Assignment 1 Solutions: Linear Models Regression, Classification and Regularization

Due Oct 9 at 11:59pm

This assignment is to be done individually.

Important Note: The university policy on academic dishonesty (cheating) will be taken very seriously in this course. You may not provide or use any solution, in whole or in part, to or by another student.

You are encouraged to discuss the concepts involved in the questions with other students. If you are in doubt as to what constitutes acceptable discussion, please ask! Further, please take advantage of office hours offered by the instructor and the TA if you are having difficulties with this assignment.

DO NOT:

- Give/receive code or proofs to/from other students
- Use Google to find solutions for assignment

DO:

- Meet with other students to discuss assignment (it is best not to take any notes during such meetings, and to re-work assignment on your own)
- Use online resources (e.g. Wikipedia) to understand the concepts needed to solve the assignment.

Submitting Your Assignment

The assignment must be submitted online at `https://canvas.sfu.ca/`. You must submit the following file:

1. An assignment report in **PDF format**, named `report.pdf`. This report should contain the solutions to questions 1-3.

---

1 Linear Regression

Optimal function (2 marks)

The main goal is to minimize the given Mean Squared Error $\mathbb{E}[(f(X) - Y)^2]$

$\mathbb{E}[(f(X) - Y)^2] = \mathbb{E}[(f(X) - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - Y)^2]$

$\mathbb{E}[(f(X) - Y)^2] = \mathbb{E}[(f(X) - \mathbb{E}[Y|X])^2 + (\mathbb{E}[Y|X] - Y)^2 + 2(f(X) - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - Y)]$

$\mathbb{E}[(f(X) - Y)^2] = \mathbb{E}[(f(X) - \mathbb{E}[Y|X])^2] + (\mathbb{E}[Y|X] - Y)^2] + 2\mathbb{E}[(f(X) - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - Y)]$

Now, using the tower property of conditional expectations, the third term of the above equation(1) can be written as follows: [Tower Rule: $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$]

$\mathbb{E}_{XY}[(f(X) - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - Y)] = \mathbb{E}_X[\mathbb{E}_{Y|X}[(f(X) - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - Y)|X]]$

$\mathbb{E}_{XY}[(f(X) - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - Y)] = \mathbb{E}_X[(f(X) - \mathbb{E}[Y|X])\mathbb{E}_{Y|X}(\mathbb{E}[Y|X] - Y)|X]] = 0$

Since conditioning on X, f(X), $\mathbb{E}[Y|X]$ are constant. Therefore,

$\mathbb{E}[(f(X) - Y)^2] = \mathbb{E}[(f(X) - \mathbb{E}[Y|X])^2] + (\mathbb{E}[Y|X] - Y)^2]$

$\mathbb{E}[(f(X) - Y)^2] >= (\mathbb{E}[Y|X] - Y)^2]$ for all f [Since the first term being a square of a quantity is non negative.]

Hence, $f^* = \mathbb{E}[Y|X]$

Gaussian Noise Regression Model (3 marks)

The likelihood function we wish to estimate is:

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

The Gaussian distribution $\mathcal{N}$ for each training data point is given by:

$$\mathcal{N}(t_n|\mathbf{w}^{\mathbf{T}}\phi(\mathbf{x_n}), \beta_n^{-1}) = \sqrt{\frac{\beta_n}{2\pi}}\, e^{-\frac{\beta_n}{2}(t_n - w^T\phi(x_n))^2} \tag{2}$$

Taking log on both sides in Eqn 2, we get

$$\ln \mathcal{N}(t_n|\mathbf{w}^{\mathbf{T}}\phi(\mathbf{x_n}), \beta_n^{-1}) = \frac{1}{2}\left(\ln \beta_n - \ln 2\pi\right) - \frac{\beta_n}{2}(t_n - w^T\phi(x_n))^2 \tag{3}$$

Substituting result of Eqn 3 in Eqn 1, we obtain:

$$\ln \boldsymbol{p}(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \frac{1}{2}\sum_{n=1}^{N}(\ln \beta_n - \ln 2\pi) - \frac{1}{2}\sum_{n=1}^{N}\beta_n(t_n - w^T\phi(x_n))^2 \tag{4}$$

An interesting observation from this result is that precision estimates β_n manifest themselves as weight terms in the squared error function (second term in Eqn 7) for each basis function $\phi(x_n)$.

