

Regression I

COMP9417, 23T1

- 1 Intro
- 2 Thinking Machine Learning
- 3 Linear Regression
- 4 Question 1 ($a \rightarrow c$)
- 5 Question 2 ($a \rightarrow h$)
- 6 3 ($a \rightarrow b$)

Section 1

Intro

Intro

What you'll get from this course:

- Understand the basis of machine learning
- ML algorithms and the math behind them
- Ability to implement these ideas in Python

How to do well:

- Fully understand tut questions from week to week (they pile up)
- Don't be afraid of math or notation, break it all down
- Keep researching

What do you prefer?

More theory, more practice (i.e Python and using packages), going through questions, consultation etc.

Section 2

Thinking Machine Learning

Thinking Machine Learning

We try to make sense of data using mathematics to help us quantify what we *know*.

A standard way to break the problem down is as follows:

- We have 'input' data X and targets/outputs y
- Our data can be modelled as $y = f(X)$
- Goal is to find the best approximation for f as \hat{f}

We define the quality of our approximation (\hat{f}) by using a error/loss function.

Section 3

Linear Regression

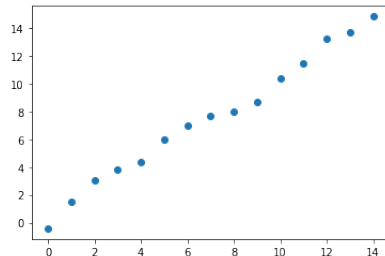
Linear Regression

We deduct and assume a linear relationship between X and y .

In this simple case, our model will take the form:

$$\hat{y} = w_0 + w_1 X$$

How do we find the optimal w_0 and w_1 ?



What will our loss function need? Boils down to the properties of the target function.

- Target function has ≈ 0 distance to all points
- We can define a basic loss function with one glaring issue:

$$L(w_0, w_1) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)$$

To make life easy, we define our loss function as:

$$\begin{aligned} L(w_0, w_1) &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 && \text{a.k.a MSE} \\ &= \frac{1}{n} \sum_{i=1}^n (y_i - w_0 - w_1 x_i)^2 && \text{by definition} \end{aligned}$$

The minimum of our loss function w.r.t w_0 and w_1 will be their optimal values respectively.

Section 4

Question 1 ($a \rightarrow c$)

1a

Derive the least-squares estimates for the univariate linear regression model.

i.e Solve:

$$\arg \min_{w_0, w_1} L(w_0, w_1)$$
$$\arg \min_{w_0, w_1} \frac{1}{n} \sum_{i=1}^n (y_i - w_0 - w_1 x_i)^2$$

First we differentiate $L(w_0, w_1)$ with respect to w_0 ,

$$\begin{aligned}\frac{\partial L(w_0, w_1)}{\partial w_0} &= -\frac{2}{n} \sum_{i=1}^n (y_i - w_0 - w_1 x_i) \\ &= -\frac{2}{n} \left(\sum_{i=1}^n y_i - nw_0 - w_1 \sum_{i=1}^n x_i \right)\end{aligned}$$

For the minimum, $\frac{\partial L(w_0, w_1)}{\partial w_0} = 0$,

$$-\frac{2}{n} \left(\sum_{i=1}^n y_i - nw_0 - w_1 \sum_{i=1}^n x_i \right) = 0$$

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n y_i - w_0 - w_1 \frac{1}{n} \sum_{i=1}^n x_i &= 0 \\ \bar{y} - w_0 - w_1 \bar{x} &= 0 \\ w_0 &= \bar{y} - w_1 \bar{x}\end{aligned}\tag{1}$$

To find w_1 , we follow a similar process and use simple simultaneous equations to solve for the final solution.

