Simon Fraser University

CMPT 726: Machine Learning, Fall 2020

Assignment #1

For
Dr. Chen, Xubo Lyu, Roshni Shaik, Ria Thomas

By
BenKun Chen (ID: 301410005 | Email: bca96@sfu.ca)

1. Linear Regression

Optimal function (2 marks)

Based on the joint probability distribution model P_{XY} for X and Y that has been given, we know that X and Y are both discrete random variables that the function given by

$$f(x, y) = P(X = x,\ Y = y) \tag{1.1}$$

For each pair of values (x_1, y_1), (x_2, y_2), (x_3, y_3), ... , (x_n, y_n) within the range of X. Since the optimal function f is linear, we know that $f(x) = mx + b$ where x is within the range of X, m is the slope and b is the y-intercept. Based on the Mean Squared Error (MSE) given by the question, we know the Squared Error (SE) is

$$SE = \sum_{n=1}^{N} (f(x_n) - y_n)^2 \tag{1.2}$$

where N is the total pair of values in P_{XY}. Expand SE we get:

$$SE = (f(x_1) - y_1)^2 + (f(x_2) - y_2)^2 + \text{......} + (f(x_N) - y_N)^2 \tag{1.3}$$

$$= ((mx_1 + b) - y_1)^2 + ((mx_2 + b) - y_2)^2 + \text{......} + ((mx_N + b) - y_N)^2 \tag{1.4}$$

$$\begin{aligned} = & ((mx_1 + b)^2 - 2y_1(mx_1 + b) + y_1^2) \\ & + ((mx_2 + b)^2 - 2y_2(mx_2 + b) + y_2^2) \\ & + \text{......} + ((mx_N + b)^2 - 2y_N(mx_N + b) + y_N^2) \end{aligned} \tag{1.5}$$

$$\begin{aligned} = & ((mx_1 + b)^2 - 2y_1(mx_1 + b) + y_1^2) \\ & + ((mx_2 + b)^2 - 2y_2(mx_2 + b) + y_2^2) \\ & + \text{......} + ((mx_N + b)^2 - 2y_N(mx_N + b) + y_N^2) \end{aligned} \tag{1.6}$$

$$\begin{aligned} = & (m^2x_1^2 + 2mx_1b + b^2 - 2y_1mx_1 - 2y_1b + y_1^2) \\ & + (m^2x_2^2 + 2mx_2b + b^2 - 2y_2mx_2 - 2y_2b + y_2^2) \\ & + \text{......} + (m^2x_N^2 + 2mx_Nb + b^2 - 2y_Nmx_N - 2y_Nb + y_N^2) \end{aligned} \tag{1.7}$$

$$\begin{aligned} = & (m^2x_1^2 + 2mx_1b + b^2 - 2y_1mx_1 - 2y_1b + y_1^2) \\ & + (m^2x_2^2 + 2mx_2b + b^2 - 2y_2mx_2 - 2y_2b + y_2^2) \\ & + \text{......} + (m^2x_N^2 + 2mx_Nb + b^2 - 2y_Nmx_N - 2y_Nb + y_N^2) \end{aligned} \tag{1.8}$$

$$\begin{aligned} = & m^2(x_1^2 + x_2^2 + \ldots + x_N^2) + 2mb(x_1 + x_2 + \ldots + x_N) + Nb^2 \\ & - 2m(x_1y_1 + x_2y_2 + \ldots + x_Ny_N) - 2b(y_1 + y_2 + \ldots + y_N) + (y_1^2 + y_2^2 + \text{......} + y_N^2) \end{aligned} \tag{1.9}$$

As we know $\overline{X} = (x_1 + x_2 + \text{......} + x_N)/N$, $\overline{Y} = (y_1 + y_2 + \text{......} + y_N)/N$, $\overline{X^2} = (x_1^2 + x_2^2 + \text{......} + x_N^2)/N$ and $\overline{Y^2} = (y_1^2 + y_2^2 + \text{......} + y_N^2)/N$ then we have $\overline{X}N = x_1 + x_2 + \text{......} + x_N$, $\overline{Y}N = y_1 + y_2 + \text{......} + y_N$, $\overline{X^2}N = x_1^2 + x_2^2 + \text{......} + x_N^2$ and $\overline{Y^2}N = y_1^2 + y_2^2 + \text{......} + y_N^2$:

$$= m^2N\overline{X^2} + 2mbN\overline{X} + Nb^2 - 2mN\overline{XY} - 2bN\overline{Y} + N\overline{Y^2} \tag{1.10}$$

