19기 정규세션

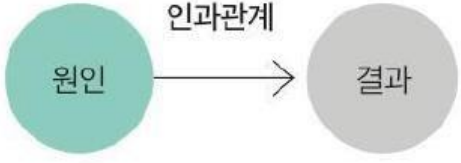
ToBig's 18기 강의자
김희경

Week 2:
Regression Analysis

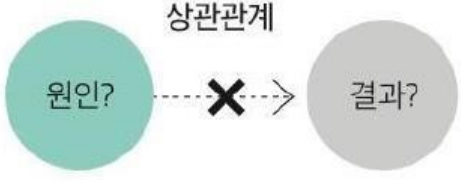
1. 머신러닝 알고리즘

지도학습 (Supervised Learning)	비지도학습 (Unsupervised Learning)	강화학습 (Reinforcement Learning)
- 입력과 결과값(Label) 이용한 학습 - 회귀(Regression) - 분류(Classification)	- 입력만을 이용한 학습 - 군집화(Clustering)	- Agent가 주어진 State에서 Action을 취했을 때, 이로부터 얻는 Reward를 최대화하는 방향으로 학습
Ex) 선형회귀, 로지스틱 회귀, KNN, SVM, Decision Tree	Ex) K-Means Clustering	

2. 인과관계 VS 상관관계



원인이 있었기 **때문에** 결과가 생겨났다.



원인과 결과의 관계가 아니다.

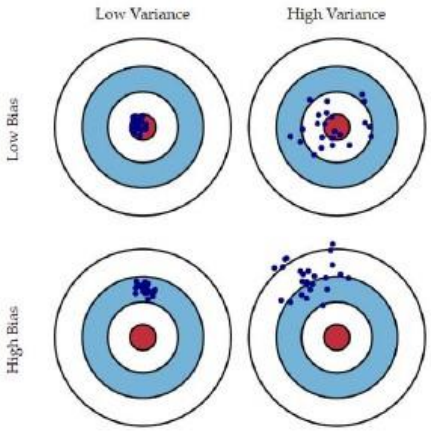
인과관계(Causality)

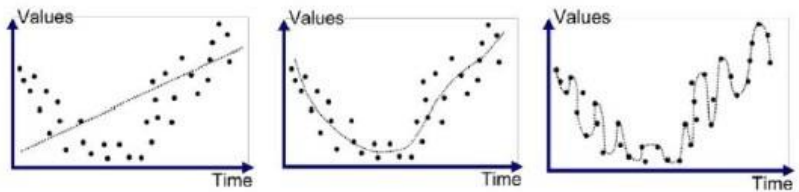
- 어떤 사실과 다른 사실 사이의 원인과 결과 관계

상관관계 (Association, Correlation)

- 두 변량 중 한쪽이 증가함에 따라, 다른 한쪽이 증가하거나 감소하는 관계
- 상관관계가 존재할 때, 필연적으로 인과관계가 존재하는 것은 아님

3. 편향(Bias) VS 분산(Variance)





high bias
low variance

medium bias
medium variance

low bias
high variance

Bias(편향)

- 데이터 내 모든 정보를 고려하지 않기에 알고리즘이 지속적으로 잘못된 내용을 학습하는 경향성
- Underfitting과 관련

Variance(분산)

- Highly flexible model에 데이터를 fit함으로써, 실제 현상과 관계없는 random한 것들까지 학습하는 알고리즘의 경향성
- Overfitting과 관련

Contents

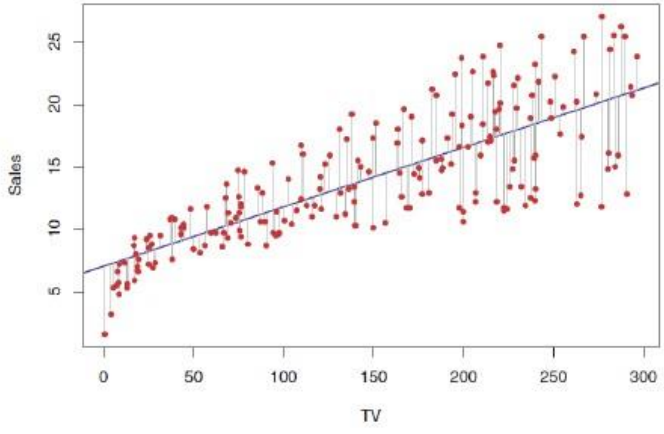
선형 회귀분석(Linear Regression)

" 회귀분석 "

- 설명변수 (X)에 대응하는 반응변수 (Y)와 가장 비슷한 값 ($\hat{Y}$)을 출력하는 함수를 찾는 과정
- 변수들의 관계를 기술하고 형태를 파악하는 통계적인 기법

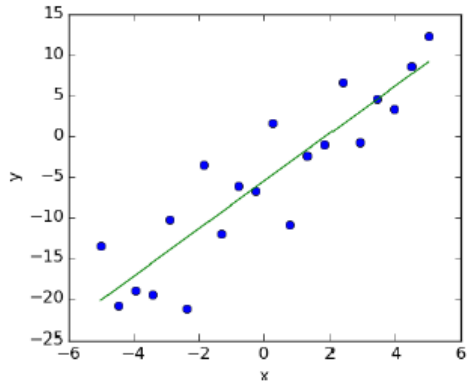
" 선형회귀분석"

- 반응변수와 한 개 이상의 설명변수와의 선형 상관관계를 모델링하는 회귀분석 기법
- Ex) 해당 연도 수확량(X)에 따른 열매 개수(Y)



Unit 01 | 선형회귀: 기본 선형 회귀

단순 선형 회귀



$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$Formulation : \hat{Y} = \widehat{\beta_0} + \widehat{\beta_1}\, \boldsymbol{x}$$

- β_0 , β_1 : 회귀계수
- $\widehat{\beta_0}$, $\widehat{\beta_1}$: 예측된 회귀계수

Least Square Method(LSE) -단순선형회귀

$$L = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

⬇ Partial differential for minimization

Normal Equation(정규방정식)