So,

$$\begin{aligned}\frac{\partial L(w_0, w_1)}{\partial w_1} &= -\frac{2}{n} \sum_{i=1}^n x_i (y_i - w_0 - w_1 x_i) \\ &= -\frac{2}{n} \left(\sum_{i=1}^n x_i y_i - w_0 \sum_{i=1}^n x_i - w_1 \sum_{i=1}^n x_i^2 \right)\end{aligned}$$

$$\frac{\partial L(w_0, w_1)}{\partial w_1} = 0,$$

$$\begin{aligned}\frac{1}{n} \left(\sum_{i=1}^n x_i y_i - w_0 \sum_{i=1}^n x_i - w_1 \sum_{i=1}^n x_i^2 \right) &= 0 \\ \overline{xy} - w_0 \bar{x} - w_1 \overline{x^2} &= 0\end{aligned}$$

$$\begin{aligned}\overline{xy} - w_0\bar{x} - w_1\overline{x^2} &= 0 \\ w_1 &= \frac{\overline{xy} - w_0\bar{x}}{\overline{x^2}}\end{aligned}\tag{2}$$

Sub (1) into (2):

$$\begin{aligned}w_1 &= \frac{\overline{xy} - (\bar{y} - w_1\bar{x})\bar{x}}{\overline{x^2}} \\ w_1 &= \frac{\overline{xy} - \bar{x}\bar{y} + w_1\bar{x}^2}{\overline{x^2}} \\ w_1\left(\frac{\overline{x^2} - \bar{x}^2}{\overline{x^2}}\right) &= \frac{\overline{xy} - \bar{x}\bar{y} + w_1\bar{x}^2}{\overline{x^2}} \\ w_1 &= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}\end{aligned}$$

Finally, we have

$$w_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \text{ and } w_0 = \bar{y} - w_1\bar{x}$$

1b

Problem: Prove (\bar{x}, \bar{y}) is on the line.

From 1(a), the equation of our line ($\hat{y} = w_0 + w_1x$) becomes:

$$\hat{y} = \bar{y} - \bar{x} \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} + \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}x$$

Sub $x = \bar{x}$,

$$\hat{y} = \bar{y} - \bar{x} \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} + \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}\bar{x}$$

$$\hat{y} = \bar{y}$$

$\therefore (\bar{x}, \bar{y})$ is on the line

1c

Similar to 1a, though take care with the partial derivatives:

$$\frac{\partial L(w_0, w_1)}{\partial w_0} = -\frac{2}{n} \sum_{i=1}^n (y_i - w_0 - w_1 x_i)$$

$$\frac{\partial L(w_0, w_1)}{\partial w_1} = -\frac{2}{n} \sum_{i=1}^n x_i (y_i - w_0 - w_1 x_i) + 2\lambda w_1$$

Final result is:

$$w_0 = \bar{y} - w_1 \bar{x}$$
$$w_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2 + \lambda}$$

Notice how the coefficients have an inverse relationship with λ .

Section 5

Question 2 ($a \rightarrow h$)

Math for multiple linear regression

Say we have our weight vector w and a constant vector c ,

$$\frac{\partial(cw)}{\partial w} = c^T$$

$$\frac{\partial(w^T cw)}{\partial w} = 2cw$$

$$\frac{\partial(cw^2)}{\partial w} = 2cw$$

2a

Problem: Show that $\mathcal{L}(w) = \frac{1}{n} \|y - Xw\|_2^2$ has critical point $\hat{w} = (X^T X)^{-1} X^T y$.

To find optimal w , solve $\frac{\partial \mathcal{L}(w)}{\partial w} = 0$

$$\begin{aligned}\mathcal{L}(w) &= \frac{1}{n} (y - Xw)^T (y - Xw) \\ &= \frac{1}{n} \left(y^T y - y^T Xw - w^T X^T y + w^T X^T Xw \right) \\ &= \frac{1}{n} \left(y^T y - 2y^T Xw + w^T X^T Xw \right)\end{aligned}$$

$$\frac{\partial \mathcal{L}(w)}{\partial w} = -\frac{1}{n}(-2X^T y + 2X^T X w)$$

To solve for \hat{w} ,

$$\begin{aligned} -2X^T y + 2X^T X \hat{w} &= 0 \\ \hat{w} &= (X^T X)^{-1} X^T y \end{aligned}$$