As we notice the squared terms in (1.10) are positive, we can say that m and b contain terms can be looked at as equations for parabolas which open upwards. In which, these parabolas can only have minima. In order to have values for m and b which minimize the value for SE, we just need to take $\frac{\partial SE}{\partial m}$ and $\frac{\partial SE}{\partial b}$ by setting them equal to 0. By using (1.10), we get:

$$\frac{\partial SE}{\partial m} = 2mN\overline{X^2} + 2bN\overline{X} - 2N\overline{XY} = 0 \tag{1.11}$$

$$m\overline{X^2} + b\overline{X} - \overline{XY} = 0 \tag{1.12}$$

$$\frac{m\overline{X^2}}{\overline{X}} + b - \frac{\overline{XY}}{\overline{X}} = 0 \tag{1.13}$$

and,

$$\frac{\partial SE}{\partial b} = 2mN\overline{X} + 2Nb - 2N\overline{Y} = 0 \tag{1.14}$$

$$m\overline{X} + b - \overline{Y} = 0 \tag{1.15}$$

then we use (1.15) - (1.13), we get:

$$m\overline{X} + b - \overline{Y} - \frac{m\overline{X^2}}{\overline{X}} - b + \frac{\overline{XY}}{\overline{X}} = 0 \tag{1.16}$$

$$m\overline{X} - \overline{Y} - \frac{m\overline{X^2}}{\overline{X}} + \frac{\overline{XY}}{\overline{X}} = 0 \tag{1.17}$$

$$m(\overline{X})^2 - m\overline{X^2} = \overline{X}\,\overline{Y} - \overline{XY} \tag{1.18}$$

$$m = \frac{\overline{X}\,\overline{Y} - \overline{XY}}{(\overline{X})^2 - \overline{X^2}} \tag{1.19}$$

then we plug (1.19) back to (1.15), we get:

$$\frac{\overline{X}\,\overline{Y} - \overline{XY}}{(\overline{X})^2 - \overline{X^2}}\,\overline{X} + b - \overline{Y} = 0 \tag{1.20}$$

$$b = \overline{Y} - \frac{\overline{X}\,\overline{Y} - \overline{XY}}{(\overline{X})^2 - \overline{X^2}}\,\overline{X} \tag{1.21}$$

As a result, we will get the optimal function which minimizes the given Mean Squared Error by plugging (1.19) and (1.21) into $f(x) = mx + b$:

$$f(x) = \frac{\overline{X}\,\overline{Y} - \overline{XY}}{(\overline{X})^2 - \overline{X^2}} * x + \left(\overline{Y} - \frac{\overline{X}\,\overline{Y} - \overline{XY}}{(\overline{X})^2 - \overline{X^2}}\,\overline{X}\right)$$

Gaussian Noise Regression Model (3 marks)

a) As the question states that $\beta_n = \frac{1}{\sigma^2}$, we use the following Gaussian distribution:

$$N(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{2.1}$$

Then we can write $p(t|X, w, \beta) = \prod_{n=1}^{N} N(t_n|w^T\phi(x_n), \beta_n^{-1})$ into:

$$p(t|X, w, \beta) = \prod_{n=1}^{N} N(t_n|w^T\phi(x_n), \beta_n^{-1}) \tag{2.2}$$

$$p(t|X, w, \beta) = \prod_{n=1}^{N} \frac{\sqrt{\beta_n}}{\sqrt{2\pi}} exp\left\{-\frac{\beta_n}{2}(t_n - w^T\phi(x_n))^2\right\} \tag{2.3}$$

take the log: $log\,[p(t|X, w, \beta)] = log\left[\prod_{n=1}^{N} \frac{\sqrt{\beta_n}}{\sqrt{2\pi}} exp\left\{-\frac{\beta_n}{2}(t_n - w^T\phi(x_n))^2\right\}\right])$

(2.4)

Note that the log is always in the base of e in this class, we can derive the (2.4) further by using properties of logarithms:

$$log\,(p(t|X, w, \beta)) = \sum_{n=1}^{N}\left[\frac{1}{2}log(\beta_n) - \frac{1}{2}log(2\pi) - \frac{\beta_n}{2}(t_n - w^T\phi(x_n))^2\right] \tag{2.5}$$

$$log\,(p(t|X, w, \beta)) = \frac{1}{2}\sum_{n=1}^{N} log(\beta_n) - \frac{N}{2}log(2\pi) - \frac{1}{2}\sum_{n=1}^{N}\left[\beta_n(t_n - w^T\phi(x_n))^2\right] \tag{2.6}$$

b) Since we want to choose a w that maximizes the likelihood, so that:

$$w^* = arg\ max_w(\frac{1}{2}\sum_{n=1}^{N} log(\beta_n) - \frac{N}{2}log(2\pi) - \frac{1}{2}\sum_{n=1}^{N}\left[\beta_n(t_n - w^T\phi(x_n))^2\right]) \quad (2.7)$$

since β_n is just a constant, terms $\frac{1}{2}\sum_{n=1}^{N} log(\beta_n)$ and $-\frac{N}{2}log(2\pi)$ don't depend on w.

