Lecture 4 - Linear models

Linear regression

Issac Lee

2021-03-26

Sungkyunkwan University

Actuarial Science

Linear models

Matrix theory

Definitions and results

Matrix

$\mathbf{A}_{n\times m} = [a_{ij}]$ is a rectangular array of elements.

- Demension of $\mathbf{A}$: n (rows) by m (columns)
- Square matrix if $n = m$.
- A vector $\mathbf{a}_{n\times 1} = [a_i]$ is a matrix consisting of one `column`.
- Our interests is on real matrices: whose elements are real numbers.

Transpose

If $\mathbf{A}_{n\times m} = [a_{ij}]$ is $n \times m$, the transpose of $\mathbf{A}$, $\mathbf{A}^T$ is $m \times n$ matrix $[a_{ji}]$.

- Symmetric if $\mathbf{A} = \mathbf{A}^T$

Propsition 1 If $\mathbf{A}$ is $n \times m$ and $\mathbf{B}$ is $m \times n$, the $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$

T.B.D

Simple linear regression

- Response variable y_i is linearly related to an independent variable x_i, given by

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, n$$

where $e_1, \ldots, e_n$ are typically assumed to be uncorrelated random variables with mean zero and constrant variance σ^2.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{pmatrix}, \boldsymbol{X}\beta = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \ldots & \ldots \\ 1 & x_{n-1} \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ \ldots \\ e_n \end{pmatrix}$$

Multiple linear regression

Response variable y_i is linearly related to p independent variables x_{ij}s, given by

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ij} + e_i, \quad i = 1, \ldots, n, j = 1, \ldots, p$$

which is the same as

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + e_i, \quad i = 1, \ldots, n$$

where

$$\begin{aligned} \mathbf{x}_1^T &= (x_{11}, \ldots, x_{1p}), \\ &\ldots \\ \mathbf{x}_n^T &= (x_{n1}, \ldots, x_{np}), \end{aligned} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \ldots \\ \beta_p \end{pmatrix}$$

Multiple linear regression

We assume

$$\mathbb{E}\left(\boldsymbol{e}\right) = \boldsymbol{0}, Var\left(\boldsymbol{e}\right) = \boldsymbol{\sigma}^2 I_n$$

where I_n is an identity matrix size of n.

Regression problem

Linear model problem can be viewed as a best approximation $\mathbf{X}\beta$ to the observed $\mathbf{y}$.

- If we define closeness or distance in Euclidean manner, then the problem becomes to find a value of the vector β that minimizes $L(\beta)$ as follows;

$$\begin{aligned} L(\beta) &= (\mathbf{y} - \boldsymbol{X}\beta)^T (\mathbf{y} - \boldsymbol{X}\beta) \\ &= \|\mathbf{y} - \boldsymbol{X}\beta\|^2 \end{aligned}$$

- Solution: Find the gradient vector of $L(\beta)$ and set it equals to zero.

$$\frac{\partial L}{\partial \beta} = \begin{pmatrix} \frac{\partial L}{\partial \beta_1} \\ \dots \\ \frac{\partial L}{\partial \beta_p} \end{pmatrix}$$

Practice

Find $\frac{\partial f}{\partial \beta}$

$$f(\beta) = \beta_1 x_1 + \beta_2 x_2$$

Find $\frac{\partial g}{\partial \beta}$

$$g(\beta) = \beta_1^2 + 4\beta_1\beta_2 + 3\beta_2^2$$

Derivative rules

Let $\mathbf{a}$ and $\mathbf{b}$ be $p \times 1$ vectors and $\mathbf{A}$ be $p \times p$ matrix of constants. Then,

- $\frac{\partial \mathbf{a}^T \mathbf{b}}{\partial \mathbf{b}} = \mathbf{a}$
- $\frac{\partial \mathbf{b}^T \mathbf{A} \mathbf{b}}{\partial \mathbf{b}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{b}$

What is the $\frac{\partial L}{\partial \beta} =?$

$$
\begin{aligned}
L(\beta) &= (\mathbf{y} - \boldsymbol{X}\beta)^T (\mathbf{y} - \boldsymbol{X}\beta) \\
&= \|\mathbf{y} - \boldsymbol{X}\beta\|^2
\end{aligned}
$$

Normal equation

Setting the gradient to zero, we obtain Normal Equation;

$$\boldsymbol{X}^T\boldsymbol{X}\beta = \boldsymbol{X}^T\mathbf{y}$$

The solution of this equation is as follows;

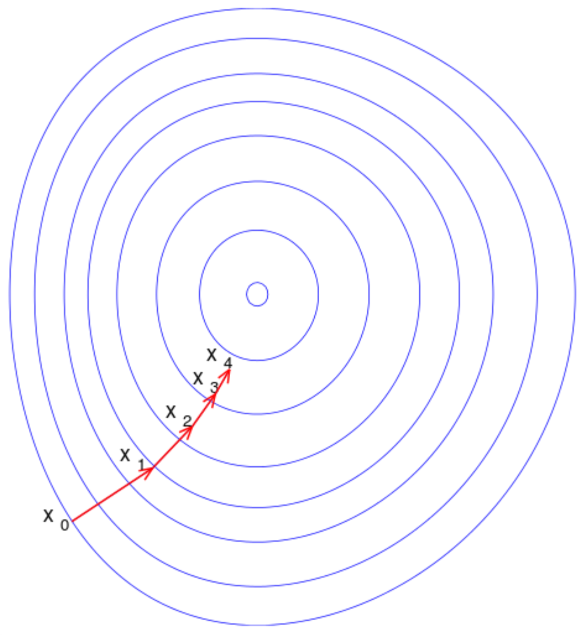
$$\hat{\beta} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Food for thought

Are we happy about this always?