$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)) = 0$$

$$\frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)) x_i = 0$$

➡ Result

Least Squares Estimator(최소제곱 추정치)

$$\widehat{\beta_1} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1} \bar{x}$$

Least Square Method(LSE) -다중선형회귀

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_k x_{1k} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_k x_{2k} + \varepsilon_2 \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_k x_{nk} + \varepsilon_n
\end{aligned}
$$

⬇

$$
\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} \cdots x_{1k} \\ 1 & x_{21} & x_{22} \cdots x_{2k} \\ \vdots & \vdots & \vdots \quad \vdots \\ 1 & x_{n1} & x_{n2} \cdots x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}
$$

⬇

$$
y = X\beta + \varepsilon \qquad \varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'
$$

$$
\begin{aligned}
\sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon &= (y - X\beta)'(y - X\beta) \\
&= y'y + \beta' X' X\beta - 2\beta' X' y
\end{aligned}
$$

⬇ Partial differential for beta

정규방정식

$$
\begin{aligned}
\frac{\partial L}{\partial \beta} &= 2X'X\beta - 2X'y = 0 \\
&\Rightarrow X'X\beta = X'y
\end{aligned}
$$

최소제곱 추정치

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

Unit 01 | 선형회귀: 기본 선형 회귀

왜 최소제곱법을 사용할까?

β의 유의확률 공명

Contents

Unit 02 | 선형회귀: 규제

선형모형의 개선이 가능할만한 case

주어진 자료가 고차원 자료일 때
- n ≥ p이되 n≈p이면 최소제곱 추정량의 분산이 급격하게 증가한다.
  따라서 회귀 계수의 추정 및 반응 변수 예측의 안정성↓
- N < p 이면 최소제곱 추정량은 유일하지 않다.

모형 해석가능성을 제고하기 위한 변수선택이 목표일 때
- P개의 변수 중 일부분만 선정하여 적합함
- 모든 가능한 회귀, 전진 선택법, 후진 소거법, 단계적 선택법

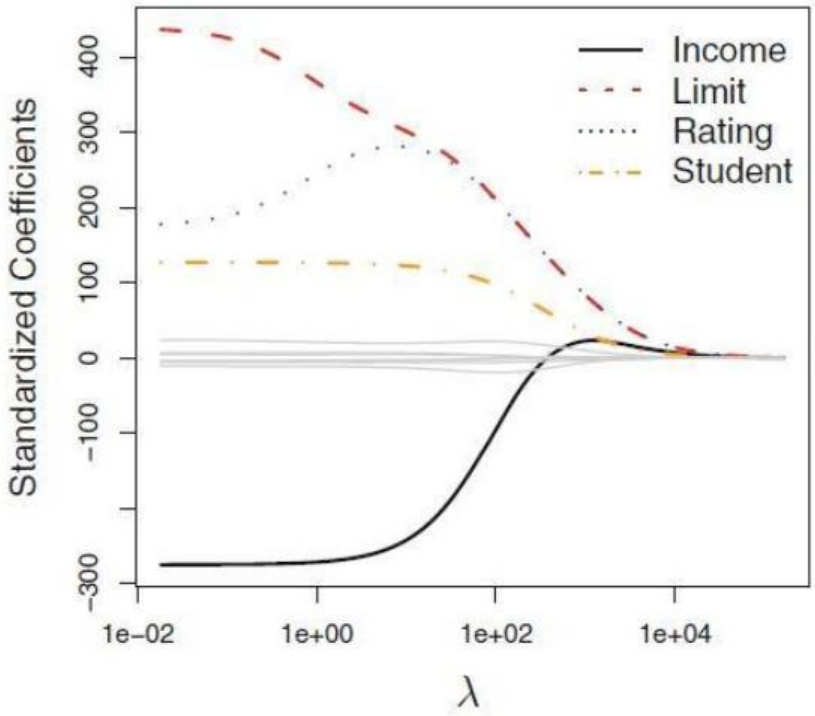
고차원 자료에서도 손실함수의 해가 정의되고 분산을 안정화시킬 수 있는 방법
예측 성능을 높이면서도 자동적으로 소수의 변수만 모형에 포함 할 수 있는 방법
→ lasso, ridge

정규화(Regularization)

- 모델이 복잡해질수록 penalty를 크게 하고자 목적함수에 항을 하나 더 추가
- 과적합된 모델을 일반성을 갖추도록 하기 위하여 사용

Ridge Regression	Lasso Regression	ElasticNet Regression

Ridge Regression(L2 Regression)



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

- λ ↑ ➔ 계수를 많이 줄이는데 집중
- λ ↓ ➔ 기존 최소 제곱법 문제
- β^2을 사용하기 때문에 완전히 0으로 수렴하지 X

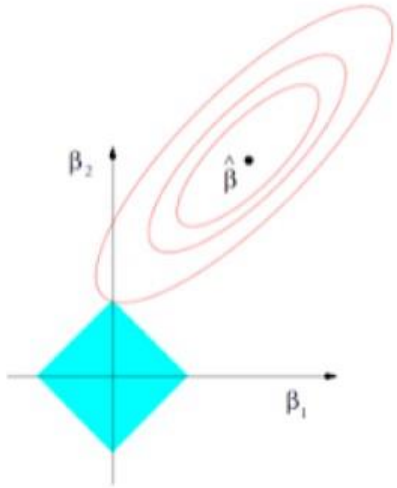
✓ 변수의 크기가 결과에 큰 영향을 미치기 때문에,
변수를 스케일링을 해주는 작업이 필요할 수 있다.

https://untitledtblog.tistory.com/96

Lasso Regression(L1 Regression)

- Rigde Regression과 다른점은 패널티 항에 절대값의 합을 주었다는 것!

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$



- 최적값은 모서리 부분에서 나타날 확률이 릿지에 비해 높아 몇몇 유의미하지 않은 변수들에 대해 계수를 0에 가깝게 추정
- 작은 값의 파라미터를 0으로 만들어 해당 변수를 삭제한다는 점이 차이점

Contents

Unit 03 | 선형회귀: 로지스틱 회귀

선형회귀분석을 분류에 사용하면 안되나?

