Signal & Noise II

Classification

D:plus

목차

Statistics in Regression에 이어서 수강하기를 권장합니다.

1. Classification [Analysis]

1. Classification [Analysis]

Linear Regression를 알아야 부드러움!

$$Y = f(X) + \epsilon \qquad \hat{Y} = \hat{f}(X)$$

가정/가설로 시작 : X의 선형적인 조합에, 어떤 경계선을 기준으로 Y값이 바뀌지 않을까? -> X의 선형 결합에 Logistic함수를 씌우자!

Chk1. 선형 결합이 뭐지?
Chk2. Logistic 함수가 뭐지?
Chk3. 그래서 전체의 의미가 뭐지?

D:plus

1. Classification [Analysis]

로지스틱 리그레션! : 손으로 적어보자.

모델의 구조를 그려서 비교해보자.

로지스틱 회귀 (수식)

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

$$y_i = f(X) + e_i = g(z_i) = g(w_1X_{i1} + w_2X_{i2} + \cdots + w_pX_{ip} + w_0) + e_i$$
$$= \frac{1}{1 + e^{-(w_1X_{i1} + w_2X_{i2} + \cdots + w_pX_{ip} + w_0)}} + e_i$$

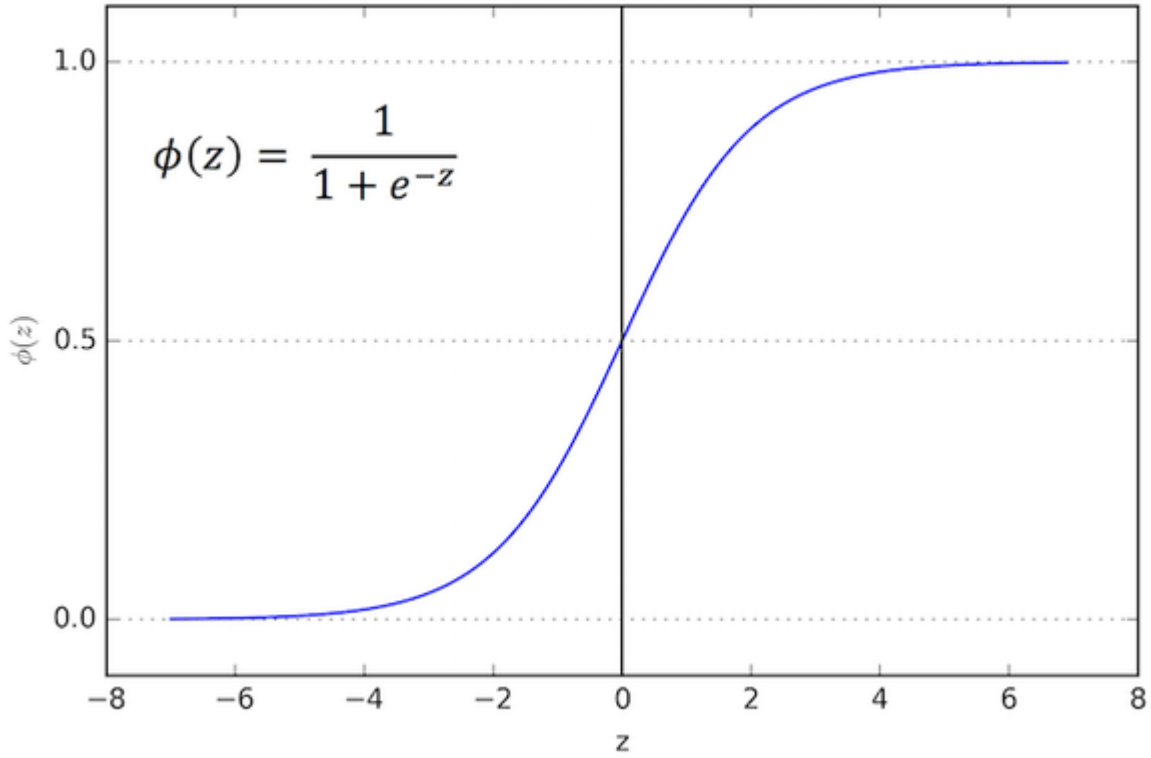
$$\hat{Y} = \hat{f}(X)$$

로지스틱 회귀 (그래프)

D:plus

1. Classification [Analysis]

결국 끝은 똑같은 로지스틱 함수인데..



장점이 있다!

Probability theory is a mathematical framework for representing uncertain statements.

