로지스틱 회귀분석

ToBig's 6기 김지수

Logistic regression

투빅스 8기 정규세션

Contents

그냥 대강 보자

$$Y = \beta_0 + \beta_1 X$$

X가 한 단위 증가하면 Y는 B_1만큼 변화!! 통계적으로 끄적대면 $E(Y|X)$

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

다중회귀는 행렬식

$$Y = XB$$

$$\beta = (X^t X)^{-1} X^t y$$

$$\begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ y_n \end{pmatrix} = \begin{pmatrix} 1\ 1\ 1\ 1 \cdots\cdots\cdots\cdots 1 \\ x_{11} \cdots\cdots\cdots\cdots x_{1n} \\ x_{21} \cdots\cdots\cdots\cdots x_{2n} \\ \\ x_{n1} \cdots\cdots\cdots\cdots x_{nn} \end{pmatrix} \begin{pmatrix} B_0 \\ B_1 \\ B_2 \\ \\ B_n \end{pmatrix}$$

(nxn) **(nx1)**

행렬의 미분

스칼라를 벡터로

$$\frac{\partial scalar}{\partial vector} = \nabla y = \frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_N} \end{bmatrix}$$

$$for\ example$$
$$f(x,y) = 2x^2 + 6xy + 7y^2 - 26x - 54y$$

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} = \begin{bmatrix} 4x + 6y - 26 \\ 6x + 14y - 54 \end{bmatrix}$$

알아두면 유용한 행렬의 미분

1. In linear model

선형 모형을 미분하면 가중치 벡터가 된다!

$$\frac{\partial w^t x}{\partial x} = \frac{\partial x^t w}{\partial x} = w$$

$$x = (x_1, x_2, \ldots, x_k)^t$$

2. qudratic from

이차 형식을 미분하면 행렬과 벡터의 곱으로 나타난다.

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}) = \mathbf{x}^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}} + \mathbf{A})$$

$$\frac{\partial(\mathbf{x} - \mathbf{A}\mathbf{s})^{\mathrm{T}}\mathbf{W}(\mathbf{x} - \mathbf{A}\mathbf{s})}{\partial \mathbf{s}} = -2\mathbf{A}^{\mathrm{T}}\mathbf{W}(\mathbf{x} - \mathbf{A}\mathbf{s})$$

http://blog.naver.com/PostView.nhn?blogId=enewltlr&logNo=220918689039&parentCategoryNo=&categoryNo=83&viewDate=&isShowPopularPosts=true&from=search

OLS(최소제곱법)

$$S(B) = \text{argmin}(\sum(Yi - \hat{Y}i)^2)$$

$$S(B) = \text{argmin}((Y - XB)^2)$$

회귀분석

종속변수가 연속형인 경우 사용
하지만, 많은 경우에 종속변수가 범주형이거나, 수치형이어도 연속형 변수의 개념으로 해석 할 수 없는 경우도 존재!!

일반화 선형모형(glm)

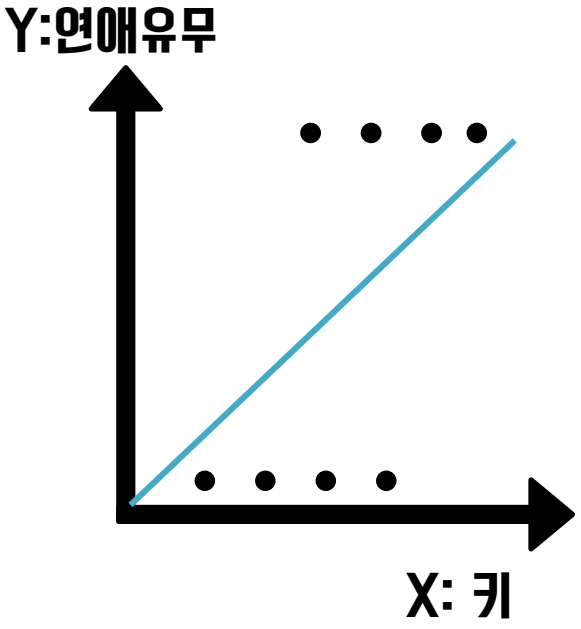
위와 같은 선형회귀분석이 제한되는 경우에도 사용이 가능하며,
회귀분석의 기초가정에 위배되는 상황에도 유연하게 대처가 가능하다.

