Homework 1

CSCI 567: Machine Learning

Seyed Mohammad Asghari Pari

Linear Regression

Solution to Question 1.1:

From linear algebra,

Fact 1: A $m \times m$ matrix S is invertible iff $rank(S) = m$.

Fact 2: $rank(X^\intercal X) = rank(X)$.

Fact 3: For a $m \times n$ matrix S, we have $rank(S) \leq \min\{m, n\}$.

We know that our design matrix X is $N \times (D+1)$(it has N rows, $D+1$ columns) and hence, $X^\intercal X$ is a $(D+1) \times (D+1)$ matrix. Now, from Fact 1, $X^\intercal X$ is invertible iff $rank(X^\intercal X) = D+1$. Combining this result with Fact 2, we conclude that $X^\intercal X$ is invertible iff $rank(X) = D+1$. Since the dimensionality of $\boldsymbol{w}$ is $D+1$, we can say that $X^\intercal X$ is invertible iff $rank(X) =$ dimensionality of $\boldsymbol{w}$.

Hence, if $X^\intercal X$ is NOT invertible, it means $rank(X) \neq$ dimensionality of $\boldsymbol{w}$. Further from Fact 3, $rank(X) \leq$ dimensionality of $\boldsymbol{w}$. Consequently, if $X^\intercal X$ is NOT invertible, it means $rank(X) <$ dimensionality of $\boldsymbol{w}$.

In a nutshell, if $X^\intercal X$ is NOT invertible, it means $rank(X) <$ dimensionality of $\boldsymbol{w}$.

Solution to Question 1.2:

In the lecture, we found that

$$RSS(\boldsymbol{w}, b) = \sum_{n=1}^{N} [y_n - (b + \sum_{d=1}^{D} w_d x_{nd})]^2. \tag{1}$$

Now, by taking the derivative of (1) w.r.t. b, we get

$$\frac{\partial RSS(\boldsymbol{w}, b)}{\partial b} = -2 \sum_{n=1}^{N} [y_n - (b + \sum_{d=1}^{D} w_d x_{nd})]. \tag{2}$$

If we set the gradient in (2) to 0 for $b = b^*$, we get

$$2 \sum_{n=1}^{N} [y_n - (b^* + \sum_{d=1}^{D} w_d x_{nd})] = 0 \tag{3}$$

which can be written as,

$$\sum_{n=1}^{N} y_n - \sum_{n=1}^{N} b^* - \sum_{n=1}^{N} \sum_{d=1}^{D} w_d x_{nd} = 0. \tag{4}$$

Note that according to condition of question, we have

$$\frac{1}{N}\sum_{n=1}^{N} x_{nd} = 0, \quad \forall d = 1, \ldots, D \tag{5}$$

$$\implies \sum_{n=1}^{N} x_{nd} = 0, \quad \forall d = 1, \ldots, D. \tag{6}$$

Further, note that we can change the order of summation in (4) to get

$$\sum_{n=1}^{N} y_n - \sum_{n=1}^{N} b^* - \sum_{d=1}^{D}\sum_{n=1}^{N} w_d x_{nd} = 0, \tag{7}$$

$$\implies \sum_{n=1}^{N} y_n - \sum_{n=1}^{N} b^* - \sum_{d=1}^{D} w_d \sum_{n=1}^{N} x_{nd} = 0 \tag{8}$$

$$\implies \sum_{n=1}^{N} y_n - \sum_{n=1}^{N} b^* - \sum_{d=1}^{D} w_d \times 0 = 0 \tag{9}$$

$$\implies \sum_{n=1}^{N} y_n - \sum_{n=1}^{N} b^* = 0 \tag{10}$$

$$\implies b^* = \frac{1}{N}\sum_{n=1}^{N} y_n, \tag{11}$$

where (8) is correct because w_d does not depend on n and can be taken out from the summation over n. Furthermore, (9) is correct according to (6).

Logistic Regression

Solution to Question 2.1:

If we don't have access to the feature $\boldsymbol{x}$ of the data, we let the probability that a test sample is labeled as 1 be $p(y = 1) = \sigma(b)$. Considering this, the cross entropy error function can be rewritten as,

$$\mathcal{E}(b) = -\sum_{n=1}^{N} \Big[y_n \log[p(y_n = 1)] + (1 - y_n)\log[p(y_n = 0)] \Big] \tag{12}$$

$$= -\sum_{n=1}^{N} \Big[y_n \log[\sigma(b)] + (1 - y_n)\log[1 - \sigma(b)] \Big]. \tag{13}$$

Now, in order to minimize $\mathcal{E}(b)$, we take its derivative w.r.t. b and set it to zero for $b = b^*$ to get,

$$\frac{\partial \mathcal{E}(b)}{\partial b} = 0 \tag{14}$$

$$\implies -\sum_{n=1}^{N} \Big[y_n \frac{\sigma'(b^*)}{\sigma(b^*)} + (1 - y_n)\frac{-\sigma'(b^*)}{1 - \sigma(b^*)} \Big] = 0. \tag{15}$$

Note that $\sigma(b) = \dfrac{1}{1 + e^{-b}}$ and hence,

$$\sigma'(b) = \frac{e^{-b}}{(1 + e^{-b})^2} = (1 - \frac{1}{1 + e^{-b}})\frac{1}{1 + e^{-b}} = [1 - \sigma(b)]\sigma(b). \tag{16}$$

Considering (16), (15) can be further simplified as,

$$\sum_{n=1}^{N} \Big[y_n[1-\sigma(b^*)] - (1-y_n)\sigma(b^*) \Big] = 0 \tag{17}$$

$$\implies \sum_{n=1}^{N} \Big[y_n - y_n\sigma(b^*) - \sigma(b^*) + y_n\sigma(b^*) \Big] = 0, \tag{18}$$

$$\implies \sum_{n=1}^{N} \Big[y_n - \sigma(b^*) \Big] = 0 \tag{19}$$

$$\implies \sigma(b^*) = \frac{1}{N}\sum_{n=1}^{N} y_n, \tag{20}$$

$$\implies b^* = \sigma^{-1}\Big(\frac{1}{N}\sum_{n=1}^{N} y_n\Big). \tag{21}$$

If we want to find the explicit form of b^*, we can substitute $\sigma(b^*) = \dfrac{1}{1+e^{-b^*}}$ in (20) and solve for b^* to get,

$$\frac{1}{1+e^{-b^*}} = \frac{1}{N}\sum_{n=1}^{N} y_n, \tag{22}$$

$$\implies e^{-b^*} = \frac{N - \sum_{n=1}^{N} y_n}{\sum_{n=1}^{N} y_n} = \frac{\sum_{n=1}^{N}(1-y_n)}{\sum_{n=1}^{N} y_n} \tag{23}$$

$$\implies b^* = -\log\Big(\frac{\sum_{n=1}^{N}(1-y_n)}{\sum_{n=1}^{N} y_n}\Big) = \log[\sum_{n=1}^{N} y_n] - \log[\sum_{n=1}^{N}(1-y_n)]. \tag{24}$$

Therefore, we get $p(y=1) = \sigma(b^*) = \dfrac{1}{N}\sum_{n=1}^{N} y_n$.