Chapter. 02

분류 분석

분류 분석과 로지스틱 회귀 모델

FAST CAMPUS
ONLINE

직장인을 위한 파이썬 데이터분석

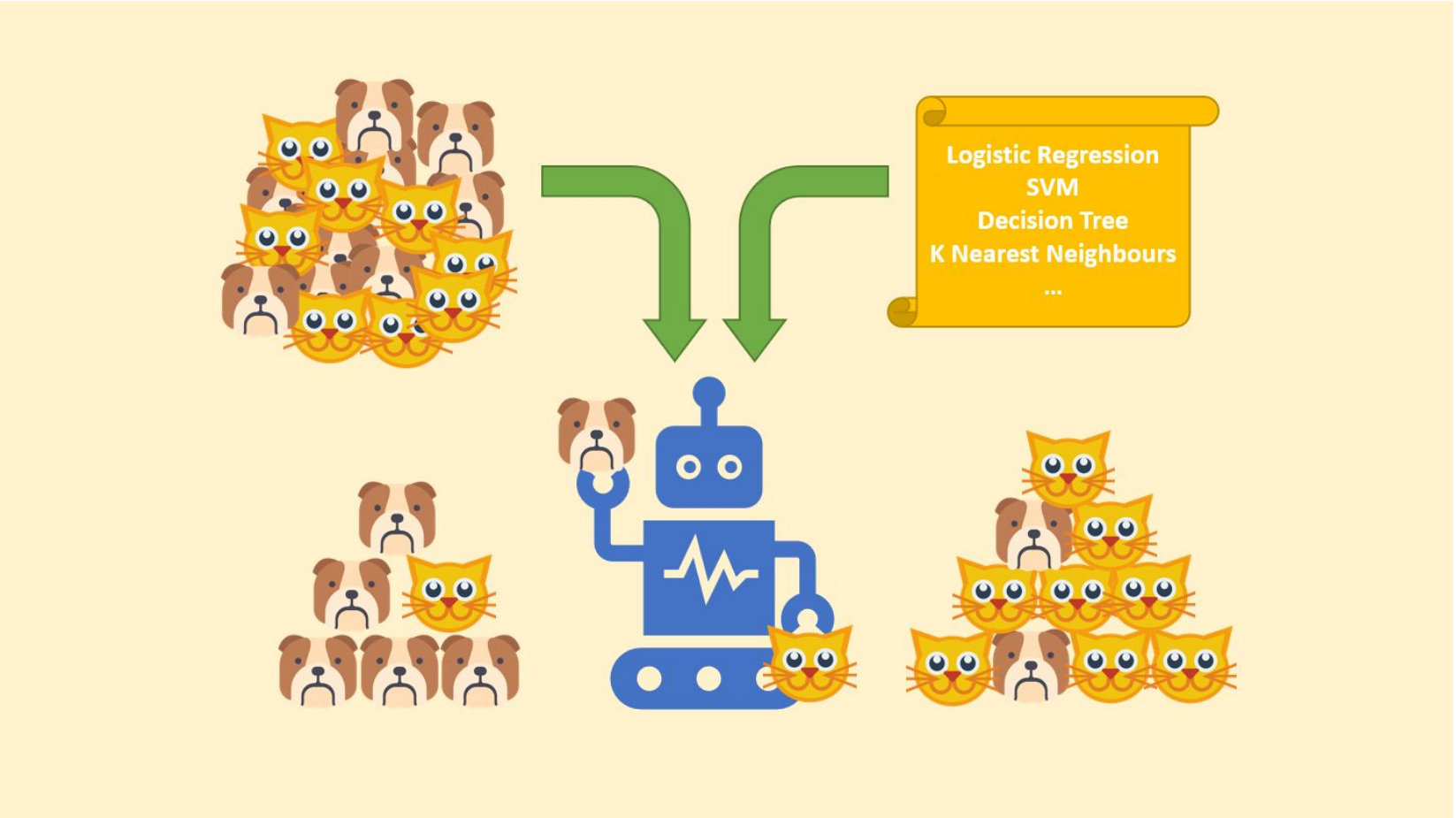
강사. 윤기태

Chapter. 02

분류 분석

FAST CAMPUS
ONLINE
윤기태 강사.