Ex) logistic reg/penalized reg/poission reg 등등

Logistic Regression

투빅스에서 회귀분석만 알려주고 분류과제를 시켰어 난 아는게 회귀분석밖에 없는데
이럴 땐 뭘 해야 할까 ㅠㅠ

로지스틱 회귀분석?



종속변수가 이산형 변수!!
이런 경우에 회귀분석이 가능할까?
실제 종속변수는 0과 1인데

$$Y = \beta_0 + \beta_1 X$$

이 모델로 예측이 가능해?
당연히 안되죠!

로지스틱 회귀분석?

Y라는 binary 변수를 확률로 바꿔서 생각해봅시다!!

왜? 확률은 연속형이니까!!

P(Y=1|x) : x가 주어졌을 때 Y가 1일 확률!!

P(Y=1|x) = $\beta_0 + \beta_1 x$ 꼴의 수식이 등장!

확률은 0과 1사이 값인데 저걸 어떻게 하지..?
라는 의문이 드신다면 이미 반쯤 이해하신 겁니다.

Odds(승산)

오즈: 실패확률 대비 **성공확률**

오즈는 0에서 Inf 사이의 값을 갖습니다!

$$Odds = \frac{p(Y=1|X)}{1-p(Y=1|X)}$$

로짓: 오즈에 로그를 씌운 값.

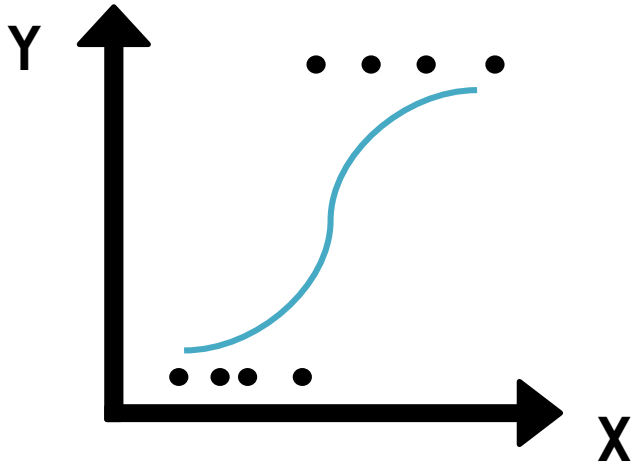
로짓은 -Inf에서 Inf 사이의 값을 갖습니다!!

$$\log it = \log(odds) = \log\frac{p}{1-p}$$

로지스틱 회귀모형

$$\log\left(\frac{p(Y=1|X)}{1-p(Y=1|X)}\right)=X\beta$$

이 `로짓` 을 반응변수로 선형 모형을 만드는 것이 로지스틱 회귀분석!!



오잉 그러면 회귀계수는 어떻게 구하죠?

저번처럼 최소제곱법(OLS)으로 구하나요?

회귀계수 구하기

최소제곱법을 사용할 수 없어요!!

왜? 오차를 구할 수가 없어서!! -> 반응변수 binary

따라서.. MLE라는 방법을 통해 회귀계수를 구해야 한다!

오차의 제곱합을 최소화하는 것은 관측될 데이터의 $likelihood$를 최대화 하는 것과 동일하다!