Weighted Linear Regression (2 marks)

(a) **Solution**: $y_i = \mathrm{x_i^T}\beta + \epsilon_i$, thus $p\left(y_i \mid \mathrm{x}_i, \beta\right) = \mathcal{N}\left(\mathrm{x_i^T}\beta, \sigma_i^2\right)$, Thus the formula for the MLE of β is:

$$\begin{aligned}\beta_{MLE} &= \arg\max_{\beta} \log \prod_i p\left(y_i \mid \mathrm{x}_i, \beta\right) \\ &= \arg\max_{\beta} \sum_i \log p\left(y_i \mid \mathrm{x}_i, \beta\right) \\ &= \arg\max_{\beta} \sum_i \log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{\left(y_i - \mathrm{x_i^T}\beta\right)^2}{2\sigma_i^2}\right)\right)\end{aligned} \tag{5}$$

(b) **Solution**: Stating from (5),

$$\begin{aligned}\beta_{MLE} &= \arg\max_{\beta} \sum_i \log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{\left(y_i - \mathrm{x_i^T}\beta\right)^2}{2\sigma_i^2}\right)\right) \\ &= \arg\max_{\beta} \sum_i \log\frac{1}{\sqrt{2\pi\sigma_i^2}} + \log\left(\exp\left(-\frac{\left(y_i - \mathrm{x_i^T}\beta\right)^2}{2\sigma_i^2}\right)\right)\end{aligned} \tag{6}$$

But first term does not involve β thus we can ignore it,

$$\begin{aligned}\beta_{MLE} &= \arg\max_{\beta} \sum_i \log\left(\exp\left(-\frac{\left(y_i - \mathrm{x_i^T}\beta\right)^2}{2\sigma_i^2}\right)\right) \\ &= \arg\max_{\beta} \sum_i -\frac{\left(y_i - \mathrm{x_i^T}\beta\right)^2}{2\sigma_i^2} \\ &= \arg\min_{\beta} \sum_i \frac{\left(y_i - \mathrm{x_i^T}\beta\right)^2}{\sigma_i^2}\end{aligned} \tag{7}$$

Note that we can remove the 2 in the denominator. Now we write the (7) in matrix notation. If we let $\mathbf{W}$ be a diagonal matrix with diagonal entry $w_{ii} = \frac{1}{\sigma_i^2}$, we get:

$$\beta_{MLE} = \arg\min_{\beta} (\mathrm{y} - \mathbf{X}\beta)^T \mathbf{W} (\mathrm{y} - \mathbf{X}\beta) \tag{8}$$

Now we take derivatives to get β_{MLE} as follows:

$$\begin{aligned}0 &= \frac{\partial}{\partial\beta}\left((\mathrm{y} - \mathbf{X}\beta)^T \mathbf{W} (\mathrm{y} - \mathbf{X}\beta)\right) \\ &= \frac{\partial}{\partial\beta}\left(\mathbf{y}^{\mathrm{T}}\mathbf{W}_{\mathrm{y}} - \mathbf{y}^{\mathrm{T}}\mathbf{W}\mathbf{X}\beta - \beta^T\mathbf{X}^{\mathrm{T}}\mathbf{W}_{\mathrm{y}} + \beta^T\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\beta\right)\end{aligned} \tag{9}$$

