{\sigma}^2$ is a biased estimator of σ^2.

□

Exercise 2 (Linear regression)

1. We want to find w_0, just need to minimize $\sum_{i=1}^{n}(y_i - w_0)^2$. The solution is the sample mean

$$w_0 = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{1}$$

Plug the data into (1), we have

$$w_0 = \bar{y} = \frac{1 + (-1) + 1}{3} = \frac{1}{3}$$

2. We want to find w_1, just need to minimize $\sum_{i=1}^{n}(y_i - w_1 x_i)^2$, which means

$$w_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}. \tag{2}$$

Plug the data into (2), we have

$$w_1 = \frac{(-1)(1) + 0(-1) + 2(1)}{(-1)^2 + 0^2 + 2^2} = \frac{1}{5}$$

3. To find w_0, w_1 in $y_i = w_0 + w_1 x_i + \epsilon_i$, we could design matrix and response

$$X = \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 2 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}.$$

By minimizing the total residual sum-of-product, we obtain that the minimizer $\hat{w}$ satisfies

$$\hat{w} = (X^T X)^{-1} X^T y \tag{3}$$

Plug the data into (3) ($X^T X = \begin{bmatrix} 3 & 1 \\ 1 & 5 \end{bmatrix}, \quad (X^T X)^{-1} = \frac{1}{14}\begin{bmatrix} 5 & -1 \\ -1 & 3 \end{bmatrix}, \quad X^T y = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$), we have

$$\hat{w} = \frac{1}{14}\begin{bmatrix} 5 & -1 \\ -1 & 3 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{14}\begin{bmatrix} 4 \\ 2 \end{bmatrix} = \begin{bmatrix} \frac{2}{7} \\ \frac{1}{7} \end{bmatrix}$$

The results are

$$w_0 = \frac{2}{7}, \quad w_1 = \frac{1}{7}$$

4. Following the reasoning in 3, we have

$$\hat{w} = (X^T X + \lambda I)^{-1} X^T y. \tag{4}$$

Plug the data into (4)

$$X^T X + \lambda I = \begin{bmatrix} 4 & 1 \\ 1 & 6 \end{bmatrix}, \quad (X^T X + \lambda I)^{-1} = \frac{1}{23} \begin{bmatrix} 6 & -1 \\ -1 & 4 \end{bmatrix}$$

$$\hat{w} = \frac{1}{23} \begin{bmatrix} 6 & -1 \\ -1 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{23} \begin{bmatrix} 5 \\ 3 \end{bmatrix}$$

The results are

$$w_0 = \frac{5}{23}, \quad w_1 = \frac{3}{23}$$

□

Exercise 3 (Properties of Linear Regression)

1. In multivariate linear problem, input X, then the corresponding output is

$$\hat{y} = Xw.$$

Then

$$RSS(w) = \|y - \hat{y}\|_2^2 = \|y - Xw\|_2^2.$$

By differentiating $RSS(w)$ with respect to w and setting it to zero, we get

$$\frac{\partial RSS(w)}{\partial w} = -2X^T(y - Xw) = 0.$$

It gives

$$\hat{w} = (X^T X)^{-1} X^T y.$$

Thus, the linear regression predictor is

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

2. By definition, we get

$$E(\hat{w}) = E\left[(X^T X)^{-1} X^T y\right] = E\left[(X^T X)^{-1} X^T (Xw + \epsilon)\right] = E\left[(X^T X)^{-1} X^T X w\right] = E[w] = w,$$

and

$$\begin{aligned} \text{Var}(\hat{w}) &= E\left[(\hat{w} - E(\hat{w}))(\hat{w} - E(\hat{w}))^T\right] = E\left[(\hat{w} - w)(\hat{w} - w)^T\right] \\ &= E\left[\left((X^T X)^{-1} X^T \epsilon\right)\left((X^T X)^{-1} X^T \epsilon\right)^T\right] \\ &= E\left[(X^T X)^{-1} X^T \epsilon \epsilon^T X (X^T X)^{-1}\right] \\ &= (X^T X)^{-1} X^T X (X^T X)^{-1} E[\epsilon \epsilon^T] \\ &= (X^T X)^{-1} \sigma^2. \end{aligned}$$

3. Since we have

$$\mathbf{P}^2 = \mathbf{P}\mathbf{P} = X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T = X(X^T X)^{-1} X^T = \mathbf{P},$$

then if (λ, x) is an eigenpair for **P**, we must have

$$\lambda x = \mathbf{P} x = \mathbf{P}^2 x = \lambda^2 x.$$

Eigenvectors are by definition nonzero, thus, $\lambda = \lambda^2$ must hold, which gives λ can only be 0 or 1, i.e., **P** has only 0 and 1 eigenvalues.