we can simplify the equation to:

$$w^* = arg\ max_w(-\frac{1}{2}\sum_{n=1}^{N}[\beta_n(t_n - w^T\phi(x_n))^2]) \quad (2.8)$$

take the negative sign out then we have:

$$w^* = arg\ min_w(\frac{1}{2}\sum_{n=1}^{N}[\beta_n(t_n - w^T\phi(x_n))^2]) \quad (2.9)$$

As we can observe in (2.9), there is an extra β_n in the log-likelihood of $p(t|X, w, \beta)$ in which the sum of squares error function doesn't have.

Weighted Linear Regression (2 marks)

a) As the question states that $y_i = x_i^T\beta + \varepsilon_i$ where ε_i are noise terms from a given distribution and we are assuming that $\varepsilon_1, \ldots, \varepsilon_N$ IID and sampled from the same zero-mean Gaussian that is, $\varepsilon_i \sim N(0, \sigma^2)$. We can write the probability density function as below:

$$p(y|x, \beta, \sigma) = \prod_{i=1}^{N} N(y_i|x_i^T\beta, \sigma_i^2) \quad (3.1)$$

Then we use (2.1) from the previous question and we get:

$$p(y|x, \beta, \sigma) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\cdot\sigma_i}exp\left\{-\frac{1}{2\sigma_i^2}(y_i - x_i^T\beta)^2\right\} \quad (3.2)$$

take the log: $log\,[p(y|x, \beta, \sigma)] = log\left[\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\cdot\sigma_i}exp\left\{-\frac{1}{2\sigma_i^2}(y_i - x_i^T\beta)^2\right\}\right]) \quad (3.3)$

Note that the log is always in the base of e in this class, we can derive the (3.3) further by using properties of logarithms:

$$log\,(p(y|x, \beta, \sigma)) = \sum_{i=1}^{N}\left[log(1) - \frac{1}{2}log(2\pi\sigma_i^2) - \frac{1}{2\sigma_i^2}(y_i - x_i^T\beta)^2\right] \quad (3.4)$$

$$log\,(p(y|x, \beta, \sigma)) = \sum_{i=1}^{N}\left[-\frac{1}{2}log(2\pi\sigma_i^2) - \frac{1}{2\sigma_i^2}(y_i - x_i^T\beta)^2\right] \quad (3.5)$$

$$log\,(p(y|x, \beta, \sigma)) = \sum_{i=1}^{N}[-\frac{1}{2}log(2\pi\sigma_i^2)] + \sum_{i=1}^{N}\left[-\frac{1}{2\sigma_i^2}(y_i - x_i^T\beta)^2\right] \quad (3.6)$$

the formula for calculating the MLE of β will then be:

$$\beta_{MLE} = arg\ max_\beta\left[\sum_{i=1}^{N}(-\frac{1}{2}log(2\pi\sigma_i^2)) + \sum_{i=1}^{N}\left(-\frac{1}{2\sigma_i^2}(y_i - x_i^T\beta)^2\right)\right] \quad (3.7)$$

b) We continue deriving (3.7). Since σ_i is just a constant, terms $\sum_{i=1}^{N}\left(-\frac{1}{2}log(2\pi\sigma_i^{2})\right)$ doesn't depend on β . we can simplify the equation to:

$$\beta_{MLE} = arg\ max_{\beta}\left[\sum_{i=1}^{N}\left(-\frac{1}{2\sigma_i^2}(y_i - x_i^{T}\beta)^2\right)\right] \tag{3.8}$$

$$\beta_{MLE} = arg\ max_{\beta}\left[-\sum_{i=1}^{N}\left(\frac{1}{2\sigma_i^2}(y_i - x_i^{T}\beta)^2\right)\right] \tag{3.9}$$

take the negative sign out then we have:

$$\beta_{MLE} = arg\ min_{\beta}\left[\sum_{i=1}^{N}\left(\frac{1}{2\sigma_i^2}(y_i - x_i^{T}\beta)^2\right)\right] \tag{3.10}$$

we then write β_{MLE} into the matrix format:

$$\beta_{MLE} = arg\ min_{\beta}\left[(y - X\beta)^{T}S(y - X\beta)\right] \tag{3.11}$$

where $S = diag(\frac{1}{2\sigma_1^2}, \frac{1}{2\sigma_2^2}, \frac{1}{2\sigma_3^2}, \dots, \frac{1}{2\sigma_N^2})$,

$$\beta_{MLE} = arg\ min_{\beta}\left[(y^{T} - \beta^{T}X^{T})(Sy - SX\beta)\right] \tag{3.12}$$

$$\beta_{MLE} = arg\ min_{\beta}\left[y^{T}Sy - y^{T}SX\beta - \beta^{T}X^{T}Sy + \beta^{T}X^{T}SX\beta\right] \tag{3.13}$$