Deep Learning, MIT Press.
Ian Goodfellow, Yoshua Bengio, Aaron Courville

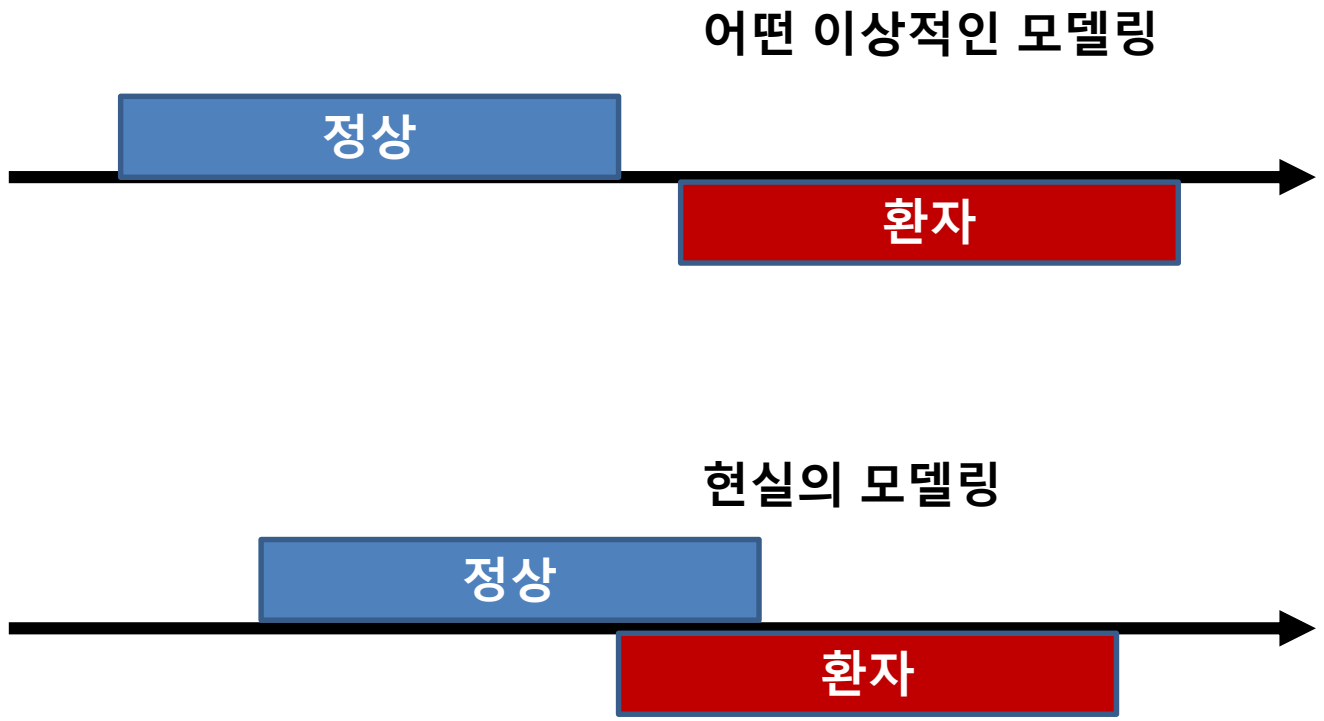
D:plus

1. Classification [Analysis]

에러를 보는 관점은 더 있다

50명의 정상인과 50명의 환자.

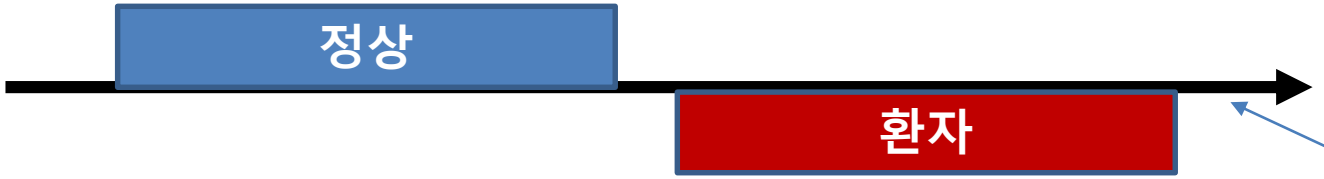
	나이	몸무게	어떤 특징	.....	질환 여부
사람1	15	..			O
사람2	23	..			O
사람3	32				X
사람4	15				X
사람5	35				O



D:plus

1. Classification [Analysis]

Confusion Matrix



	실제 환자	실제 정상
환자일거야	50	0
정상일거야	0	50

너무나 완벽! 분류 모델링이란 것은
정상과 환자를 온전히 분리해낼 수 있는
어떤 특성을 잡아내는 것이다

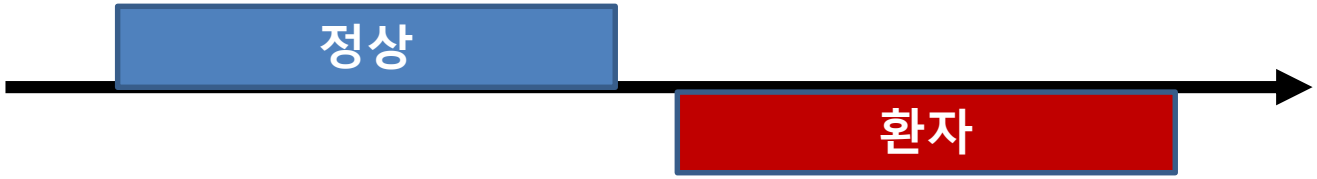
생각해봅시다.

정확도는?

D:plus

1. Classification [Analysis]

Confusion Matrix

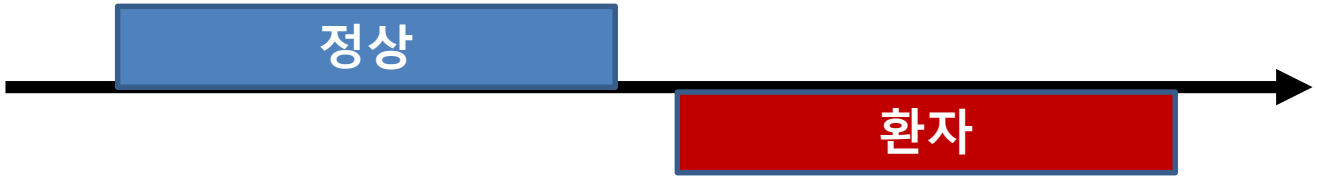


	실제 환자	실제 정상
환자일거야	50	0
정상일거야	0	50

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

1. Classification [Analysis]

Confusion Matrix



Wikipedia

	실제 환자	실제 정상
환자일거야	50	0
정상일거야	0	50

몇 개만 기억해보자.