I 분류 분석이란



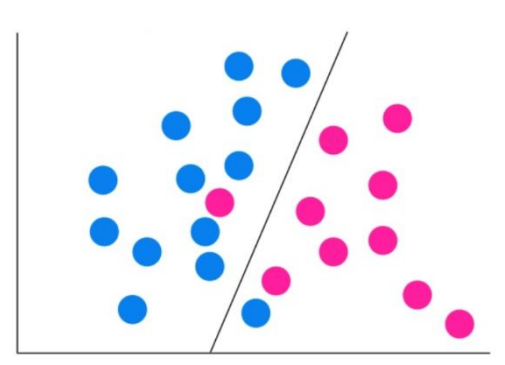
출처 - https://towardsdatascience.com/analytics-building-blocks-binary-classification-d205890314fc

FAST CAMPUS
ONLINE
윤기태 강사.

I 분류 분석의 종류

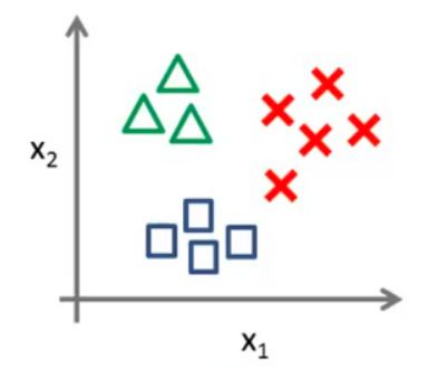
STEP 1

이진 분류
(Binary Classification)



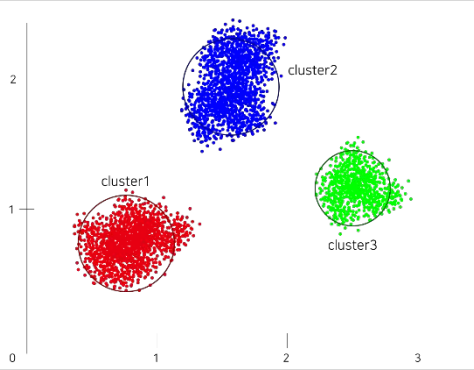
STEP 2

다중 분류
(Multi-Class Classification)



STEP 3

군집 분류
(Clustering)



FAST CAMPUS
ONLINE
윤기태 강사.

I 분류 분석의 예시

“◆◆사다리
게임◆◆으로 앉아서
수익…”

“[일정 공유] 오늘 오후
7시 30분…”

I 분류 분석의 예시

“◆◆사다리
게임◆◆으로 앉아서
수익…”

텍스트 분석

“[일정 공유] 오늘 오후
7시 30분…”

FAST CAMPUS
ONLINE
윤기태 강사.

I 분류 분석의 예시

스팸 메일

"◆◆사다리
게임◆◆으로 앉아서
수익…"

일반 메일

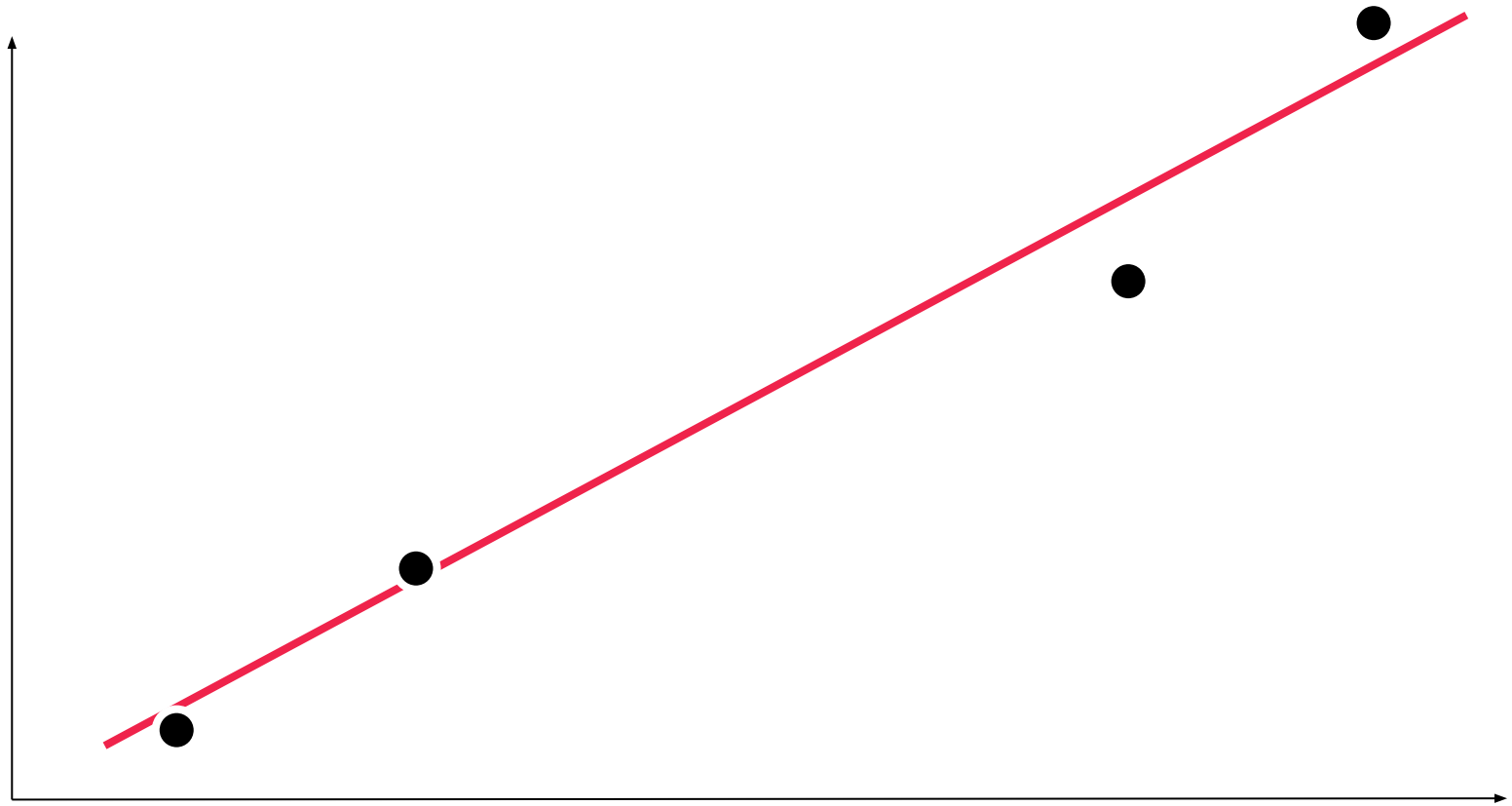
"[일정 공유] 오늘 오후
7시 30분…"