We can observe that terms $y^{T}SX\beta$ and $\beta^{T}X^{T}Sy$ is similar. Since we know that $(y^{T}SX\beta)^{T} = y^{T}SX\beta$ because S is a diagonal matrix, we then can derive the term into:

$$(y^{T}SX\beta)^{T} = \beta^{T}X^{T}S^{T}y = \beta^{T}X^{T}S\,y \tag{3.14}$$

by plugging (3.14) back to (3.13) we get:

$$\beta_{MLE} = arg\ min_{\beta}\left[y^{T}Sy - \beta^{T}X^{T}S\,y - \beta^{T}X^{T}Sy + \beta^{T}X^{T}SX\beta\right]$$

(3.15)

$$\beta_{MLE} = arg\ min_{\beta}\left[y^{T}Sy - 2\beta^{T}X^{T}Sy + \beta^{T}X^{T}SX\beta\right] \tag{3.16}$$

in order to find the minimum value of β, we take the partial derivative on the equation based on β :

$$\frac{\partial}{\partial\beta}(\beta_{MLE}) = \frac{\partial}{\partial\beta}\left[y^{T}Sy - 2\beta^{T}X^{T}Sy + \beta^{T}X^{T}SX\beta\right] \tag{3.17}$$

$$0^{T} = -2X^{T}Sy + 2X^{T}SX\beta \tag{3.18}$$

$$2X^{T}SX\beta = 2X^{T}Sy \tag{3.19}$$

$$\beta = (2X^{T}Sy)/(2X^{T}SX) \tag{3.20}$$

As a result, the MLE of β based on the question (a) is $(2X^{T}Sy)/(2X^{T}SX)$ where $S = diag(\frac{1}{2\sigma_1^2}, \frac{1}{2\sigma_2^2}, \frac{1}{2\sigma_3^2}, \dots, \frac{1}{2\sigma_N^2})$.

2. Regularization

a) As the question given that number of features M is much larger than the number of training instances N, then $tank(X) \leq min(n, m)$. As a result, the matrix X might not be invertible if it is not a full-rank matrix. We will not be able to calculate the term $(X^T X)^{-1}$ inside the $\beta^* = (X^T X)^{-1} X^T y$.

b) We have the following equation based on the question:

$$J_R(\beta) = \sum_i^N (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^M \beta_j^2 = (X\beta - y)^T (X\beta - y) + \lambda \|\beta\|^2 \quad (4.1)$$

we will keep deriving $(X\beta - y)^T (X\beta - y) + \lambda \|\beta\|^2$ to get the value of β that minimizes (4.1):

$$J_R(\beta) = (X\beta - y)^T (X\beta - y) + \lambda \|\beta\|^2 \quad (4.2)$$

$$J_R(\beta) = (\beta^T X^T - y^T)(X\beta - y) + \lambda \beta^T \beta \quad (4.3)$$

$$J_R(\beta) = \beta^T X^T X\beta - \beta^T X^T y - y^T X\beta + y^T y + \lambda \beta^T \beta \quad (4.4)$$

in order to find the minimum value of β, we take the partial derivative on the equation based on β:

$$\frac{\partial}{\partial\beta}(J_R(\beta)) = \frac{\partial}{\partial\beta}(\beta^T X^T X\beta - \beta^T X^T y - y^T X\beta + y^T y + \lambda \beta^T \beta) \quad (4.5)$$

$$0^T = 2X^T X\beta - X^T y - y^T X + 2\lambda\beta \quad (4.6)$$

$$0^T = 2X^T X\beta - X^T y - X^T y + 2\lambda\beta \quad (4.7)$$

$$0^T = 2X^T X\beta - 2X^T y + 2\lambda\beta \quad (4.8)$$

$$0^T = 2X^T X\beta - 2X^T y + 2\lambda\beta \quad (4.9)$$

times the identity matrix to λ we get:

$$\beta(X^T X + \lambda I) = X^T y \quad (4.10)$$

$$\beta = (X^T X + \lambda I)^{-1} X^T y \quad (4.11)$$

As a result, we have derived the value of β that minimizes (4.1) and proved that the β is equal to the equation given by the question.