$$p = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

$$L(X, \beta) = \prod p_i^{y_i} (1 - p_i)^{1 - y_i}$$

얘를 미분

회귀계수의 해석

회귀분석: X의 변화에 따른 Y의 변화량

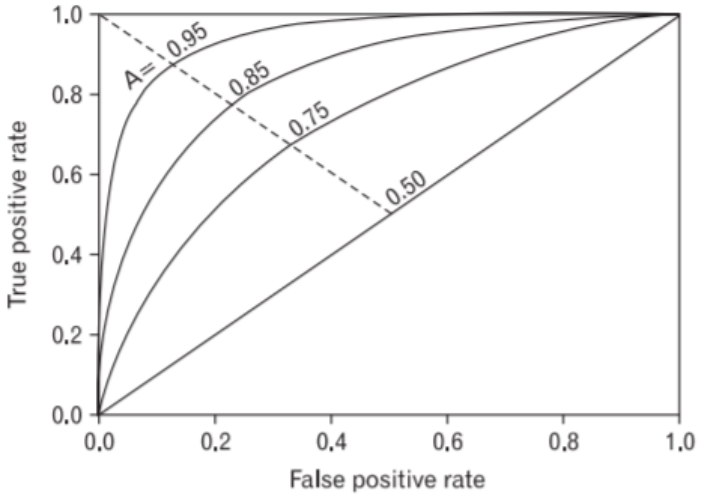
로지스틱 회귀분석: X의 변화에 따른 Y의 로짓/오즈의 변화!!

X가 한 단위 증가하면, y의 로짓이 B만큼 변화!!

➔ Y의 오즈는 e^B 배 변화!!

R에서는 function을 실행하면
베타계수를 반환!!
=> 고려해서 해석

ROC Curve



내가 만든 model이 얼마나 적합한 모델인가를 판단하는 graph

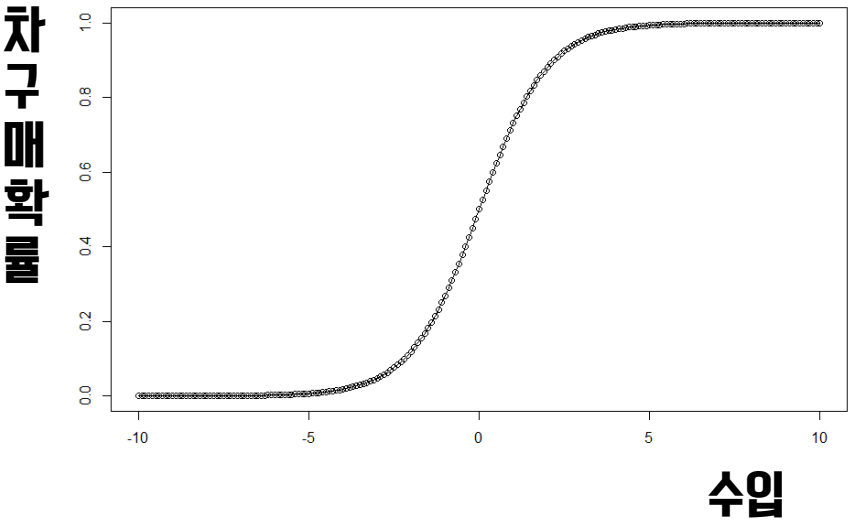
아래의 면적이 넓을수록 더 좋은 모델이다!!

Y축 : sensitivity(true positive)

X축 : 1-specify(False positive)

https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Sigmoid function



$$sigmoid(x) \qquad \frac{1}{1+\exp(-z)}$$

$$sigmoid(\mathrm{W}^t X) \qquad \frac{1}{1+\exp(-w^t X)}$$

실제 세상에는 많은 데이터가 저런 모양으로 생김!!
sigmoid 혹은 s자 형태라고 하는데
로지스틱 회귀분석이 실제 데이터에는 더 잘 적합
될 수 있음

Softmax Regression

투빅스에서 이번엔 로지스틱만 알려주고 다범주 모형 분류를 시켰어 ㅠㅠ

Multinomial classification

$$p(y = A\ |x) = \frac{1}{1+\exp(-w^t X)}$$ A or not A

$$p(y = B\ |x) = \frac{1}{1+\exp(-w^t X)}$$ B or not B

$$p(y = C\ |x) = \frac{1}{1+\exp(-w^t X)}$$ C or not C

Softmax regression

$binary\ classification * num\ of\ level(k)$만큼 수행!!

$tensorflow$로 하면 한방에 가능함..