(True/False) & (Positive/Negative) 구별 법

True Positive Rate = Sensitivity = Recall 의 의미

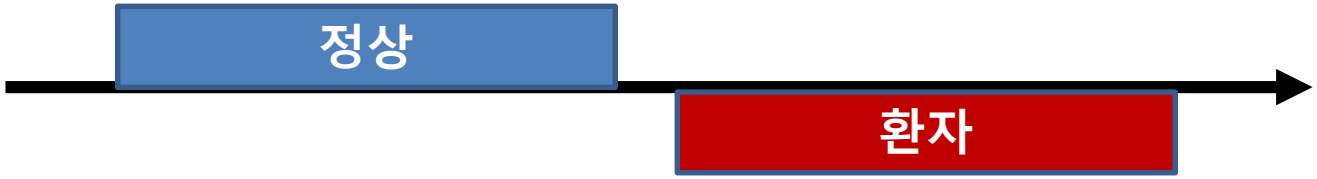
True Negative Rate = specificity 의 의미

Precision 의 의미

D:plus

1. Classification [Analysis]

Confusion Matrix



Wikipedia

	실제 환자	실제 정상
환자일거야	50	0
정상일거야	0	50

몇 개만 기억해보자.

(True/False) & (Positive/Negative) 구별 법

True Positive Rate = Sensitivity = Recall 의 의미

참을 참이라 판단한 비율

True Negative Rate = specificity 의 의미

거짓을 거짓이라 판단한 비율

Precision 의 의미

참이라 판단한 것 중 참인 것의 비율

D:plus

1. Classification [Analysis]

Confusion Matrix

10	10	10	10	10			
			10	10	10	10	10

기준선1	실제 환자	실제 정상
환자일거야		
정상일거야		

True Positive Rate :
False Positivie Rate :

기준선2	실제 환자	실제 정상
환자일거야		
정상일거야		

True Positive Rate :
False Positivie Rate :

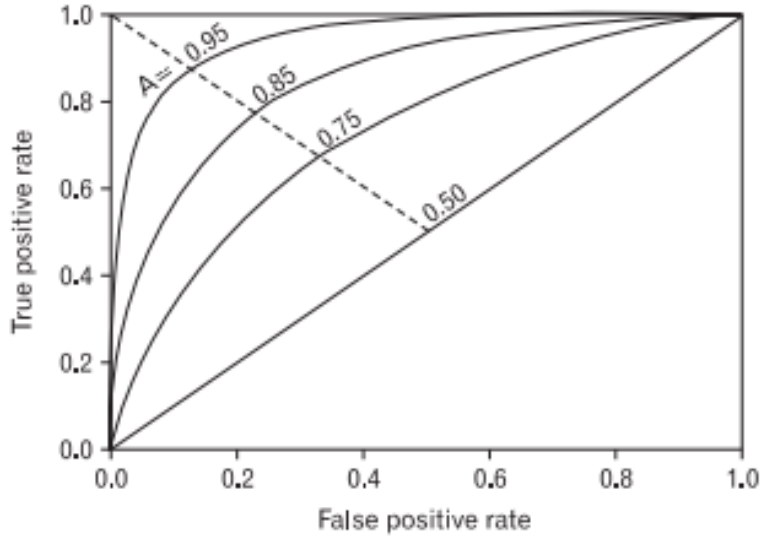
기준선3	실제 환자	실제 정상
환자일거야		
정상일거야		

True Positive Rate :
False Positivie Rate :

D:plus

1. Classification [Analysis]

Confusion Matrix -> ROC Curve & AUC ?