3. Classification

Logistic regression (2 marks)

a) First, we can assume that the probability of a cell being unaffected is μ, reversely the probability of a cell being affected will be $1 - \mu$. The odds for the cell being unaffected is then $\mu / (1 - \mu)$ since it is defined as the probability of success divided by the probability of failure. As the equation is given for the log of odds for the cells being unaffected:

$$log(odds) = -10 + 2 * x_1 \quad (5.1)$$

since we know the $odds$ for the cell being unaffected is $\mu / (1 - \mu)$ and x_1 is 6, we can derive (5.1) as:

$$log(\mu (1 - \mu)^{-1}) = -10 + 2 * 6 \quad (5.2)$$

$$log(\mu (1 - \mu)^{-1}) = 2 \quad (5.3)$$

Note that the log is always in the base of e in this class, we can derive the (5.3) further by using properties of logarithms:

$$exp\left\{log(\mu(1-\mu)^{-1})\right\} = exp\{2\} \tag{5.4}$$

$$\mu(1-\mu)^{-1} = exp\{2\} \tag{5.5}$$

$$\mu = exp\{2\} \cdot (1 + exp\{2\})^{-1} \tag{5.6}$$

$$\mu \approx 0.880797078 \tag{5.7}$$

As a result, the probability of a cell being unaffected is approximately 88.0797078% with its diameter of dye stain of 6.

b) As we know the probability of making unaffected cell (μ) is 90%, we can derive the equation as:

$$log(\mu(1-\mu)^{-1}) = -10 + 2 * x_1 \tag{5.8}$$

$$x_1 = \frac{10+log(\mu(1-\mu)^{-1})}{2} \tag{5.9}$$

$$x_1 = \frac{10+log(0.9(0.1)^{-1})}{2} \tag{5.9}$$

$$x_1 \approx 6.098612289 \tag{5.10}$$

as a result, the minimum diameter of the dye stain on a cell which secure the cell is unaffected with a probability of 90% is approximately 6.098612289.

Softmax for Multi-Class Classification (3 marks)

a) The activation function a_k for Type A, Type B and Type C are as follows:

$$a_{Type_A}(x_1, x_2) = 2x_1 + 5x_2 + 5 \tag{6.1}$$

$$a_{Type_B}(x_1, x_2) = 5x_1 + 10x_2 + 1.5 \tag{6.2}$$

$$a_{Type_C}(x_1, x_2) = 5x_1 + 2x_2 + 1 \tag{6.3}$$

b) Based on the softmax function (6) provided by the question, we can derive the class probabilities for Type A, Type B and Type C cells for a sample cell with diameter 10 and depth 2 as follow:

$$p(C_{Type_A} | x) = \frac{exp(a_{Type_A})}{exp(a_{Type_A})+exp(a_{Type_B})+exp(a_{Type_C})} \tag{6.4}$$

$$= \frac{exp(a_{Type_A})}{exp(a_{Type_A})+exp(a_{Type_B})+exp(a_{Type_C})} \tag{6.5}$$

$$= \frac{exp(2x_1+5x_2+5)}{exp(2x_1+5x_2+5)+exp(5x_1+10x_2+1.5)+exp(5x_1+2x_2+1)} \tag{6.6}$$

plug $x_1 = 10, x_2 = 2$:

$$= \frac{exp(2*10+5*2+5)}{exp(2*10+5*2+5)+exp(5*10+10*2+1.5)+exp(5*10+2*2+1)} \tag{6.7}$$

$$= \frac{exp(2x_1+5x_2+5)}{exp(2x_1+5x_2+5)+exp(5x_1+10x_2+1.5)+exp(5x_1+2x_2+1)} \tag{6.8}$$

$$\approx 1.40686162 * 10^{-16} \tag{6.9}$$

$$p(C_{Type_B} | x) = \frac{exp(a_{Type_B})}{exp(a_{Type_A})+exp(a_{Type_B})+exp(a_{Type_C})} \tag{6.10}$$

$$\approx 0.9999999317 \tag{6.11}$$

$$p(C_{Type_C} | x) = \frac{exp(a_{\ Type_C})}{exp(a_{\ Type_A})+exp(a_{\ Type_B})+exp(a_{\ Type_C})} \tag{6.12}$$

$$\approx 6.82560291 * 10^{-8} \tag{6.13}$$

Since $p(C_{Type_B} | x) < p(C_{Type_C} | x) < p(C_{Type_B} | x)$, the predicted type for the mentioned sample cell is Type B.