Chapter. 02

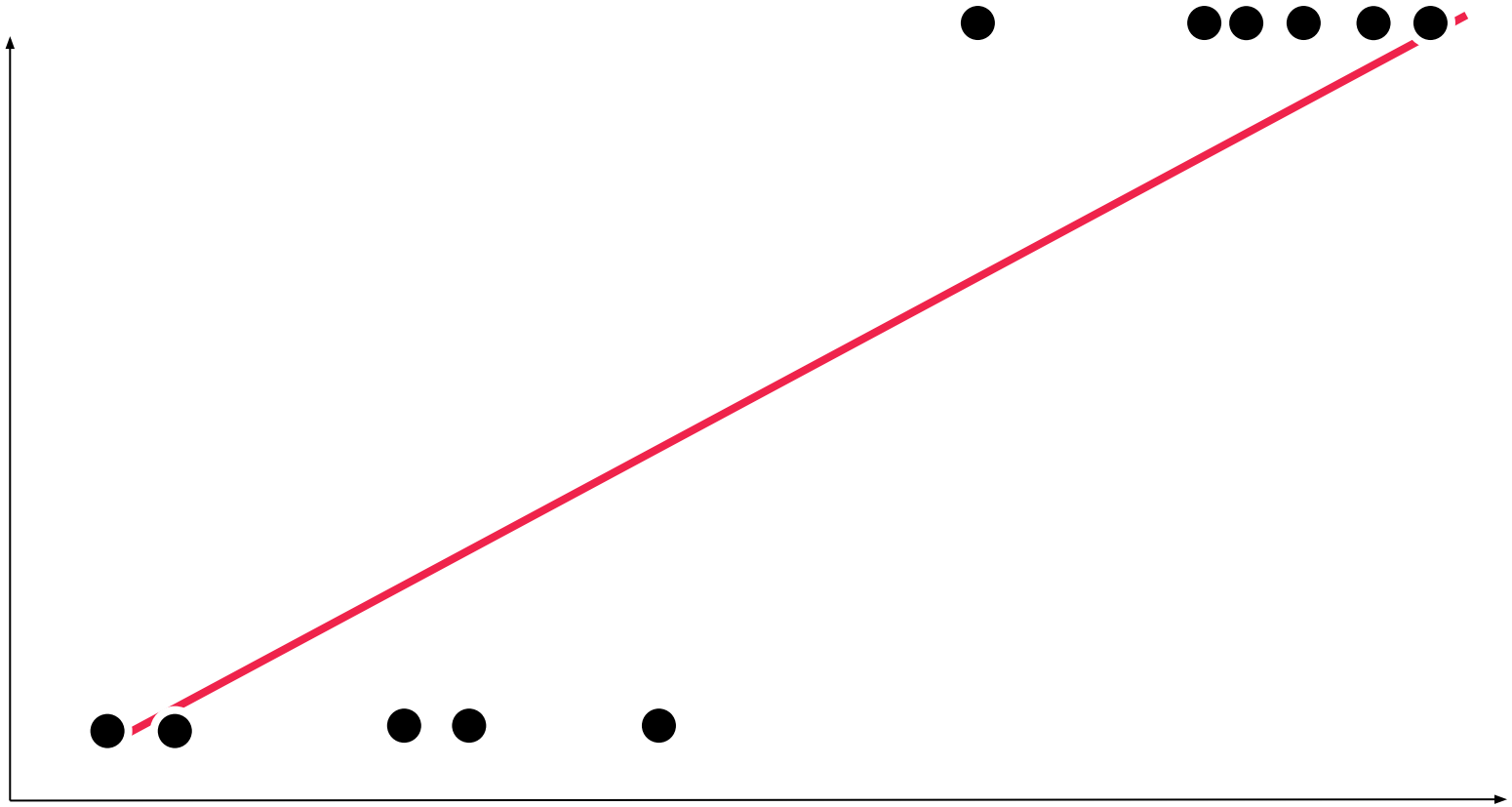
로지스틱 회귀 모델

FAST CAMPUS
ONLINE
윤기태 강사.

