THE CHINESE UNIVERSITY OF HONG KONG, SHENZHEN

DDA 3020

MACHINE LEARNING

Assignment 1 Report

Author:

Ma Kexuan

Student Number:

ID 120090651

October 15, 2022

1 Written Problems

1.1. ① Proof: Assume X is a $d \times h$ matrix, then

$$\frac{d(x_j^T w)}{dw_i} = \frac{d(x_{j1}w_1 + x_{j2}w_2 + \cdots + x_{jd}w_d)}{dw_i} = x_{ji}$$

$$\frac{d(X^T w)}{dw} = \begin{bmatrix} \frac{dx_1^T w}{dw_1} & \cdots & \frac{dx_h^T w}{dw_1} \\ \vdots & & \vdots \\ \frac{dx_1^T w}{dw_d} & & \frac{dx_h^T w}{dw_d} \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{h1} \\ \vdots & & \vdots \\ x_{1d} & \cdots & x_{hd} \end{bmatrix} = X$$

② Proof: We let $a^T = y^T X$, then $a = X^T y$, since a is also not a function of w, we can obtain:

$$\frac{d(y^T X w)}{dw} = \frac{d(a^T w)}{dw} \xrightarrow[①]{\text{Apply}} a = X^T y$$

Thus, $\frac{d(y^T X w)}{dw} = X^T y$

③ Proof: w is $d \times 1$, X is $d \times d$, then by definition,

$$w^T X w = \sum_{j=1}^{d} \sum_{i=1}^{d} x_{ij} w_i w_j$$

Then, if we differentiate to the k^{th} element of w,

$$\frac{d(w^T X w)}{dw_k} = \sum_{j=1}^{d} x_{kj} w_j + \sum_{i=1}^{d} x_{ik} w_i \qquad \forall k.$$

$$= w^T X_k^T + w^T X_k$$

Consequently, $\frac{d(w^T X w)}{dw} = w^T X^T + w^T X = w^T (X^T + X) = (X + X^T) w$

1.2 (1) We can pack $f_{w,b}(x) = Wx + b$ as

$f_{w,b}(x) = X\bar{w}$, where

$$X = \begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \vdots \\ 1 & x_N^T \end{bmatrix} \in R^{N\times(d+1)} \quad , \quad \bar{w} = [b \;\; w_1 \;\; w_2 \cdots w_d]^T \in R^{(d+1)\times k}$$

and also $Y = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_N^T \end{bmatrix} \in R^{N\times k}$, $A = \begin{bmatrix} a_1 & & 0 \\ & a_2 & \\ 0 & & \ddots \; a_N \end{bmatrix} \in R^{N\times N}$

We can reformulate $\sum_{i=1}^{N} a_i \|y_i - Wx_i - b\|^2$ as:

$$\begin{aligned} &= (X\bar{w} - Y)^T A (X\bar{w} - Y) \\ &= \bar{w}^T X^T A X \bar{w} - \bar{w}^T X^T A Y - Y^T A X \bar{w} + Y^T A Y \\ &= f(\bar{w}) \end{aligned}$$

$$\frac{df(\bar{w})}{d\bar{w}} = 2X^T A X \bar{w} - 2X^T A Y$$

Let $\frac{df(\bar{w})}{d\bar{w}} = 0 \Rightarrow 2X^T A X \bar{w} = 2X^T A Y$

$$\hat{\bar{w}} = (X^T A X)^{-1} X^T A Y$$

Thus, the closed form solution is $\hat{\bar{w}} = (X^T A X)^{-1} X^T A Y$,

where $\bar{w} = [b \;\; w_1 \;\; w_2 \cdots w_d]^T$

(2) From (1), $J(\bar{w}) = (X\bar{w} - Y)^T A (X\bar{w} - Y)$,

$$\frac{dJ(\bar{w})}{d\bar{w}} = 2X^T A X \bar{w} - 2X^T A Y$$

We can update $\bar{w}$: $\bar{w}^* = \bar{w} - 2r \cdot X^T A (X\bar{w} - Y)$

where r is the appropriate step size, which can be determined by backtracking algorithm.

1.3 (1) $f'(x) = 4x^3 \quad f''(x) = 12x^2 \geq 0$

Since the second-order derivative is non-negative, then $f(x) = x^4$ is convex.

(2) $f(x) = |x|$ is not second-order differentiable on R, we prove by definition: $\forall x_1, x_2 \in R$, $\alpha \in [0, 1]$:

$$\begin{aligned} f(\alpha x_1 + (1-\alpha)x_2) &= |\alpha x_1 + (1-\alpha)x_2| \\ &\leq |\alpha x_1| + |(1-\alpha)x_2| \\ &= \alpha|x_1| + (1-\alpha)|x_2| \\ &= \alpha f(x_1) + (1-\alpha) f(x_2) \end{aligned}$$

We conclude that $f(x) = |x|$ is convex.