통계과에서는 $multi\ logit$모형이라고 합니다 (거의 같음!)

$$softmax(y_i) = \frac{\exp(y_i)}{\sum_{j=1}^{k} \exp(y_j)}$$

이 친구가 최대인 i로 분류를 하면 됩니다.

무슨말인고 하니..

$$p(y = A\ |x) = \frac{1}{1+\exp(-w_1^t X)}$$

$$p(y = B\ |x) = \frac{1}{1+\exp(-w_2^t X)}$$

$$p(y = C\ |x) = \frac{1}{1+\exp(-w_3^t X)}$$

클래스마다 그때그때 Gradient decent(로지스틱으로 하면 MLE)에 의해 구해지는 w,b (회귀분석으로 치면 베타계수)의 값이 다름!!
➔ 이 값들을 비교하면 됩니다.

Cost function

In logistic
$misclassification\ rate$**(오분류율)** : $y * \log\big(H(x)\big) - (1-y) * \log(1-H(x))$

In softmax
$cross-entropy$**(교차 엔트로피)** : $-\sum_k y_k * \log(\widehat{y_k})$

사실 두개 똑같은 말임 ㅎㅎ

정리

1. 다수의 binary classification 을 사용
2. 1을 통해 나온 결과를 0~1사이 값으로 변경
3. 2를 통해 나온 결과들의 총 합은 1이 되고
4. 가장 큰 값을 지니는 클래스로 때려 박는다.

회귀식의 추정이 쉽지 않은경우

1.설명변수들간의 연관관계가 심한 경우!!(다중 공선성)

2.샘플 수(n) << 변수 개수(p) 인 경우

조금 더 수식적으로 보면

$$\beta = (X^t X)^{-1} X^t y$$

$$V(\beta) = \sigma^2 (X^t X)^{-1}$$

우선 1번의 경우에는 베타계수의 분산이 무진장 커진다!!
➔ 추정의 정도가 크게 하락한다

2번의 경우에는 $(X'X)$의 역행렬이 존재하지 않는다.
➔ 애초에 추정부터가 안된다.

능형 회귀란?

위와 같은 OLS 추정의 문제를 해결하기 위한 방법론!!

베타계수를 축소시켜서 true한 effect만 남긴다.

⇒ 어떻게 축소시킬까?

Penalty를 회귀계수 추정식에 부여함!!

Penalty가 포함된 회귀계수 추정식을 최소화하는 베타계수를 찾으면 됩니다!!

능형 회귀란?

$S(b(\lambda)) = (XX + \lambda I)^{-1}X'Y$ 꼴이 주어짐.

=> 이렇게 구해진 능형 추정량의 베타계수는 해석에는
어려움이 있지만, 더 좋은 예측도를 가진다.

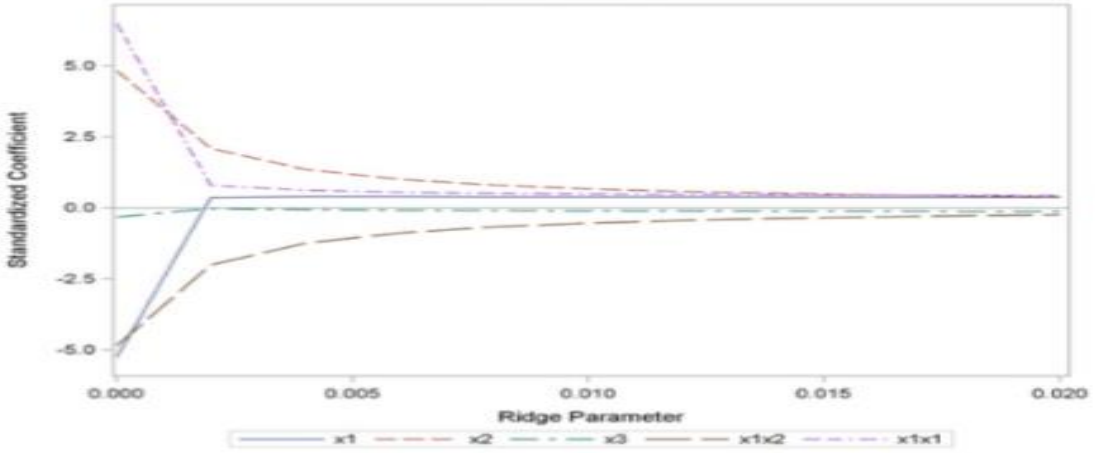
$$S(B) = argmin(||\mathrm{Y} - XB|| + \lambda B^2)$$

$$\hat{\beta}^{ridge} = \min_{\beta}\left\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\}$$