Y를 종속변수, X를 독립변수로 두고 선형회귀분석을 하여 $\hat{Y} > 0.5$ 이면 yes로 분류하면 되지 않을까?
선형 판별 분석이 이런 작업을 잘 수행한다.
$E(Y|X = x) = P(Y = 1|X = x)$이므로, 이진 분류인 경우에는 선형회귀 분석으로 충분히 가능.

But 선형회귀 분석은 적합값으로, 0보다 작거나 1보다 큰 값을 내놓을 수도 있음

또한 오차항의 정규성과 등분산성 가정도 깨지게 되어, 통계적 추론을 위한 가정도 위배가 된다.

이진형이 아닌 다진형 Y에 대해서는 적절한 활동법도 아님.

Unit 03 | 선형회귀: 로지스틱 회귀

Logistic regression 모델

모형계수 (β, β_0) 의 추정에는 최대가능도법 (maximum likelihood) 사용
새로운 feature X에 대하여 로짓모형은 먼저 Y=1일 조건부 확률을 추정함:
분류함수 C(X)의 추정과 Y의 예측: 사건 지정된 cutoff value C에 기반하여 최종 분류함

Estimating coefficients

Contents

회귀진단 (Regression Diagnostics)

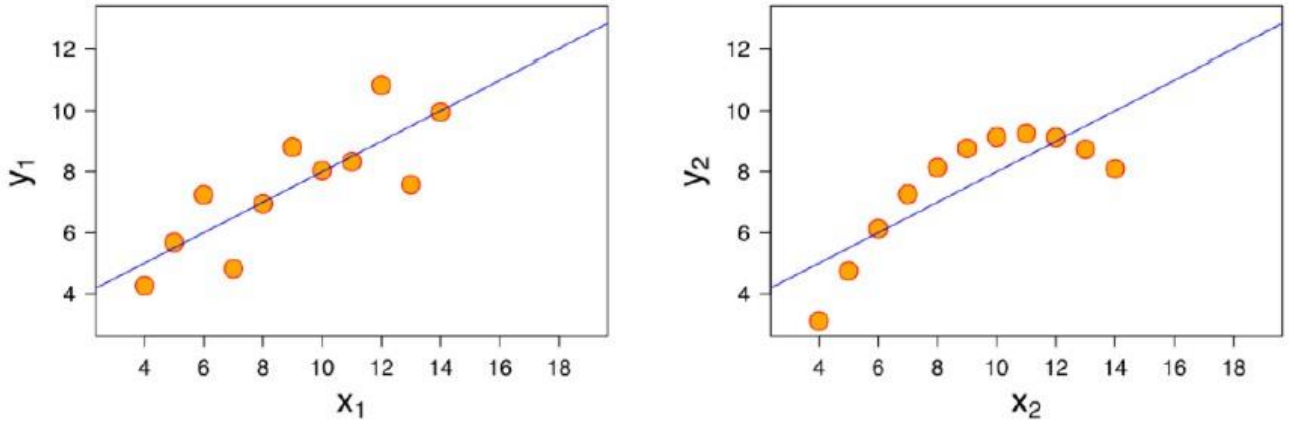
1) 회귀모형의 가정이 타당한가?
2) 각각의 관측값이 모형 및 가정에 어떠한 영향을 미치는가?

[회귀모형 기본 가정]

1. 선형성(Linearity) : 설명변수(X)와 반응변수(Y) 간 선형관계
2. 정규성(Normality) : 오차(Error)의 정규성
3. 등분산성(Homoscedasticity) : 오차의 등분산성
4. 독립성(Independence) : 오차의 독립성

그래프적 방법

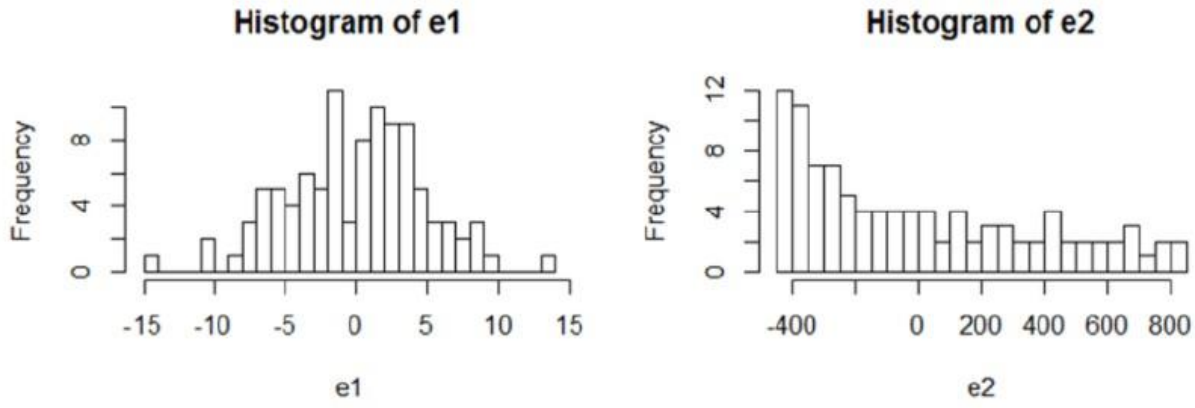
1. 선형성(설명변수와 반응변수 간 선형관계) 판단



- 산점도(Scatter plot)을 통해 선형성 판단 가능
- x1과 y1 간에는 선형관계가 존재하지만, x2와 y2 사이엔 선형관계가 있다고 보기 어려움