기준1
True Positive Rate : 1
False Positivie Rate : 0.4

기준2
True Positive Rate : 0.8
False Positivie Rate : 0.2

기준3
True Positive Rate : 0.6
False Positivie Rate : 0

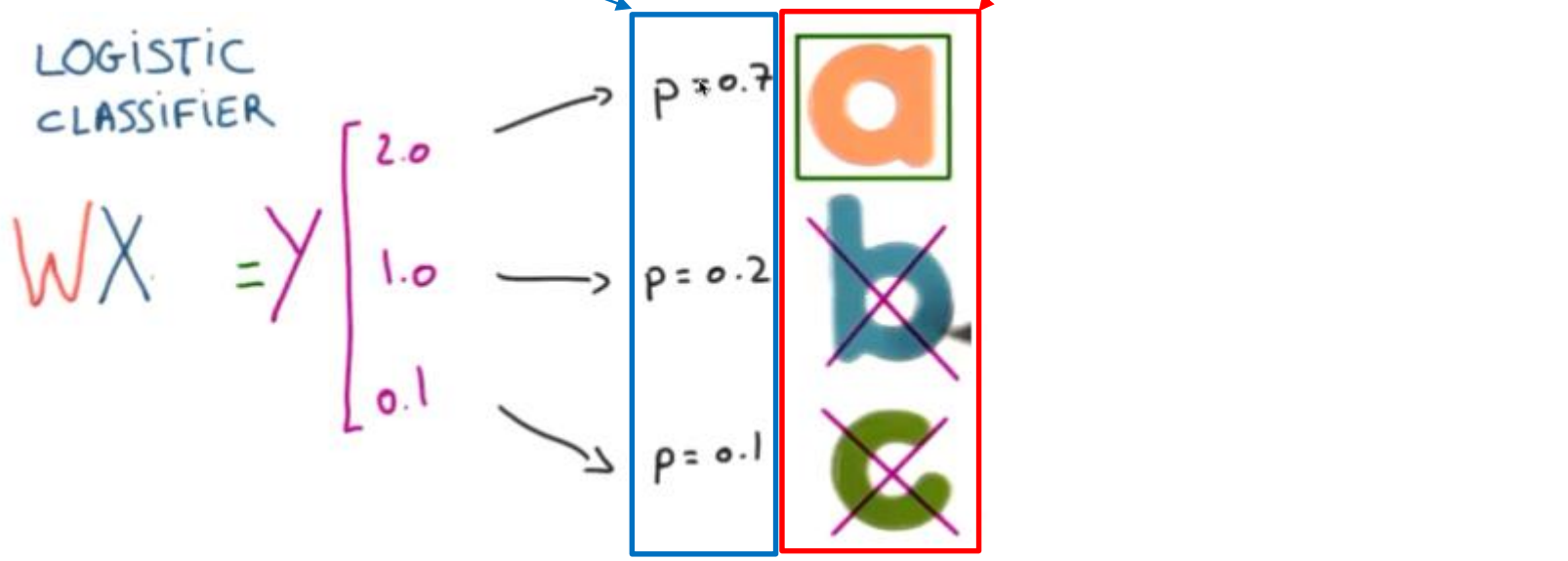
2. Cost Function 뜯어보기

2. Cost Function 뜯어보기

분류문제에서 에러를 바라보는 관점

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

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



2. Cost Function 뜯어보기

두 확률 분포의 차이를 계산할 때는 상대 엔트로피를 계산한다고 하는데....

엔트로피?

Probability theory is a mathematical framework for representing uncertain statements.

Deep Learning, MIT Press.
Ian Goodfellow, Yoshua Bengio, Aaron Courville

Information theory enables us to quantify the amount of uncertainty in a probability distribution

Deep Learning, MIT Press.
Ian Goodfellow, Yoshua Bengio, Aaron Courville

D:plus

2. Cost Function 뜯어보기

두 확률 분포의 차이를 계산할 때는 상대 엔트로피를 계산한다고 하는데….

정보이론..?

1. 자주 발생하는 사건은 낮은 정보량을 가진다.
2. 덜 자주 발생하는 사건은 더 높은 정보량을 가진다.

정보량…?!

$$I(x) = -\log P(x)$$

동전을 던져 앞면이 나오는 사건의 정보량은? $-\log_2 0.5 = 1$

주사위를 던져 3이 나오는 사건의 정보량은? $-\log_2 1/6 = 2.5849$

밑이2인 경우 정보량의 단위를 Shannon 또는 bit라고 함

D:plus

2. Cost Function 뜯어보기

두 확률 분포의 차이를 계산할 때는 상대 엔트로피를 계산한다고 하는데....

어떤 확률분포 P에 대한 섀넌 엔트로피 : 모든 사건 정보량의 기대값 혹은 평균

$$H(P) = H(x) = -E_{X\sim P}[I(x)] = E[-logP(x)]$$
$$= \sum_x \big(-logP(x)\big)P(x)$$

D:plus

2. Cost Function 뜯어보기

두 확률 분포의 차이를 계산할 때는 상대 엔트로피를 계산한다고 하는데....

수식을 뜯어보자!

$$H(P) = H(x) = -E_{X\sim P}[I(x)] = E[-logP(x)]$$
$$= \sum_x \big(-logP(x)\big)P(x)$$

당첨금	빈도	확률
-1000원	80	0.8
100원	12	0.12
200원	6	0.06
400원	2	0.2

평균을 구해보고 이해해보자. 그를 통해, 정보량의 평균이란 것을 이해해보자.

D:plus

2. Cost Function 뜯어보기

두 확률 분포의 차이를 계산할 때는 상대 엔트로피를 계산한다고 하는데....

말을 바꾸어 이해해보자.

$$H(P) = H(x) = -E_{X\sim P}[I(x)] = E[-logP(x)]$$
$$= \sum_x (-logP(x))P(x)$$

이 분포를 기준으로 (이 분포가 진짜라고 했을 때)

관찰된 정보량의 평균

이 경우는 관찰된 것도 기준도 전부 동일한 분포.

D:plus

2. Cost Function 뜯어보기

두 확률 분포의 차이를 계산할 때는 상대 엔트로피를 계산한다고 하는데....

그렇다면, 이런 것을 이해해볼 수도 있다.

$$H(P,Q) = \sum_{x} (-logQ(x))P(x)$$

이 분포를 기준으로 (이 분포가 진짜라고 했을 때)

관찰된 정보량의 평균

D:plus

2. Cost Function 뜯어보기

두 확률 분포의 차이를 계산할 때는 상대 엔트로피를 계산한다고 하는데....

상대 엔트로피는 다음과 같게 될 것이다.

$$D_{KL}(P||Q) = H(P,Q) - H(P)$$

$$= \sum_x \big(-logQ(x)\big)P(x) - \sum_x \big(-logP(x)\big)P(x)$$

실제 분포 P와, 관찰된 값으로부터 추측한 분포 Q가 같다면
P를 기준으로 Q의 정보량을 계산해도, P를 기준으로 P의 정보량을
계산해도 같은 값이 나올 것이다.