What is the problem?

Gradient descent



$$\beta_{n+1} = \beta_n - \gamma \nabla L(\beta_n)$$

Linear Basis function models

$$f(x) = \sum_{j=0}^{M-1} \beta_j \phi_j(x) = \boldsymbol{\Phi}(x)\,\beta$$

where $\phi_j(x)$ are known as **basis functions**.

typically, $\phi_0(x) = 1$ so that β_0 becomes a bias.

Example of basis functions

- Polynomial basis functions (global)

$$\phi_j(x) = x^j$$

- Gaussian basis (local)

$$\phi_j(x) = exp\left(-\frac{(x-\mu_j)^2}{2\sigma^2}\right)$$

- Sigmoidal basis functions (local)

$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right)$$

where

$$\sigma(x) = \frac{1}{1+exp(-x)}$$

Easy Example

Polynomial Curve Fitting

$$y = sin(2\pi x) + \epsilon$$

```
## # A tibble: 10 x 2
##        x      y
##    <dbl>  <dbl>
##  1  0.3   1.00
##  2  0.25  1.18
##  3  0.65 -0.806
##  4  1     0.346
##  5  0.55 -0.525
##  6  0.15  0.754
##  7  0.95 -0.273
##  8  0.7  -0.649
##  9  0.5   0.321
## 10  0.9  -0.956
```

0th order polynomial

$f(x) = \beta_0$

1th order polynomial

$f(x) = \beta_0 + \beta_1 x$

3th order polynomial

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

Feel so good~! Let's do 9th!!

Is this looks okay? Why?

Avoid Over fitting: Regularization

Priviously we looked at linear models. Let's extend our candidates!

$$RSS(f) = (\mathbf{y} - f(X))^T (\mathbf{y} - f(X))$$

To avoid the overfitting, we will consider the following penalized RSS, PRSS;

$$PRSS(f;\lambda) = RSS(f) + \lambda J(f)$$

where the functional $J(f)$ represents a regularization term.

Bias-Variance trade off

We observe a quantitative responds Y and p different perdictors, $X_1, \ldots, X_p$.

$$Y = f(X) + \epsilon$$

where $X = (X_1, \ldots, X_p)$. ϵ is a random error term, which is independent of X and has mean zero.

We can predict Y using

$$\hat{Y} = \hat{f}(X),$$

where $\hat{f}$ represents our estimate for f, and $\hat{Y}$ represents the resulting prediction for Y.

Accuracy of $\hat{Y}$

The accuracy of $\hat{Y}$ as a predicton for Y depends on two quantities;

- Reducible error
- Irreducible error

$$\mathbb{E}\left(Y-\hat{Y}\right)^2 = \mathbb{E}\left[\left(f\left(X\right)+\epsilon-\hat{f}\left(X\right)\right)^2\right]$$
$$= \left[f\left(X\right)-\hat{f}\left(X\right)\right]^2 + Var\left(\epsilon\right)$$

Expected test error

Expected test error can be decomposed as the following three terms;

- $Variance, Noise, Bais^2$

$$
\begin{aligned}
&\mathbb{E}_{D,X,y}\left[\left(\hat{f}_D(X)-y\right)^2\right] \\
=&\mathbb{E}_{X,D}\left[\left(\hat{f}_D(X)-\bar{f}(X)\right)^2\right]+ \\
=&\mathbb{E}_{X,y}\left[\left(\hat{f}(X)-y\right)^2\right]+ \\
=&\mathbb{E}_{X}\left[\left(\bar{f}(X)-\hat{f}(X)\right)^2\right]
\end{aligned}
$$

Ridge regression

Ridge regression use L_2 norm

$$\underset{\beta}{min}(y - X\beta)^T (y - X\beta) + \frac{\lambda}{2}\|\beta\|_2^2$$

H.W. What is the optimal $\beta_\star$?

Lasso regression

Lasso regression use L_1 norm

$$\underset{\beta}{min}(y - X\beta)^T (y - X\beta) + \frac{\lambda}{2}\|\beta\|_1$$

Elastic Net

Why don't we have the both of the two?

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

The loss function can be parameterized with the two parameters; λ, α

- λ controls the magnitude
- α controls the weights of the two panalty functions

$$\underset{\beta}{min}(y - X\beta)^T(y - X\beta) + \frac{\lambda}{2}\left(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2\right)$$

Problem

So we have the two models like `Lasso` and `Ridge` regression, and more extended model called `Elastic net`. These models have the parameters.

How do we determine these parameters?

- We can't use test dataset. (That's cheating and in Kaggle we don't know the dependent variables)

Validation set

Make our own validation set using `train data set`.

- Assumption: train and test data set have the same data distribution.

Hyperparameter Tuning

If our model perform well on the validation set, it will work well in the test data!

- Tunning the hyperparameter using validation set.

Thanks!