(3) $f(x) = ||Ax-b||^2 = (Ax-b)^T(Ax-b) = x^TA^TAx - x^TA^Tb - b^TAx + b^Tb$

$$\nabla f(x) = 2A^TAx - 2A^Tb$$

$$\nabla^2 f(x) = 2A^TA$$

Since $||Ax||^2 = (Ax)^TAx = x^T(A^TA)x \geq 0$,

A^TA is always a PSD matrix.

Thus, $f(x) = ||Ax-b||^2$ is convex.

1.4. The pdf is given by $f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

The likelihood function is

$$L(\mu, \sigma^2) = \prod_{n=1}^{N} f(x_n; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{\sum_{n=1}^{N}(x_n-\mu)^2}{2\sigma^2}\right)$$

We take natural log on both sides,

$$l(\mu, \sigma^2) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{\sum_{n=1}^{N}(x_n-\mu)^2}{2\sigma^2}$$

Take partial derivative w.r.t μ, σ^2.

$$\frac{\partial}{\partial \mu} l(\mu, \sigma^2) = -\frac{1}{2\sigma^2} \cdot (-1) \cdot \sum_{n=1}^{N} 2(x_n - \mu) = \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu) = 0$$

$$\frac{\partial}{\partial \sigma^2} l(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{n=1}^{N} (x_n - \mu)^2 = 0$$

Set derivatives equal to 0 and solve them, we get:

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n = \bar{x} \ ,$$

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N} (x_n - \bar{x})^2 = \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu_{ML})^2$$

2 Programming

2.1

1. Data Preprocessing

After loading the data, I drop the first two columns that represent “station” and “date”. What’s more, since there are still some NAN values in the original data, I drop all the rows that have missing columns.

2. Model Selection

I use the Linear Regression with multiple outputs to estimate the parameters, then do the prediction.

3. How to estimate the parameter W?

Since the amount of data is still relatively small, I choose the closed form solution to estimate the parameter W, that is, to add an additional column which contains 1s to the left side of the training data, then add an additional row to the top of the W matrix that we need to estimate. Finally, compute the matrix $\widehat{W} = (X^T X)^{-1} X^T Y$ by the function that I defined in the code.

4. Model Evaluation

By the train_test_split function provided by sklearn, I set up 10 distinct random seeds in each iteration to generate different set of training data. Then I estimate the parameter, and predict the y_hat_test as well as y_hat_train. After that, I separately calculate the train and test RMSE for Next_Tmax and Next_Tmin. Finally, I take the average of the two-dimensional RMSE to generate the result.

The RMSE definition:

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \widehat{y_t})^2}$$

Where y_t is the true value and $\widehat{y_t}$ is the predicted value for one column in the output matrix.

The result dataframe is shown below:

Index	Training_RMSE	Testing_RMSE
0	1.2293749	1.247396
1	1.2205682	1.2816744
2	1.225238	1.2626857
3	1.2365733	1.2175749
4	1.2327009	1.2332352
5	1.2289127	1.2478378
6	1.2333427	1.2309739
7	1.2313633	1.2382074
8	1.2201993	1.2809426
9	1.2330323	1.2321962

In summary, the model is quite robust between the training data and testing data. The temperature difference between the prediction and actual values are small. It's about 1.2-1.3 degree of error. So the model is relatively precise to do the regression of this data.

2.2

1. Data Preprocessing

Since there is no NAN value in the dataset, so there's no need to drop any value. Then, I transfer the class variables by get_dummies in pandas, which contain the value 1, 2 and 3, into one-hot encoding. Then the Y matrix contains 3 columns of binary values. After that, I continue to do the further analysis.

2. Model Selection

I use Linear Regression for classification to estimate the parameters, and do the prediction.

3. How to estimate the parameter W?

As for the measure in 2.1, I also use the closed form solution for 2.2. However, before the standard process of calculating $\widehat{W}$, I first change the Y variable to one-hot assignment. Then the remaining process is the same as 2.1.

4. Model Evaluation

By the train_test_split function provided by sklearn, I set up 10 distinct random seeds in each iteration to generate different set of training data. Then I estimate the parameter and calculate the regression result, and I use the argmax function provided by numpy to get the classification result. Finally I compute the error rate defined on the assignment requirement. Below is the result dataframe:

Index	Training_error	Testing_error
0	0.13333333	0.13333333
1	0.15833333	0.13333333
2	0.14166667	0.13333333
3	0.16666667	0.13333333
4	0.15	0.2
5	0.14166667	0.23333333
6	0.125	0.2
7	0.15833333	0.033333333
8	0.15833333	0.26666667
9	0.15	0.16666667

In summary, since two classes of the three aren't linearly separable, we cannot use the simple linear model to perfectly classify all the iris plants. Advanced non-linear model may be used to classify the data.