그래프적 방법

2. 정규성(오차가 정규분포를 따르는지) 판단

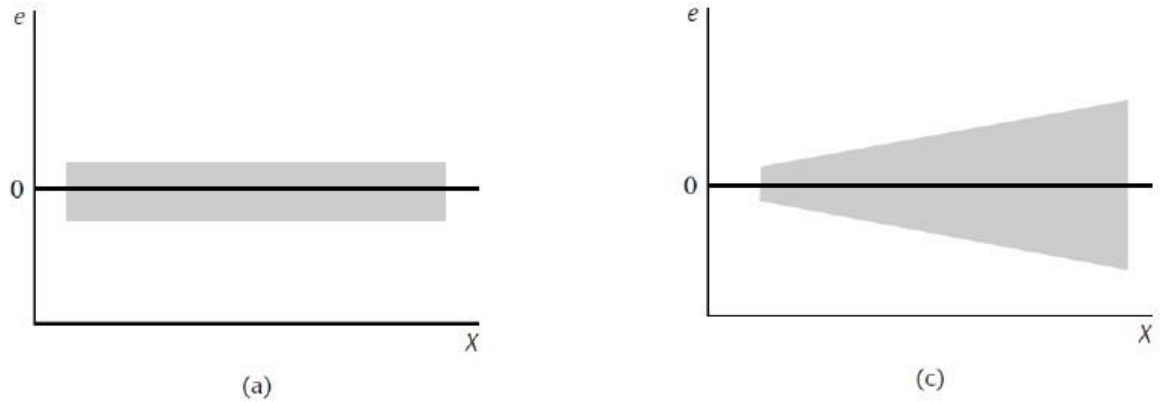


✓ 용어정리
모수 → 오차
표본 → 잔차

- 잔차의 히스토그램 ➔ 오차의 정규성 파악 가능
- e1은 정규성 가정을 만족하고, e2은 정규성 가정을 위배한다고 볼 수 있음
- [R] Shaprio-Wilk Normality Test

그래프적 방법

3. 등분산성(오차의 분산이 일정한지) 판단



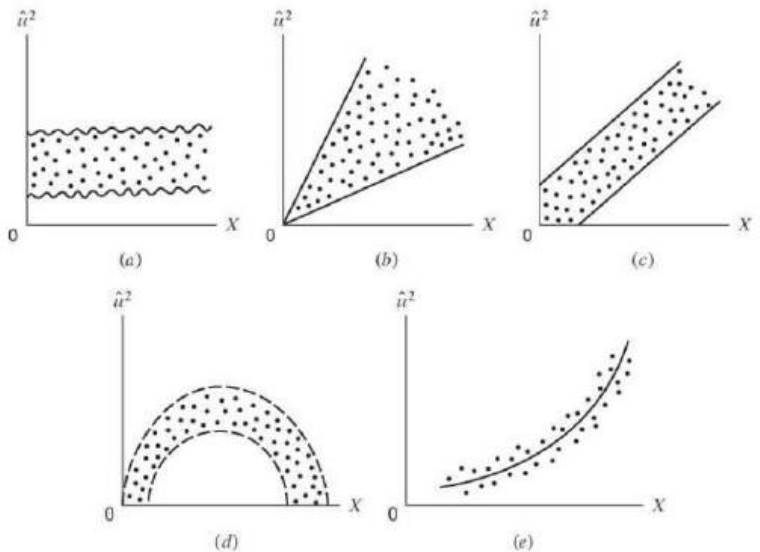
설명변수에 대한 그림으로 오차의 등분산성 판단 가능

band width가 일정해서
등분산성 가정 만족

band width가 넓어져서
등분산성 가정을 만족 X

그래프적 방법

4. 독립성(오차가 서로 독립인지) 판단



- 설명변수와의 상관성과 자기상관성 확인 ➔ 독립성 판단
- 잔차에 패턴이 존재한다면 독립적이지 않음
- a : 잔차에 어떤 패턴도 X ➔ 독립적이라 판단 가능
- Durbin-Watson 검정, ACF

OLS : Ordinary Least Square, 최소자승법

Python내의 statsmodel package – OLS class

- R-squared & Adj. R-squared
- F-statistic
- t-statistic
- Durbin-Watson(오차의 자기상관)

OLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.741
Model:	OLS	Adj. R-squared:	0.734
Method:	Least Squares	F-statistic:	108.1
Date:	Mon, 18 Nov 2019	Prob (F-statistic):	6.72e-135
Time:	21:54:23	Log-Likelihood:	-1498.8
No. Observations:	506	AIC:	3026.
Df Residuals:	492	BIC:	3085.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	36.4595	5.103	7.144	0.000	26.432	46.487
CRIM	-0.1080	0.033	-3.287	0.001	-0.173	-0.043
ZN	0.0464	0.014	3.382	0.001	0.019	0.073
INDUS	0.0206	0.061	0.334	0.738	-0.100	0.141
CHAS	2.6867	0.862	3.118	0.002	0.994	4.380
NOX	-17.7666	3.820	-4.651	0.000	-25.272	-10.262
RM	3.8099	0.418	9.116	0.000	2.989	4.631
AGE	0.0007	0.013	0.052	0.958	-0.025	0.027
DIS	-1.4756	0.199	-7.398	0.000	-1.867	-1.084
RAD	0.3060	0.066	4.613	0.000	0.176	0.436
TAX	-0.0123	0.004	-3.280	0.001	-0.020	-0.005
PTRATIO	-0.9527	0.131	-7.283	0.000	-1.210	-0.696
B	0.0093	0.003	3.467	0.001	0.004	0.015
LSTAT	-0.5248	0.051	-10.347	0.000	-0.624	-0.425

Omnibus:	178.041	Durbin-Watson:	1.078
Prob(Omnibus):	0.000	Jarque-Bera (JB):	783.126
Skew:	1.521	Prob(JB):	8.84e-171
Kurtosis:	8.281	Cond. No.	1.51e+04

Partition of Sum of Squares(제곱합 분해)