I 회귀 분석 모델

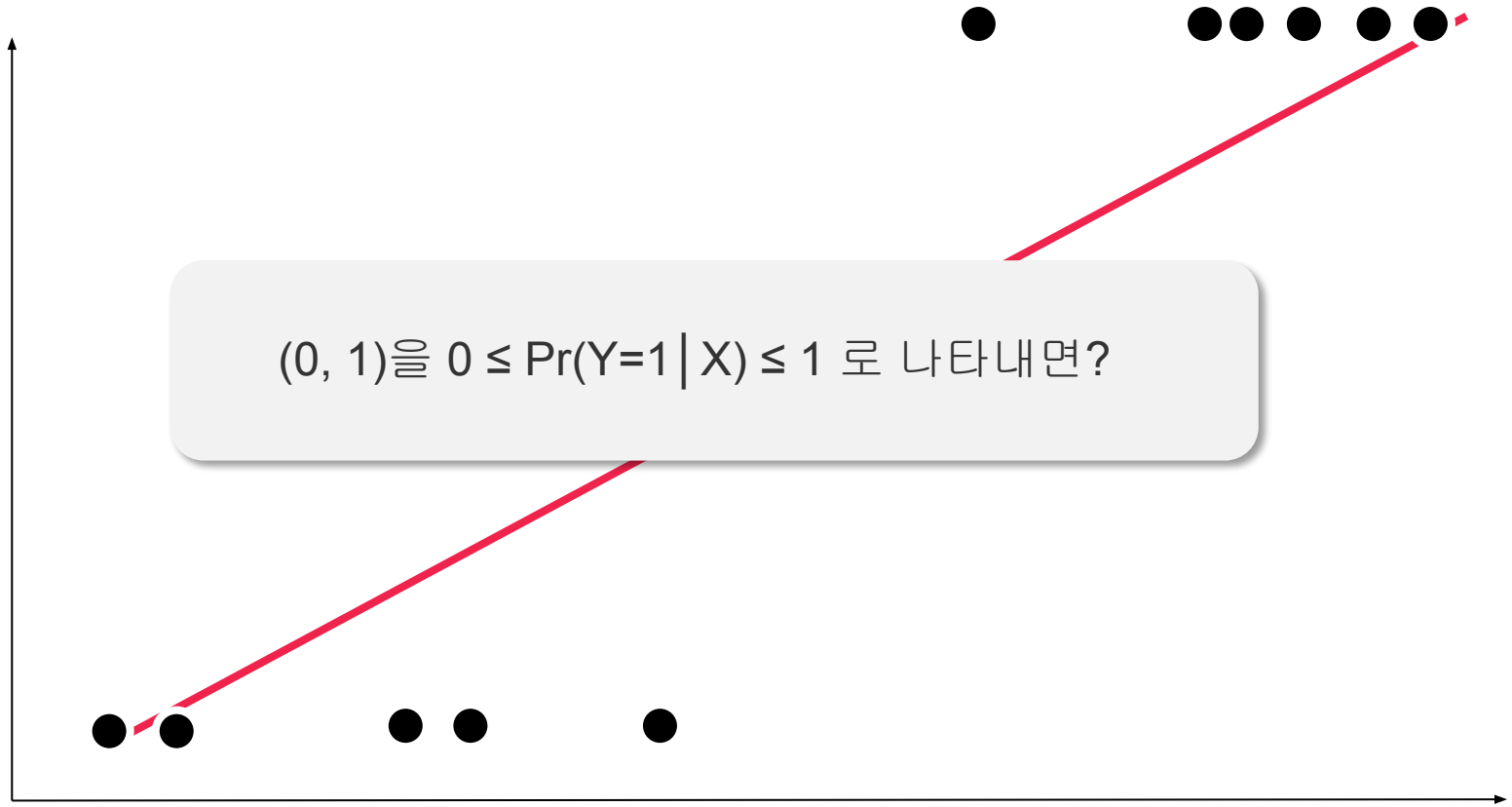


I 회귀 분석 모델