For any scalar $z, z = z^T$. therefore, $\left(\left(\beta^T \mathbf{X}^{\mathrm{T}}\right)(\mathbf{W}\mathbf{y})\right)^T = \mathbf{y}^{\mathrm{T}}\mathbf{W}^T\mathbf{X}\beta = \mathbf{y}^{\mathrm{T}}\mathbf{W}\mathbf{X}\beta$ since $\mathbf{W}^T = \mathbf{W}$ as $\mathbf{W}$ is diagonal. Now putting this back in (9) and taking derivatives, we get:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \beta}\left(\mathbf{y}^{\mathrm{T}}\mathbf{W}_{\mathrm{y}} - 2\beta^T\mathbf{X}^{\mathrm{T}}\mathbf{W}_{\mathrm{y}} + \beta^T\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\beta\right) \\ 0 &= -2\mathbf{X}^{\mathrm{T}}\mathbf{W}_{\mathrm{y}} + 2\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\beta \end{aligned} \tag{10}$$

Which means that

$$\beta_{MLE} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\right)^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{y}\right) \tag{11}$$

2 Regularization (3 marks)

(a) **Solution:** In this case, $\mathrm{rank}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right) = \mathrm{rank}(\mathbf{X}) \leq N << M$, thus the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, which is $M \times M$, is not full rank and thus can not be inverted.

(b) **Solution:**

$$\begin{aligned} \frac{\partial}{\partial \beta}\mathrm{J}_R(\beta) &= \frac{\partial}{\partial \beta}\left((\mathbf{X}\beta - \mathrm{y})^{\mathrm{T}}(\mathbf{X}\beta - \mathrm{y}) + \lambda\beta^T\beta\right) \\ &= \frac{\partial}{\partial \beta}\left(\mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\beta^T\mathbf{X}^{\mathrm{T}}\mathbf{y} + \beta^T\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta + \lambda\beta^T\beta\right) \\ &= -2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta + 2\lambda\beta \end{aligned} \tag{12}$$

Equating (12) with 0 and solving for β we get $\beta* = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$.

Note that $\lambda\beta = \lambda I\beta$.

3 Classification

Logistic regression (2 marks)

$$log(odds) \;=\; -10 + 2 * x_1 \tag{13}$$

(a) The probability of a cell being affected can be found by using the formula for logistic regression -

$$p(\mathcal{C}_A|\boldsymbol{x}_1) \;=\; \frac{1}{1 + \exp(-log(odds))} \tag{14}$$

Calculate log(odds) by substituting x_1 = 6, which is 2.
Substitute log(odds) in the equation to find probability as mentioned above.

Answer : **0.88** (rounded to nearest 2 decimal points)

(b) In the above equation of logistic regression, substitute 0.90 in $p(\mathcal{C}_A|\boldsymbol{x}_1)$ and calculate log(odds)
log(odds) = 2.197 Substitute log(odds) to derive x_1
Answer: **6.098**

Softmax for Multi-Class Classification (3 marks)

(a) The activation functions can be represented as the linear functions of input features.
a_A = $2x_1$ + $5x_2$ + 5
a_B = $5x_1$ + $10x_2$ + 1.5
a_C = $5x_1$ + $2x_2$ + 1

The class probabilities for Type A, Type B and Type C.
$p(\mathcal{C}_A|\boldsymbol{x}) = \frac{\exp(a_A)}{\exp(a_A)+\exp(a_B)+\exp(a_C)}$
$p(\mathcal{C}_B|\boldsymbol{x}) = \frac{\exp(a_B)}{\exp(a_A)+\exp(a_B)+\exp(a_C)}$
$p(\mathcal{C}_C|\boldsymbol{x}) = \frac{\exp(a_C)}{\exp(a_A)+\exp(a_B)+\exp(a_C)}$
Substitute the activation functions in the above equations.

(b) Substitute x_1 =10 and x_2=2 in the activation function equations and then the class probability equations
a_A = 35
a_B = 71.5
a_C = 55

$p(\mathcal{C}_A|\boldsymbol{x}_1)$ = 1.406 * 10^-16
$p(\mathcal{C}_B|\boldsymbol{x}_1)$ = 1.00
$p(\mathcal{C}_C|\boldsymbol{x}_1)$ = 6.825 * 10^-8
The cell belongs to Type B as the probability is the highest for Type B.