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

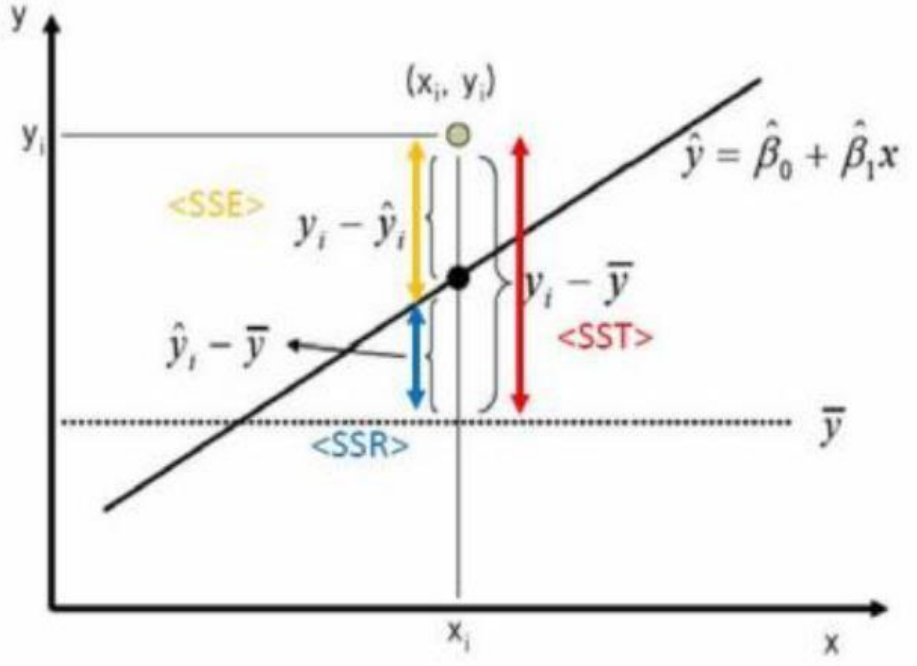
<SST> <SSR> <SSE>

SST : 총 제곱합

SSR : 회귀 제곱합 (전체 제곱합 중 회귀식으로 설명가능)

SSE : 잔차 제곱합 (전체 제곱합 중 회귀식으로 설명불가)

→ 회귀식이 데이터를 잘 나타낼수록 SSR↑ SSE↓

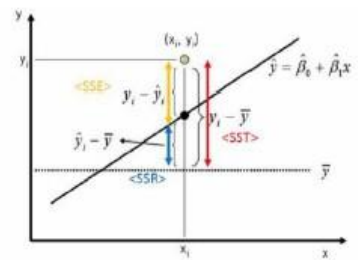


1. 결정계수(R-squared) & 조정된 결정계수 (Adj. R-squared)

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$adj\ R^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SST}{n-1}}$$

전체 제곱합(SST) 중 회귀식으로 설명 가능한 부분
→ 결정계수(R^2)은 클수록 좋다 !



SST : 총 제곱합
SSR : 회귀 제곱합
SSE : 잔차 제곱합

결정계수의 단점 : 설명변수를 추가하면 항상 SSR이 커지기 때문에 R^2가 증가

→ $\boldsymbol{Adjusted\ R^2}$: 설명변수가 증가하면 값이 감소하도록 패널티 부과

2. F-Statistics

OLS Regression Results

Dep. Variable:	MEDV	R-squared:	0.741
Model:	OLS	Adj. R-squared:	0.734
Method:	Least Squares	F-statistic:	108.1
Date:	Mon, 18 Nov 2019	Prob (F-statistic):	6.72e-135
Time:	21:54:23	Log-Likelihood:	-1498.8
No. Observations:	506	AIC:	3026.
Df Residuals:	492	BIC:	3085.
Df Model:	13		
Covariance Type:	nonrobust		

$$F = \frac{\text{표본 평균 간 변동}}{\text{표본 내 변동}}$$

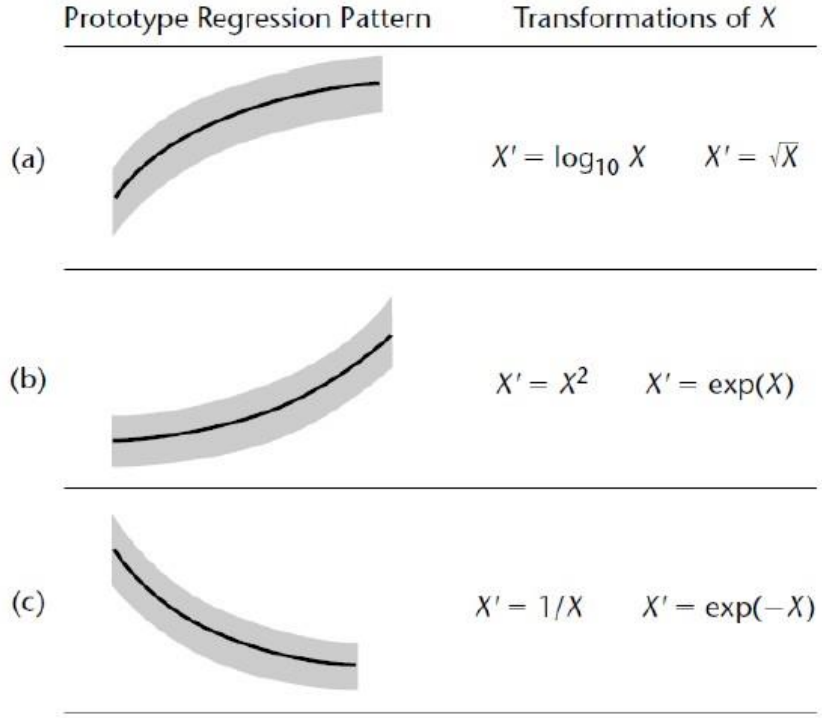
- 귀무가설(H_0) : $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ VS 대립가설(H_1) : $\beta_j \neq 0$, for some j
- 모형 자체의 유의미함을 판단하는 기준
- 모든 독립변수의 계수가 0인지 혹은 하나라도 0이 아닌지를 판별

4. Durbin-Watson (오차의 자기상관)

Omnibus:	178.041	Durbin-Watson:	1.078
Prob(Omnibus):	0.000	Jarque-Bera (JB):	783.126
Skew:	1.521	Prob(JB):	8.84e-171
Kurtosis:	8.281	Cond. No.	1.51e+04

- 오차의 독립성을 검정하기 위한 방법
- 더빈 왓슨 통계량은 0~4의 값을 가짐
  - 0에 가까울수록 : 잔차끼리 양의 상관관계를 가진다.
  - 4에 가까울수록 : 잔차끼리 음의 상관관계를 가진다.
  - 2에 가까울수록 : 오차항의 자기상관이 없음 (=독립성 만족)

변수변환(Transformation)



- 비선형적인 함수 관계를 선형으로 바꿔 다룰 수 있다.
- 분포 모양을 정규분포와 유사하도록 만들 수 있다.
- 변환을 통해 자기상관 문제를 해결할 수 있다.

