BLOCK II EXERCISES:

LINEAR REGRESSION

[SOLUTIONS]

Ex. 2

i) $L(x, y, w) = \frac{1}{2} \sum_{n=1}^{N} r_n (y_n - w^T \varphi(x_n))^2$

Matrix dimensions

$w_{D\times 1}$
$\phi(x_n)_{D\times 1}$
$\Phi_{N\times D}$
$y_{N\times 1}$
$R_{N\times N}$

* $L(x, y, w)$ can be written in matrical form

$$= \frac{1}{2}(y - \Phi w)^T R (y - \Phi w)$$

* R is a diagonal matrix with $R_{nn} = r_n$

$$= \frac{1}{2}(y^T - (\Phi w)^T) R (y - \Phi w)$$

$$= \frac{1}{2}(y^T - w^T \Phi^T) R (y - \Phi w)$$

$$= \frac{1}{2}(y^T R y - w^T \Phi^T R y - y^T R \Phi w + w^T \Phi^T R \Phi w)$$

* Notice that
$[y^T (R\Phi w)]^T = (R\Phi w)^T y = (\Phi w)^T R^T y = w^T \Phi^T R^T y$,
which happens to be a scalar

$$= \frac{1}{2}(y^T R y - 2 w^T \Phi^T R y + w^T \Phi^T R \Phi w)$$

Now we derivate this expression with respect to w

$$\frac{\partial L(w)}{\partial w} = -\frac{\partial}{\partial w}(w^T \Phi^T R y) + \frac{1}{2}\frac{\partial}{\partial w}(w^T \Phi^T R \Phi w)$$

$$= -\Phi^T R y + \Phi^T R \Phi w$$

Finally

$$\frac{\partial L(w)}{\partial w} = 0 \rightarrow \Phi^T R \Phi w = \Phi^T R y \rightarrow \boxed{w = (\Phi^T R \Phi)^{-1} \Phi^T R y}$$

ii) If we approach the problem as likelihood maximization under a Gaussian distribution, then R can be interpreted as a precision matrix, i.e. inverse of the covariance matrix.

$$\max \log N(y \mid \Phi w, R) = \max -\frac{1}{2}(y-\Phi w)^T R (y-\Phi w)$$

so r_n can be interpreted as the precision parameter of the n-th sample (x_n, y_n). ← inverse of the variance.

iii) r_n can represent the number of times (x_n, t_n) appears in our dataset.

Ex. 3

We have $f_\varepsilon(x,w) = w_0 + \sum_{i=1}^{D} w_i(x_i + \varepsilon_i) = f(w,x) + \sum_{i=1}^{D} w_i \varepsilon_i$

$\varepsilon_i \sim N(0,\sigma)$

We need to show

$$\min E_\varepsilon[L_\varepsilon(x,y,w)] = \min L(x,y,w) + \|w\|_2^2 .$$

Notation.

$\tilde{y} \equiv f_\varepsilon(x,w)$

$\hat{y} \equiv f(x,w)$

Let us define the loss with the noise

$$L_\varepsilon(x,y,w) = \frac{1}{2}\sum_{n=1}^{N}(\tilde{y}_n - y_n)^2$$

$$= \frac{1}{2}\sum_n (\tilde{y}_n^2 - 2\tilde{y}_n y_n + y_n^2)$$

$$= \frac{1}{2}\sum_n \Big(\underbrace{\hat{y}_n^2}_{(1)} + \underbrace{\Big(\sum_i w_i \varepsilon_{ni}\Big)^2}_{(2)} + 2\hat{y}_n \underbrace{\sum_i w_i \varepsilon_{ni}}_{(3)} - 2\underbrace{\hat{y}_n y_n}_{(4)} - 2y_n \underbrace{\sum_i w_i \varepsilon_{ni}}_{(5)} + \underbrace{y_n^2}_{(6)}\Big)$$