2b

Problem: Prove $\hat{w} = (X^T X)^{-1} X^T y$ is a global minimum.

$$\begin{aligned}\nabla_w^2 \mathcal{L}(w) &= \nabla_w (\nabla_w \mathcal{L}(w)) \\ &= \nabla_w (-2X^T y + 2X^T X w) \\ &= 2X^T X\end{aligned}$$

So, for a vector $u \in \mathbb{R}^p$,

$$\begin{aligned}u^T (2X^T X) u &= 2(u^T X^T)(Xu) \\ &= 2(Xu)^T (Xu) \\ &= 2\|Xu\|_2^2 \geq 0\end{aligned}$$

Therefore, \mathcal{L} is convex and \hat{w} is the unique global minimum.

2c

$$x_i = \begin{bmatrix} 1 \\ x_{i1} \end{bmatrix} \text{ to represent our input \& the bias } (w_0)$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \text{ to represent the target variable}$$

$$w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \text{ to represent the parameters}$$

$$X = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$X^T y = \begin{bmatrix} n\bar{y} \\ n\overline{xy} \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \end{bmatrix} \begin{bmatrix} 1 & x_{11} \\ 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}$$
$$= \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & n\overline{x^2} \end{bmatrix}$$

$$X^T X = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & n\overline{x^2} \end{bmatrix}$$

$$\begin{aligned} (X^T X)^{-1} &= \frac{1}{n^2\overline{x^2} - n^2\bar{x}^2} \begin{bmatrix} n\overline{x^2} & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \\ &= \frac{1}{n(\overline{x^2} - \bar{x}^2)} \begin{bmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \end{aligned}$$

2d

$$\begin{aligned}(X^T X)^{-1} X^T y &= \frac{1}{n(\overline{x^2} - \bar{x}^2)} \begin{bmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} n\bar{y} \\ n\overline{xy} \end{bmatrix} \\ &= \frac{1}{\overline{x^2} - \bar{x}^2} \begin{bmatrix} \overline{x^2}\bar{y} - \bar{x}\overline{xy} \\ \overline{xy} - \bar{x}\bar{y} \end{bmatrix} \\ &= \begin{bmatrix} \bar{y} - \hat{w}_1\bar{x} \\ \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \end{bmatrix}\end{aligned}$$

2e - Lab

Onto Jupyter.

2f

Say we have the classic regression problem with data $X \in \mathbb{R}^{n \times p}$ and target variable $y \in \mathbb{R}^n$. We can define a feature mapping $\phi : \mathbb{R}^p \rightarrow \mathbb{R}^K$. For example, say we have $p = 1$ and we choose $K = 4$, our mapping can be as follows

$$\phi(x) = \begin{bmatrix} x, & x^2, & x^3, & x^4 \end{bmatrix}^T$$

So our original model for a data point $i \in [1, n]$ becomes

$$\hat{y}_i = w^t \phi(x_i).$$

We can generalise our transformation to the matrix:

$$\Phi(x) = \begin{bmatrix} \phi(x_1)^T \\ \phi(x_2)^T \\ \vdots \\ \phi(x_n)^T \end{bmatrix} \in \mathbb{R}^{n \times K}$$

As we use the transpose of our transformation, our model now takes the form $\hat{y} = \Phi w$.

This allows us to solve

$$\hat{w} = \arg \min_w \frac{1}{n} \|y - \Phi w\|_2^2$$

Which gives us the classic form of the LS solution:

$$\hat{w} = (\Phi^T \Phi)^{-1} \Phi^T y$$

2h

$$\text{MSE}(w) = \arg \min_w \frac{1}{n} \|y - Xw\|_2^2 \text{ and } \text{SSE}(w) = \arg \min_w \|y - Xw\|_2^2$$

i) Is the minimiser of MSE and SSE the same?

ii) Is the minimum value of MSE and SSE the same?

Section 6

3 ($a \rightarrow b$)

3a

What is the difference between a population and a sample?

3b

What is population parameter? How can we estimate it?