다중공선성(Multicollinearity) 제거

제거 이유

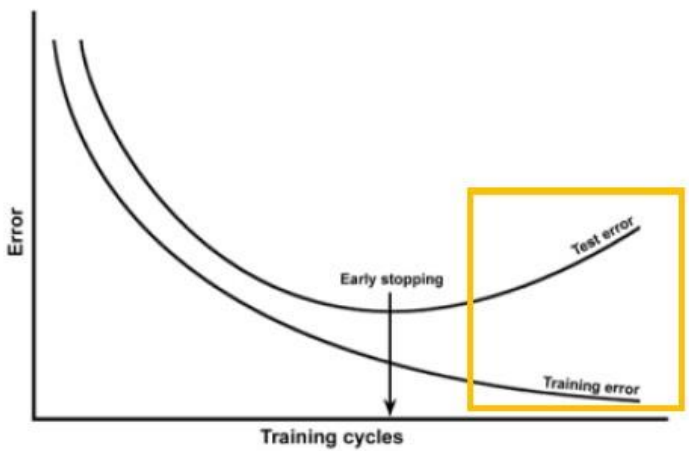
1. 설명변수 간 독립적이지 않으면 회귀계수의 추정이 불안정함.
2. 추정값이 존재하지 않거나, 추정값의 분산이 커지는 문제점을 가져올 수 있음.

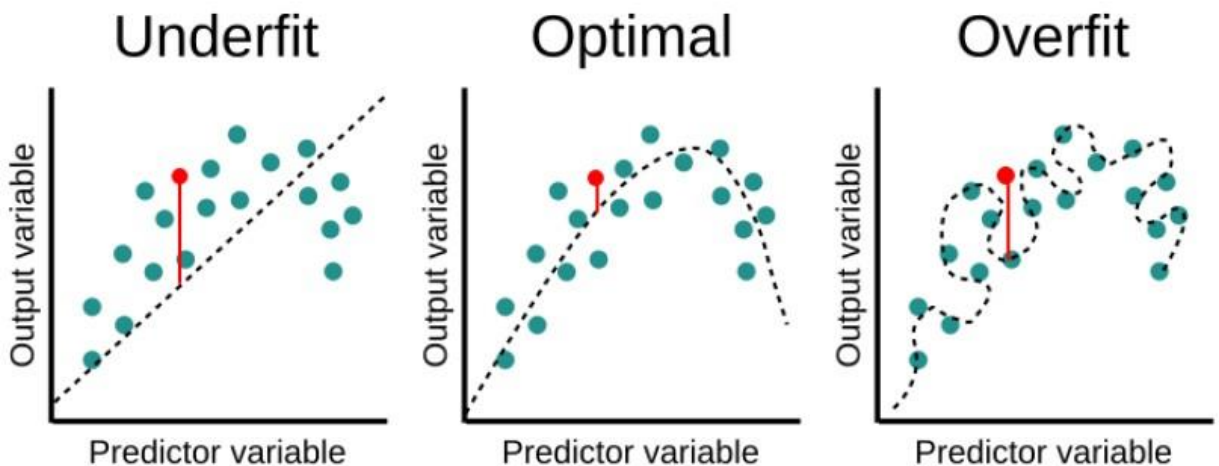
제거 방법

1. 더 많은 데이터 수집
2. 다중공선성을 유발하는 주요 변수 2개를 찾아, 각 변수 제거시 R-squared의 변동 확인하여 제거해도 상관계수가 유지되는 변수 제거
3. PCA(Principal Component Analysis, 주성분 분석) ➔ 차원 축소
4. Ridge / Lasso Regression

과적합(Overfitting)