4. *Proof.* From (1.), we know that $X^T(y-\hat{y}) = X^T(y - X\hat{w}) = 0$. Since the first column of X is just 1, we know that $1^T(y-\hat{y}) = 0$. Therefore,

$$\begin{aligned} SS_{tot} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\ &= SS_{res} + SS_{reg} + 2(y-\hat{y})^T(X\hat{w} - \bar{y}1) \\ &= SS_{res} + SS_{reg} + 2\underbrace{(y-\hat{y})^T X\hat{w} - 2\bar{y}(y-\hat{y})^T 1}_{=0} \\ &= SS_{res} + SS_{reg}. \end{aligned}$$

5. $\hat{\mathbf{w}}_{\text{ridge}}$ is a **biased** estimator.

   proof: The ridge regression estimator is given by

$$\hat{\mathbf{w}}_{\text{ridge}} = (X^T X + \lambda I_d)^{-1} X^T \mathbf{y},$$

   which has expectation

$$E[\hat{\mathbf{w}}_{\text{ridge}}] = E[(X^T X + \lambda I_d)^{-1} X^T (X\mathbf{w} + \epsilon)] = E[(X^T X + \lambda I_d)^{-1} X^T X\mathbf{w}] \neq \mathbf{w}.$$

6. Following the information given in 5, we have

$$\hat{y} = X\hat{\mathbf{w}}_{\text{ridge}} = X(X^T X + \lambda I_d)^{-1} X^T \mathbf{y} = Qy.$$

   Let X have the singular value decomposition (SVD)

$$X = USV^\top, \tag{5}$$

   where $U \in \mathbb{R}^{n\times d}$ is column-orthogonal ($U^\top U = I_d$), $S = \text{diag}(\sigma_1, \ldots, \sigma_d) \in \mathbb{R}^{d\times d}$ contains singular values, $V \in \mathbb{R}^{d\times d}$ is orthogonal ($VV^\top = I_d$).

   Substitute the SVD into Q, we obtain

$$\begin{aligned} Q &= USV^\top \left(VS^2V^\top + \lambda I_d\right)^{-1} VSU^\top \\ &= US\left(S^2 + \lambda I_d\right)^{-1} SU^\top \\ &= U \cdot \frac{S^2}{S^2 + \lambda I_d} \cdot U^\top. \end{aligned} \tag{6}$$

   The k-th power of Q is

$$Q^k = (UDU^\top)^k = UD^kU^\top, \tag{7}$$

   where $D^k = \text{diag}\left(\left(\frac{\sigma_1^2}{\sigma_1^2+\lambda}\right)^k, \ldots, \left(\frac{\sigma_d^2}{\sigma_d^2+\lambda}\right)^k\right)$.

   Since $\lambda > 0$, followed by $0 < \frac{\sigma_i^2}{\sigma_i^2+\lambda} < 1$ for all σ_i, we have

$$\lim_{k\to\infty}\left(\frac{\sigma_i^2}{\sigma_i^2+\lambda}\right)^k = 0, \quad \forall i. \tag{8}$$

   Thus

$$\lim_{k\to\infty} D^k = 0 \implies \lim_{k\to\infty} Q^k = U \cdot 0 \cdot U^\top = 0. \tag{9}$$

7. When $\lambda \to 0$, it approaches OLS estimates, reduces bias but increases variance.

   When $\lambda \to \infty$, it shrinks coefficients to zero, increases bias but reduces variance.