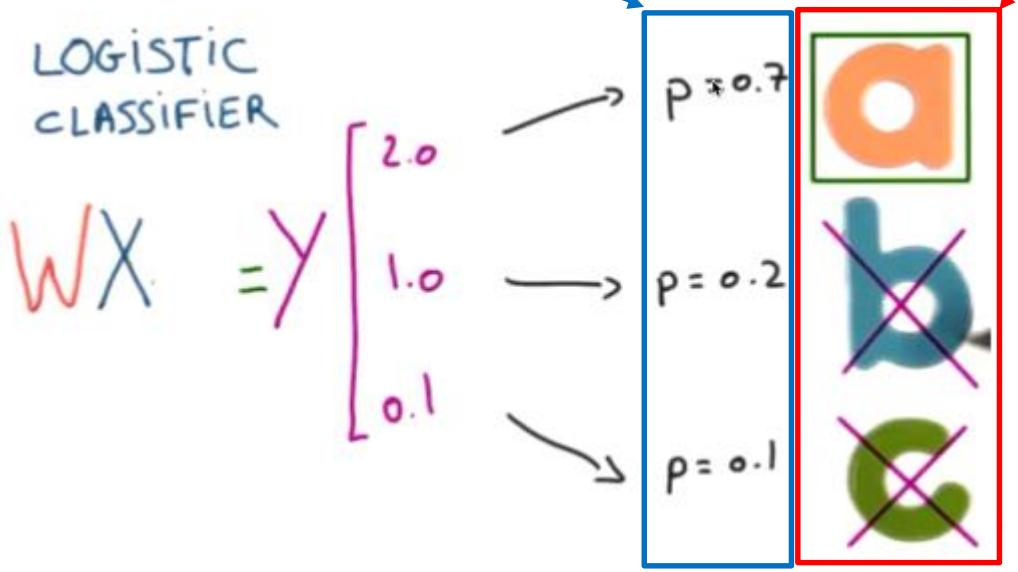
D:plus

2. Cost Function 뜯어보기

먼 길 왔지만 아직 이 이야기 하는 중.

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

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



여기는 사실 날아가버리고.

$$D_{KL}(P||Q) = H(P,Q) - H(P)$$

$$= \sum_x \big(-logQ(x)\big)P(x) - \sum_x \big(-logP(x)\big)P(x)$$

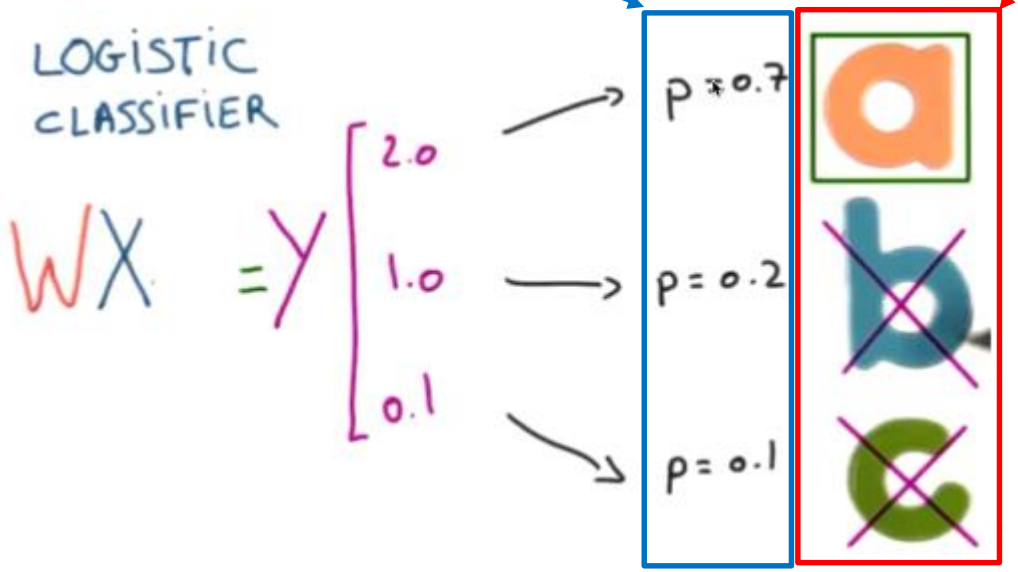
Cross Entropy만 남는다.

2. Cost Function 뜯어보기

그렇다. 분류 모델링 할 때는,
모델과 실제 사이의 에러를 크로스 엔트로피를 써도 되는 것이다!

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

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



$$H(P,Q) = \sum_{x} \big(-logQ(x)\big)P(x)$$

최적화 대상으로
얼마나 잘 편리하게 디자인 되어있는지 보자!

D:plus

2. Cost Function 뜯어보기

크로스 엔트로피를 줄인다는 뜻은 결국?

$$
\begin{aligned}
E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\
&= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}
\end{aligned}
$$