- 학습 데이터에 과하게 학습하여 실제 데이터에 대한 오차가 증가하는 현상





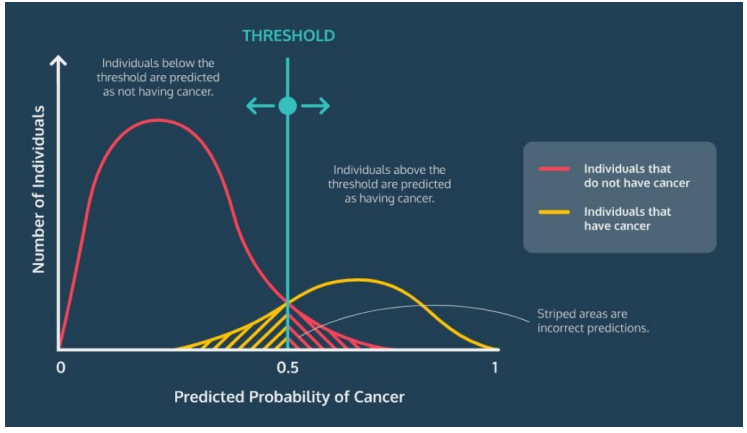
선형 회귀분석 마무리

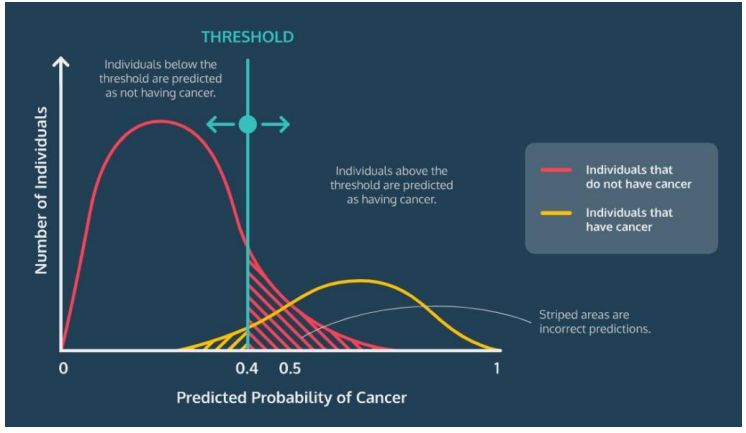
1. 회귀모형 설정 : 반응변수 및 주요 설명변수 파악
2. 선형성 검토 : 산점도를 통해 상관관계 파악
3. 설명변수 검토 : 각 설명변수 분포 확인 + 다중공선성 점검
4. 모델 적합 : 모델 회귀계수 추정 및 모형 적절성 검토
5. 변수 선택 : 주요 설명변수 선택
6. 적합된 모형 검토 : 오차에 대한 기본 가정 확인
7. 최종 모형 선택

Contents

Cutoff (Threshold)

- 분류(Classification)를 위한 기준
- 로지스틱 함수로 구한 확률이 cutoff 이상이면 1, cutoff 이하이면 0으로 분류
- Cutoff을 조정하여 성능 조절 가능





Model Evaluation(1)

Accuracy

		예측결과	
		Positive	**False**
실제값	**Positive**	True Positive (TP)	False Negative (FN)
	False	False Positive (FP)	True Negative (TN)

※ True : 옳은 예측(정답) / False : 틀린 예측(오답)

- 예측결과가 True일 때, 실제값도 True인 것
- 실제 분포가 편향 되어 있는 경우엔 적합하지 않음

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN}$$

- 웹사이트 판매량 데이터
- 학습 데이터 : 99% 물건 사지 않음 (Y=0), 1% 물건 구매(Y=1)
  - ➔ 실제 데이터와 무관하게 Y=0이라고 예측할 확률이 높아짐
  - ➔ 99%의 정확도를 가지기에 좋은 결과처럼 보임

Model Evaluation(2)

		예측결과	
		Positive	**False**
실제값	**Positive**	True Positive (TP)	False Negative (FN)
	False	False Positive (FP)	True Negative (TN)

※ True : 옳은 예측(정답) / False : 틀린 예측(오답)

Precision(정밀도)

- 모델이 True로 분류한 것 중에서 실제값이 True인 비율

$$\text{Precision} = \frac{TP}{TP + FP}$$

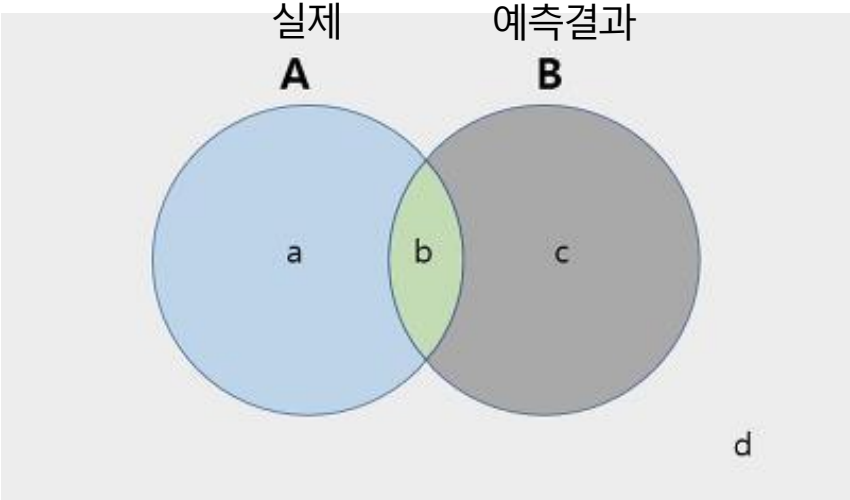
Recall(재현율)

- Sensitivity
- 실제 True인 것 중에서 모델이 True라고 분류한 것의 비율

$$\text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

(cf) Precision과 Recall은 Trade-Off 관계

* 날씨 예측(맑다 / 흐리다) 모델

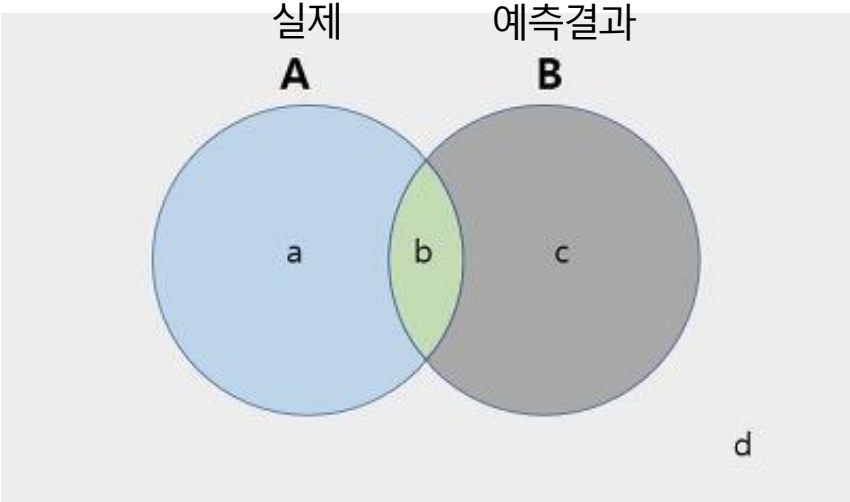


- A : 실제 날씨가 맑은 날
- B : 모델이 날씨가 맑다고 예측(분류)한 날
- b = TP = 실제 날씨가 맑은 날을 모델이 날씨가 맑다고 예측(제대로 예측)한 날

- 이때,
  - ✓ Precision = $\frac{b}{b+c}$
  - ✓ Recall = $\frac{b}{a+b}$
  - ✓ a의 영역이 줄어들면 c의 영역이 커지게 됨 = 두 지표 간 Trade-off 관계