□

Exercise 4 (Generalized Cross-Validation)

1. Let

$$f^{[k]}(w) = \sum_{i=1, i\neq k}^{n} (y_i - x_i^T w)^2 + \lambda ||w||_2^2 = ||y - Xw||_2^2 - (y_k - x_k^T w)^2 + \lambda ||w||_2^2,$$

then by differential $f^{[k]}(w)$ with respect to w and let it be zero, we get

$$\frac{\partial f^{[k]}(w)}{\partial w} = -2X^T(y - Xw) + 2x_k(y_k - x_k^T w) + 2\lambda w = 0.$$

It follows

$$(X^T X + \lambda I - x_k x_k^T)w = X^T y - x_k y_k,$$

which gives

$$\hat{w}^{[k]} = (X^T X + \lambda I - x_k x_k^T)^{-1}(X^T y - x_k y_k).$$

2. Denote $A = X^T X + \lambda I$, which is clearly nonsingular and $-x_k^T A^{-1} x_k \neq -1$ (by choosing proper λ), applying the Sherman-Morrison formula, we get

$$\begin{aligned}(X^T X + \lambda I - x_k x_k^T)^{-1} &= (A + (-x_k)x_k^T)^{-1} = A^{-1} - \frac{A^{-1}(-x_k)x_k^T A^{-1}}{1 + x_k^T A^{-1}(-x_k)} \\ &= (X^T X + \lambda I)^{-1} + \frac{(X^T X + \lambda I)^{-1} x_k x_k^T (X^T X + \lambda I)^{-1}}{1 - x_k^T (X^T X + \lambda I)^{-1} x_k}.\end{aligned}$$

Notice that

$$x_k^T (X^T X + \lambda I)^{-1} x_k = p_{kk} \quad \text{and} \quad \hat{y}_k = x_k^T (X^T X + \lambda I)^{-1} X^T y,$$

then we have

$$\begin{aligned}& x_k^T \hat{w}^{[k]} - y_k \\ =& x_k^T \left[(X^T X + \lambda I)^{-1} + \frac{(X^T X + \lambda I)^{-1} x_k x_k^T (X^T X + \lambda I)^{-1}}{1 - x_k^T (X^T X + \lambda I)^{-1} x_k} \right] (X^T y - x_k y_k) - y_k \\ =& x_k^T (X^T X + \lambda I)^{-1} X^T y - x_k^T (X^T X + \lambda I)^{-1} x_k y_k \\ & + \frac{x_k^T (X^T X + \lambda I)^{-1} x_k x_k^T (X^T X + \lambda I)^{-1} X^T y}{1 - x_k^T (X^T X + \lambda I)^{-1} x_k} \\ & - \frac{x_k^T (X^T X + \lambda I)^{-1} x_k x_k^T (X^T X + \lambda I)^{-1} x_k y_k}{1 - x_k^T (X^T X + \lambda I)^{-1} x_k} - y_k \\ =& \hat{y}_k - p_{kk} y_k + \frac{p_{kk} \hat{y}_k}{1 - p_{kk}} - \frac{p_{kk} p_{kk} y_k}{1 - p_{kk}} - y_k \\ =& \frac{\hat{y}_k - y_k}{1 - p_{kk}}.\end{aligned}$$

Hence

$$V_0(\lambda) = \frac{1}{n} \sum_{k=1}^{n} (x_k^T \hat{w}^{[k]} - y_k)^2 = \frac{1}{n} \sum_{k=1}^{n} \left(\frac{\hat{y}_k - y_k}{1 - p_{kk}} \right)^2.$$

3. We can write $V(\lambda)$ as

$$\begin{aligned}
V(\lambda) &= \frac{1}{n}\sum_{k=1}^{n} w_k\left(x_k^T\hat{x}^{[k]}-y_k\right)^2 = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{1-p_{kk}}{\frac{1}{n}\mathrm{tr}(I-P)}\right)^2\left(\frac{\hat{y}_k-y_k}{1-p_{kk}}\right)^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\left(\frac{1-p_{kk}}{\frac{1}{n}\mathrm{tr}(I-P)}\cdot\frac{\hat{y}_k-y_k}{1-p_{kk}}\right)^2 = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{\hat{y}_k-y_k}{\frac{1}{n}\mathrm{tr}(I-P)}\right)^2 \\
&= \frac{1}{n}\left(\frac{1}{\frac{1}{n}\mathrm{tr}(I-P)}\right)^2\sum_{k=1}^{n}(\hat{y}_k-y_k)^2 = \frac{1}{n}\left(\frac{1}{\frac{1}{n}\mathrm{tr}(I-\mathrm{tr}(P))}\right)^2\|\hat{y}-y\|^2 \\
&= \frac{1}{n}\left(\frac{1}{\frac{1}{n}(n-\mathrm{tr}(P))}\right)^2\|Py-y\|^2 = \frac{\frac{1}{n}\|(I-P)y\|^2}{[1-\mathrm{tr}(P)/n]^2}.
\end{aligned}$$

□