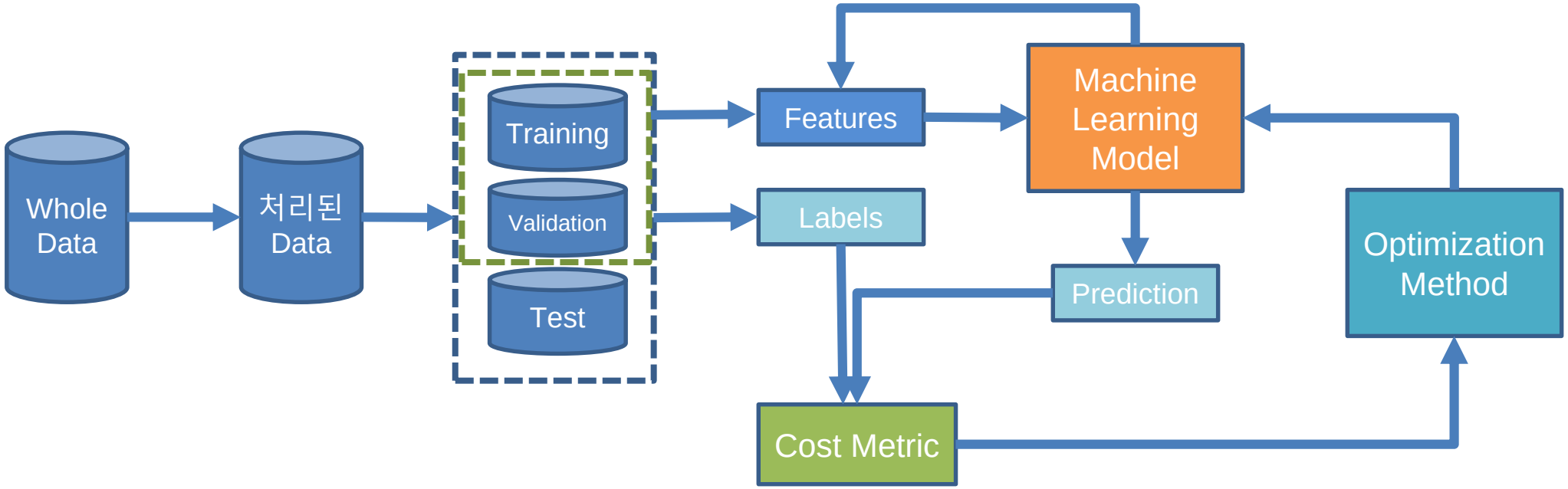
$$
H(P,Q) = \sum_{x} \big(-logQ(x)\big)P(x)
$$

D:plus

3. EDA for Feature Engineering

3. EDA for Feature Engineering

Phase 1 :
모델링을
위한 작업들

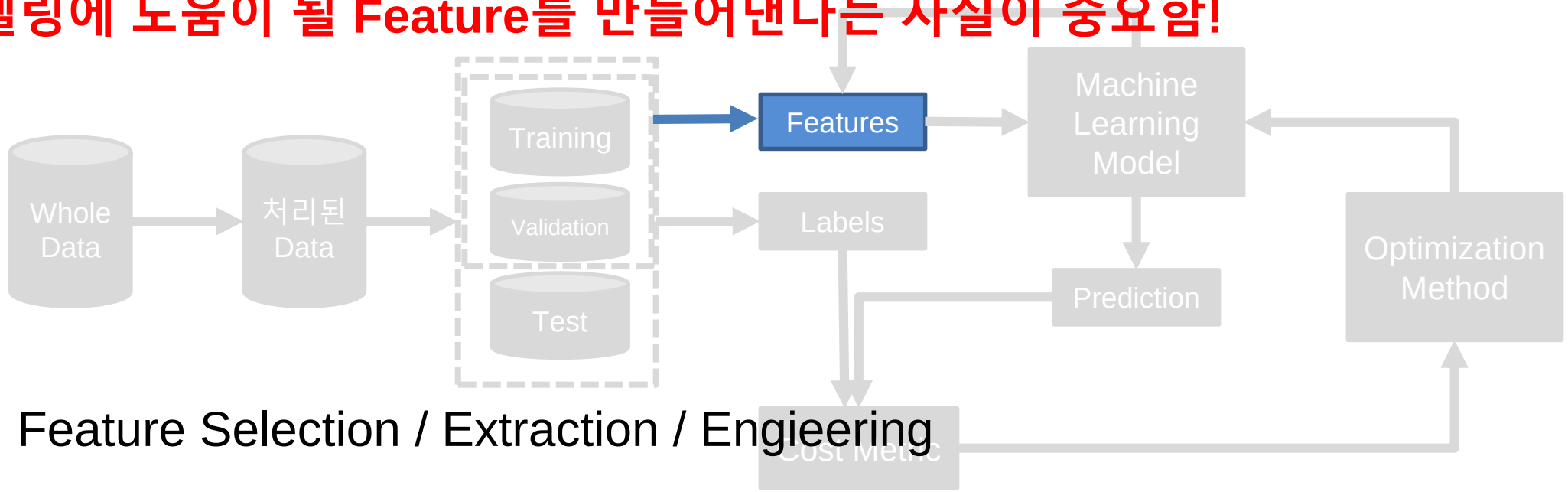


D:plus

3. EDA for Feature Engineering

Feature Engineering을 다른 것과 구별하는 것이 중요한 것이 아님.

모델링에 도움이 될 Feature를 만들어낸다는 사실이 중요함!