(cf) Precision과 Recall은 Trade-Off 관계

* 날씨 예측(맑다 / 흐리다) 모델



		실제 정답	
		True	False
분류 결과	True	TP(20)	FP(40)
	False	FN(30)	TN(10)

$\text{Precision} = \frac{20}{60} = 33.3\%$

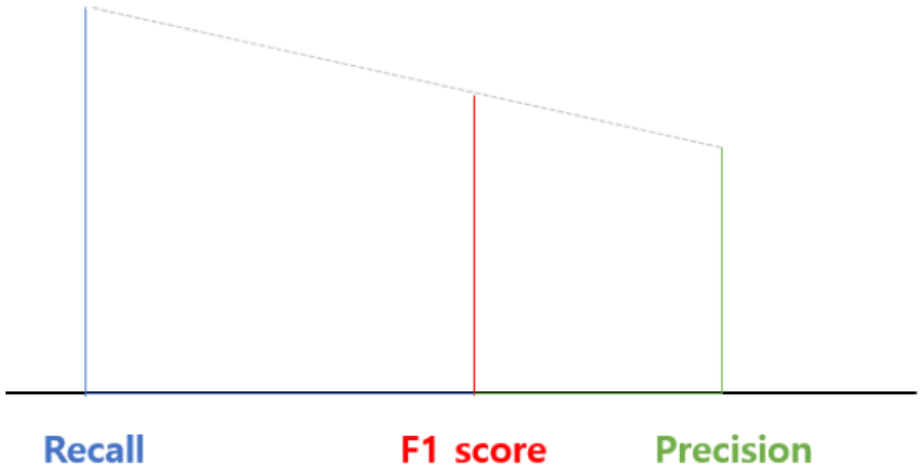
$\text{Recall} = \frac{20}{50} = 40\%$

		실제 정답	
		True	False
분류 결과	True	TP(20)	FP(80)

$\text{Precision} = \frac{20}{100} = 20\%$

$\text{Recall} = \frac{20}{20} = 100\%$

Model Evaluation(3)



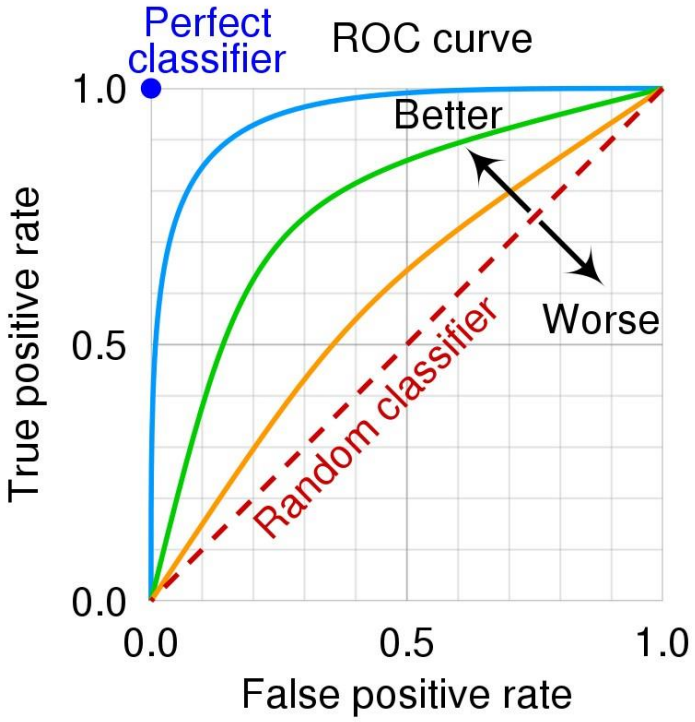
F1 Score

- Precision과 Recall의 조화 평균

$$\text{F1 Score} = 2 \times \frac{1}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- 불균형한 데이터에서 잘 동작

Model Evaluation(4)



ROC Curve

- Confusion Matrix에서 FPR, Recall(Sensitivity) 값 계산
- 그래프가 좌 상단에 위치할수록 좋은 모델
  - FPR = $\frac{FP}{FP+TN}$ (False Positive)
- AUC(=Area Under Curve) : ROC Curve 아래 면적
  - ✓ 1에 가까울 수록 좋은 모델

과제

[과제 1]

- LSE normal equation, MSE 구현

[과제 2] 회귀분석 – Used Car Priced Prediction

- Ch 1, Ch 2를 토대로 자유롭게 회귀분석 & 회귀진단 진행
- 주석으로 설명 및 근거 자세하게 달아주세요 ☺

[과제 3] 로지스틱 회귀분석 – Credit Card Fraud Detection

- 파이썬 sklearn 패키지를 활용해 로지스틱 회귀분석 진행
- 성능지표 계산 및 해석
  - Sklearn ➔ mean accuracy, f1 score 등
  - confusion matrix ➔ tp, fp, fn, tn 값
- 성능 개선 시도 (어떤 성능지표를 기준으로 했는지, 해당 지표 선택 이유 등)
- 주석으로 설명 및 근거 자세하게 달아주세요 ☺

참고

[강의안]
투빅스 14기 강재영님 강의안
투빅스 15기 장아연님 강의안
투빅스 16기 이예림님 강의안
연세대학교 응용통계학과 김현태 교수님<회귀분석> 강의안

[교재]
Michael H. Kutner,Christopher J. Nachtsheim, John Neter, <Applied Linear Regression Models>

[참고자료]
[선형, 로지스틱] 데이터사이언스스쿨4장, 6장(https://datascienceschool.net/intro.html)
[로지스틱] https://ratsgo.github.io/machine%20learning/2017/04/02/logistic/
[Ridge/Lasso Regression] Ridge regression(능형회귀) 간단한설명과장점(tistory.com)
[회귀진단] Regression(03) -회귀진단| DataLatte'sIT Blog (heung-bae-lee.github.io)
+) 숙명여대 통계학과 최영근 교수님 데이터 마이닝 강의 https://youtu.be/6Pm9dtECFrs

감사합니다

문의: 18기 김희경

ppt 제작 : 김희경
ppt 테마 : 투빅스 정규세션