여기서 저 lambda는 능형 모수라고 불리우며, 어떠한 경우에도
'추정'의 측면에서 simple OLS보다 우수한 적합도를 보이는
양수인 lambda값이 항상 존재한다!! => lower 해석도/higher 예측도

How to find λ?

그렇다면 이 λ를 구하는게 Beta 추정에서 중요!!
위에서 말한 것처럼 베타계수는 축소되어 추정된다.



Ridge parameter λ값의 변화에 따른 베타계수들의 축소량을 보여주는 graph

일반적으로 축소되어지는 베타계수들의 기울기가 안정되어 가는 점의 λ값을 선택한다.
➔ r이 알아서 골라줍니당!

Model 결정

Ridge에서 model은 선택한 λ값에서의 coef에 의해 결정된다.

➔ coef가 0에 가까이 축소되는 변수들은 자연스럽게 모델에서 영향력이 감소합니다.

물론 coef가 완전히 0에 수렴하진 않아서 모델에 여전히 변수들은 포함된 상태이다.

➔ 변수선택을 따로 하지 않은 상태에서 변수의 coef의 크기만을 조정한 상태로 모형을 적합시킨다.

Sparse linear model

$$y = x\beta + \xi = x_a * \beta_a + \ldots + x_n\beta_n = x_a * \beta_a + \xi$$

x_a = signal/true/relevant variable

β_a = true coef with nonzero element

x_n = noisy/false/irrelevant variable

β_n = true coef with zero element

이렇게 true 한 effect들만 남긴 model이 정확도/해석도 측면에서 더 우수하다

=> 어떻게 골라낼까?

LASSO

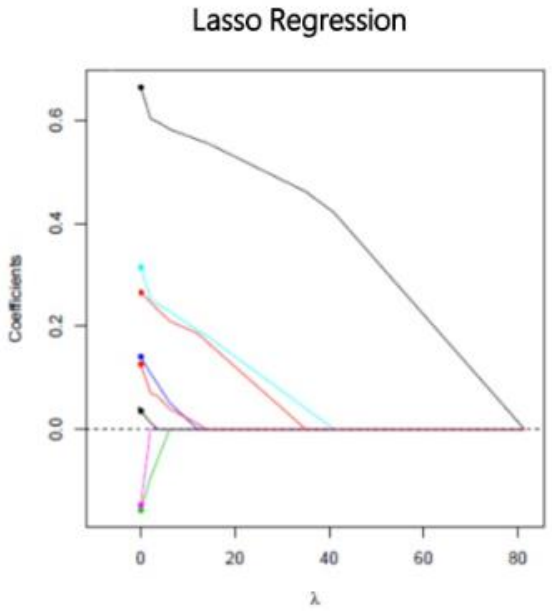
Ridge와 비슷한 개념(penalty form이 조금 다름)을 가져간 상태에서 coef를 아예 0까지 완전 축소시킨다.

➔ Coef가 0이 된 변수들은 모형에서 자연스럽게 제거된다.
➔변수선택이 가능하다.

S(B) = argmin($||Y - XB|| + \lambda|\boldsymbol{b}|$**)**

$$\hat{\beta}^{lasso} = \min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

LASSO



Ridge regression과 다르게 λ값의 증가에 따라 변수들의 coef가 0으로 완벽히 축소됨!!

➔ Coef가 0에 가까이 축소되는 변수들은 자연스럽게 모델에서 빠지게 된다.

summary

Penalty form

Ridge : λB^2

LASSO : $\lambda |B|$

OLS 식에 저 페널티를 추가한 식을 최소화 시켜주는 베타를 찾는게 곧 릿지이고 라쏘!!