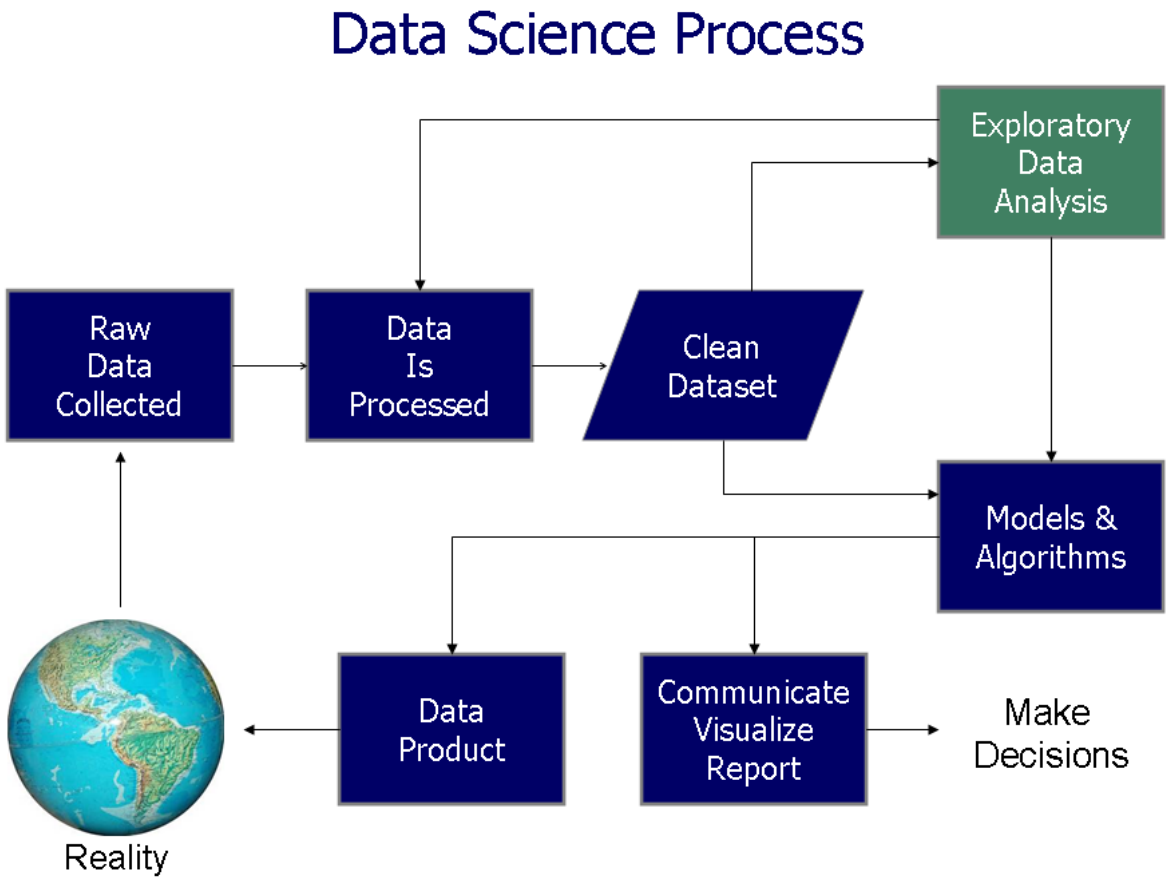
3. Feature Selection / Extraction / Engieering

Feature Selection : 사용할 것만, 중요한 것 위주로 선택 (인풋 데이터 형태 그대로)
Feature Extraction : 원본 데이터를 조합하여 새로운 Feature를 뽑아냄. (주로 차원에 관련됨)
Feature engineering : **좀더 현실적인 관점에서 필요한 특징들을 만들어냄**.

D:plus

3. EDA for Feature Engineering

EDA는 Machine Learning에 들어가기 전에 중요한 역할을 차지한다.



Wikipedia에서도
EDA 혼자만 색깔이 다름!

D:plus

3. EDA for Feature Engineering

EDA는 Machine Learning에 들어가기 전에 중요한 역할을 차지한다.

통계기법들 중 너무 가설 검정(CDA)만 강조되어 있다.

데이터를 이용해 검정할 가설을 만드는 것

그 것에 좀 더 **집중**할 필요가 있다.

탐색적인 데이터 분석을 해야 한다.
현상에 대한 **가설을 세우기** 위해.
가설 검정의 토대가 될 **가정들을 확인**하기 위해
올바른 통계 기법을 선택하기 위해
추가적인 데이터 수집의 기반을 닦기 위해

데이터 과학, 데이터 마이닝, 빅데이터 분석의 토대가 되는 기술!
어린 학생들에게 통계적 사고 방식을 가르칠 때 사용!

D:plus

3. EDA for Feature Engineering

추가적으로 무슨 데이터가 필요할까 상상해야 하는 시간이니까!

통계기법들 중 너무 가설 검정(CDA)만 강조되어 있다.

데이터를 이용해 검정할 가설을 만드는 것

그 것에 좀 더 **집중**할 필요가 있다.

탐색적인 데이터 분석을 해야 한다.
현상에 대한 **가설을 세우기** 위해.
가설 검정의 토대가 될 **가정들을 확인**하기 위해
올바른 통계 기법을 선택하기 위해
추가적인 데이터 수집의 기반을 닦기 위해

데이터 과학, 데이터 마이닝, 빅데이터 분석의 토대가 되는 기술!
어린 학생들에게 통계적 사고 방식을 가르칠 때 사용!

D:plus

3. EDA for Feature Engineering

이 시간은, 바로 토론과 실습으로 갑니다.