If we take the expectation with respect to ε

(3) and (5) disappear $E[\varepsilon_{ni}] = 0$

$$E_\varepsilon[L_\varepsilon(x,y,w)] = \underbrace{\frac{1}{2}\sum_n (\hat{y}_n - y_n)^2}_{(1)} + \underbrace{\frac{1}{2}E_\varepsilon\Big[\sum_n \Big(\sum_i w_i \varepsilon_{ni}\Big)^2\Big]}_{(2)}$$

$$(2)\ \Big(\sum_i w_i \varepsilon_{ni}\Big)^2 = \sum_{i=1}^{D} w_i^2 \varepsilon_{ni}^2 + 2\sum_{j=1}^{D}\sum_{k=1}^{j-1} w_j \varepsilon_{nj} w_k \varepsilon_{nk}$$

We know that $E[\varepsilon_i \varepsilon_j] = 0$ $i \neq j$, then

$$E_\varepsilon\Big[\sum_n \Big(\sum_i w_i \varepsilon_{ni}\Big)^2\Big] = \sum_i w_i^2 \sigma^2$$

$\sigma^2 = E[\varepsilon^2] + \underbrace{E[\varepsilon]^2}_{0}$

So finally

$$E_\varepsilon[L_\varepsilon(x,y,w)] = L(x,y,w) + \frac{1}{2}N\sum_i w_i^2 \sigma^2$$

Ex. 5

Recall

LSE: $\hat{w} = (X^T X)^{-1} X^T Y \rightarrow \hat{f}(\alpha) = \alpha^T \hat{w} = \hat{\theta}$ (new observation)

$\text{Var}(\hat{\theta}) = E_D[(\hat{\theta} - E_D[\hat{\theta}])^2]$

$\text{Bias}(\hat{\theta}, \theta) = E_D[\hat{\theta}] - \theta$

$D = \{X_{N\times 1}, y_{N\times 1}\}$

i)

We need to proof $\text{Var}(\hat{\theta}) \leq \text{Var}(\tilde{\theta})$ where $\tilde{\theta} = \alpha^T \tilde{w} \neq \hat{\theta}$

LSE ↑ any linear unbiased estimator

Step 0 Assumptions

True model $y_{N\times 1} = \theta + \varepsilon = Xw + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I)$

$\text{Var}(y) = \sigma^2 I$

$\text{Bias}(\hat{\theta}, \theta) = 0 \rightsquigarrow E[\hat{\theta}] = \theta$

Any other linear unbiased estimator $\rightsquigarrow \tilde{\theta} = c^T y$ (cte, training response)

Step 1 Is $\hat{\theta}$ unbiased? I.e. $E[\tilde{\theta}] = \theta$

$E[\tilde{\theta}] = E[c^T y] = E[c^T(\theta + \varepsilon)]$

$= E[c^T \theta] + E[c^T \varepsilon] = c^T \theta = c^T X w = \alpha^T w = \theta$

$\tilde{\theta}$ is unbiased only if $c^T X = \alpha^T$

Step 2 $\text{Var}(\hat{\theta}) \leq \text{Var}(\tilde{\theta})$

Trick: $\tilde{\theta} = \hat{\theta} + (\tilde{\theta} - \hat{\theta})$

$\text{Var}(\tilde{\theta}) = \cdots$

Hint: $\text{Var}(X_1 + X_2) = \sum_{i=1}^{2} \sum_{j=1}^{2} \text{Cov}(x_i, x_j)$

Ex. 6

i) $p(w) = N(0, \tau I)$ $\quad p(y|X,w) = N(Xw, \sigma^2)$

$$w^* = \underset{w}{\operatorname{argmax}}\, p(w|X,y) = \underset{w}{\operatorname{argmax}}\, p(w)p(y|X,w)$$

$$= \underset{w}{\operatorname{argmax}} \left[\log p(w) + \log p(y|X,w)\right]$$

$$= \underset{w}{\operatorname{argmax}} \left[\frac{1}{2\sigma^2}\|w\|_2^2 + \frac{1}{2\tau^2}\|y - Xw\|_2^2\right]$$

$$\times(2\sigma^2) \quad = \underset{w}{\operatorname{argmax}} \left[\|w\|_2^2 + \frac{\sigma^2}{\tau^2}\|y - Xw\|_2^2\right]$$

ii) $\lambda = \frac{\sigma^2}{\tau^2}$