Cross-Validation

한정된 데이터에서
반복 측정/검증을 통하여
모델을 적합하는 과정!!

Validation

일반적으로 train/test set의 비율은 7:3으로 설정

Train set으로 종속변수를 예측하는 모델을 설정한 후

해당 모델로 test set을 적합하고 참값과 비교하여
Prediction rate 를 도출한다.

Cross - validation

Overfit problem(과적합) : train set에서의 noise 한 값들에 대한 예측은 잘하지만, test set에 적합했을 때 예측률이 낮아지는 현상

혹은 train set에서 모델을 적합했는데, test 셋의 noise한 값들에 대해서 제대로 적합하지 못하는 현상

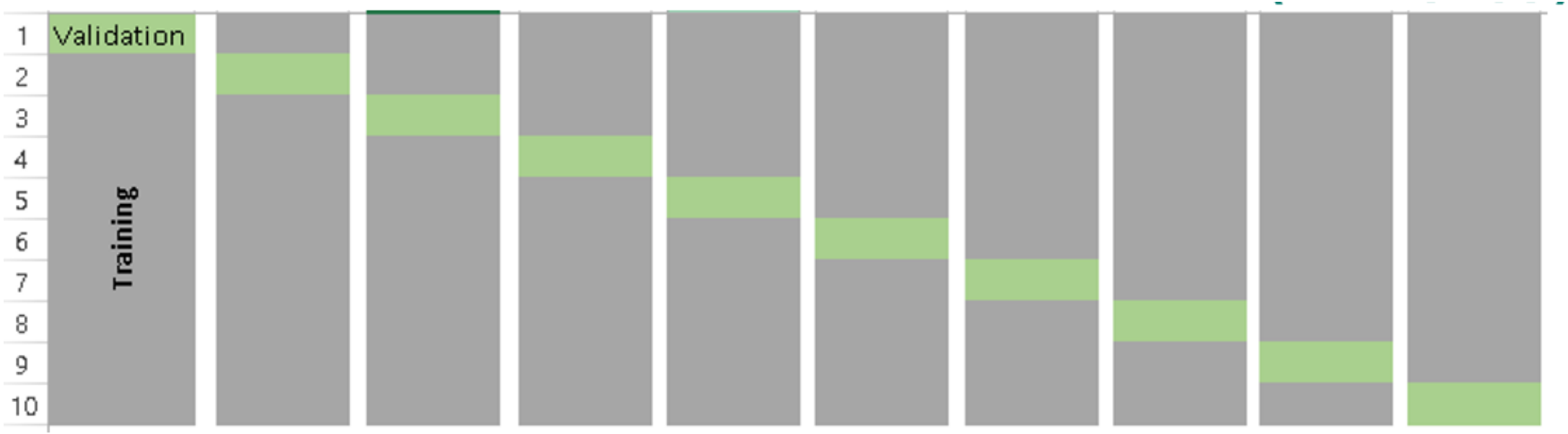
Cv의 종류

Cross-validation(교차 타당성) : 여러 개의 fold로 쪼개서 데이터를 적합하자!

Validation : 데이터 하나를 train/test로 쪼개서 적합한다.

LOOCV : sample수 만큼 모델을 만들어서 적합

Cross-Validation



데이터를 k개 fold로 분할하여 남은 데이터로 모델을 적합하여 테스트하는 방식!!

Leave One Out cv

Sample 수 만큼 모델을 만드는 방식

총 N번의 test를 거쳐 모델 적합

모든 경우를 다 고려할 수 있지만, 굉장히 비효율적인 방법

말 최소 이동횟수 찾기!
데이터 분석해오기!

(a,b)에서 (x,y)까지 가는 최소 이동경로 구하기(말은 나이트)

중간중간 말을 놓을 수 없는 구역(제가 보드 드릴 거에요)

분석은 데이터 드린 거 분석 해오시면 됩니당!

Q&A

고생하셨습니다

다들 Rstudio를 켜주세요~