UserId	Age	Address	Gender	UserType	TransactionId	Timestamp	ItemId	Quantity	Value
2105345	D	F	Unknown	Unknown	1215553	12/31/2000 12:00:00 AM	4710040000000	1	149
2105345	D	F	Unknown	Unknown	1216545	12/31/2000 12:00:00 AM	4711090000000	1	179
2105345	D	F	Unknown	Unknown	1216590	12/31/2000 12:00:00 AM	9556000000000	1	28
2105345	D	F	Unknown	Unknown	1217249	12/31/2000 12:00:00 AM	4711800000000	1	199
2105345	D	F	Unknown	Unknown	1217259	12/31/2000 12:00:00 AM	4710030000000	1	139
2105345	D	F	Unknown	Unknown	1217263	12/31/2000 12:00:00 AM	4710190000000	2	10
2105345	D	F	Unknown	Unknown	1217322	12/31/2000 12:00:00 AM	4710630000000	1	33
2105345	D	F	Unknown	Unknown	1218254	12/31/2000 12:00:00 AM	4711260000000	1	36
2085920	F	F	Unknown	Unknown	928374	11/15/2000 12:00:00 AM	4714080000000	5	400
2085920	F	F	Unknown	Unknown	1589714	2/23/2001 12:00:00 AM	4714980000000	2	26
2085920	F	F	Unknown	Unknown	1591342	2/23/2001 12:00:00 AM	4719860000000	6	474
2085920	F	F	Unknown	Unknown	1591490	2/23/2001 12:00:00 AM	50000108190	1	56
1976717	C	C	Unknown	Unknown	1267778	1/9/2001 12:00:00 AM	4710060000000	1	189
1976717	C	C	Unknown	Unknown	1268726	1/9/2001 12:00:00 AM	4710110000000	1	45
1976717	C	C	Unknown	Unknown	1269319	1/9/2001 12:00:00 AM	4710630000000	2	54
1976717	C	C	Unknown	Unknown	1269440	1/9/2001 12:00:00 AM	4711120000000	5	140
1976717	C	C	Unknown	Unknown	1269816	1/9/2001 12:00:00 AM	4719590000000	1	125

유통 업체라면 흔하게 볼 구매 데이터.

목표 : 이탈을 예방하기 위해, 이탈자를 미리 예측해보자.

First : [ ] 을 [ ]해야 한다.

Second : 관심사는 [ ]

D:plus

3. EDA for Feature Engineering

전부 적어봅시다. // 그리고 예시를 보러 갑시다!

UserId	Age	Address	Gender	UserType	TransactionId	Timestamp	ItemId	Quantity	Value
2105345	D	F	Unknown	Unknown	1215553	12/31/2000 12:00:00 AM	4710040000000	1	149
2105345	D	F	Unknown	Unknown	1216545	12/31/2000 12:00:00 AM	4711090000000	1	179
2105345	D	F	Unknown	Unknown	1216590	12/31/2000 12:00:00 AM	9556000000000	1	28
2105345	D	F	Unknown	Unknown	1217249	12/31/2000 12:00:00 AM	4711800000000	1	199
2105345	D	F	Unknown	Unknown	1217259	12/31/2000 12:00:00 AM	4710030000000	1	139
2105345	D	F	Unknown	Unknown	1217263	12/31/2000 12:00:00 AM	4710190000000	2	10
2105345	D	F	Unknown	Unknown	1217322	12/31/2000 12:00:00 AM	4710630000000	1	33
2105345	D	F	Unknown	Unknown	1218254	12/31/2000 12:00:00 AM	4711260000000	1	36
2085920	F	F	Unknown	Unknown	928374	11/15/2000 12:00:00 AM	4714080000000	5	400
2085920	F	F	Unknown	Unknown	1589714	2/23/2001 12:00:00 AM	4714980000000	2	26
2085920	F	F	Unknown	Unknown	1591342	2/23/2001 12:00:00 AM	4719860000000	6	474
2085920	F	F	Unknown	Unknown	1591490	2/23/2001 12:00:00 AM	50000108190	1	56
1976717	C	C	Unknown	Unknown	1267778	1/9/2001 12:00:00 AM	4710060000000	1	189
1976717	C	C	Unknown	Unknown	1268726	1/9/2001 12:00:00 AM	4710110000000	1	45
1976717	C	C	Unknown	Unknown	1269319	1/9/2001 12:00:00 AM	4710630000000	2	54
1976717	C	C	Unknown	Unknown	1269440	1/9/2001 12:00:00 AM	4711120000000	5	140
1976717	C	C	Unknown	Unknown	1269816	1/9/2001 12:00:00 AM	4719590000000	1	125

유통 업체라면 흔하게 볼 구매 데이터.
목표 : 이탈을 예방하기 위해, 이탈자를 미리 예측해보자.
First : [ ] 을 [ ]해야 한다.
Second : 관심사는 [ ]

Third : [ ] 를 설명하기 위해서는 [ ]가 필요할 것 같다.

D:plus

못다한 이야기들

사실은... 못다한 이야기가 너무 많습니다!

- 통계 기본, 확률 기본, 분포이론
- 큰 수의 법칙, 중심극한정리, 대표본 근사이론
- 우도, 베이지안
- 확률과정, Probabilistic Graphical Model

D:plus

못다한 이야기들

사실은... 빼버리고 싶은 내용도 많았습니다!

- 직관적인 이해를 방해하기 시작하는 생소한 용어
- 직관적인 이해를 방해하기 시작하는 생소한 수식

그럼에도 불구하고, 연결고리가 되는 지점들은 생략하지 않았습니다.

혹은

혼자 공부할 때 넘어갔다가 고생할만한 부분은 생략하지 않았습니다.

D:plus

못다한 이야기들

수학적인, 알고리즘의 디테일은 다른 강의들에도 많이 있습니다.

이 강의에서는

쉽게 접근할 수 있는 자료들에서

생략된 앞 뒤 내용을 전달 드리고자 했습니다.

D:plus