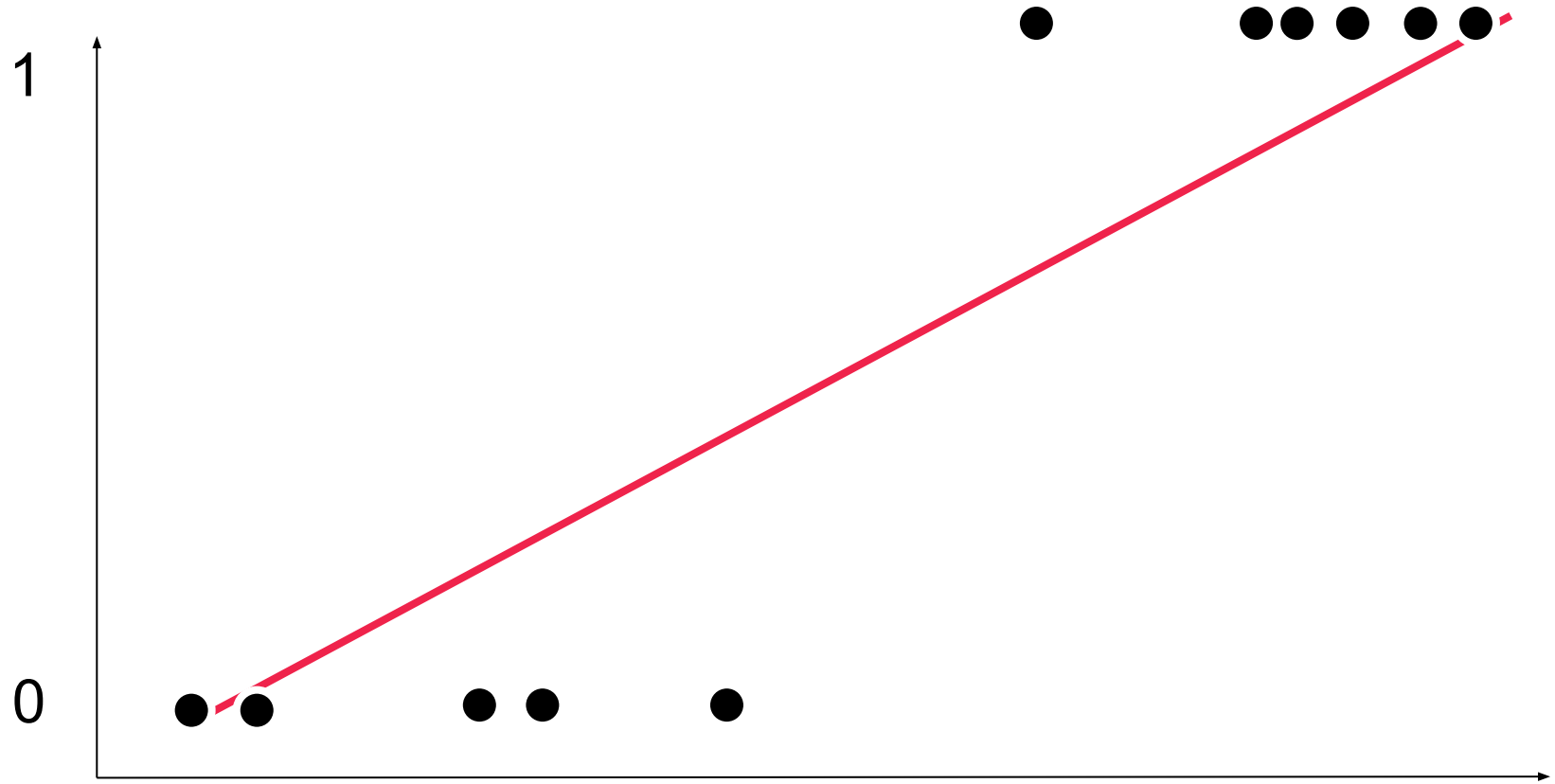
FAST CAMPUS
ONLINE
윤기태 강사.

I 회귀 분석 모델



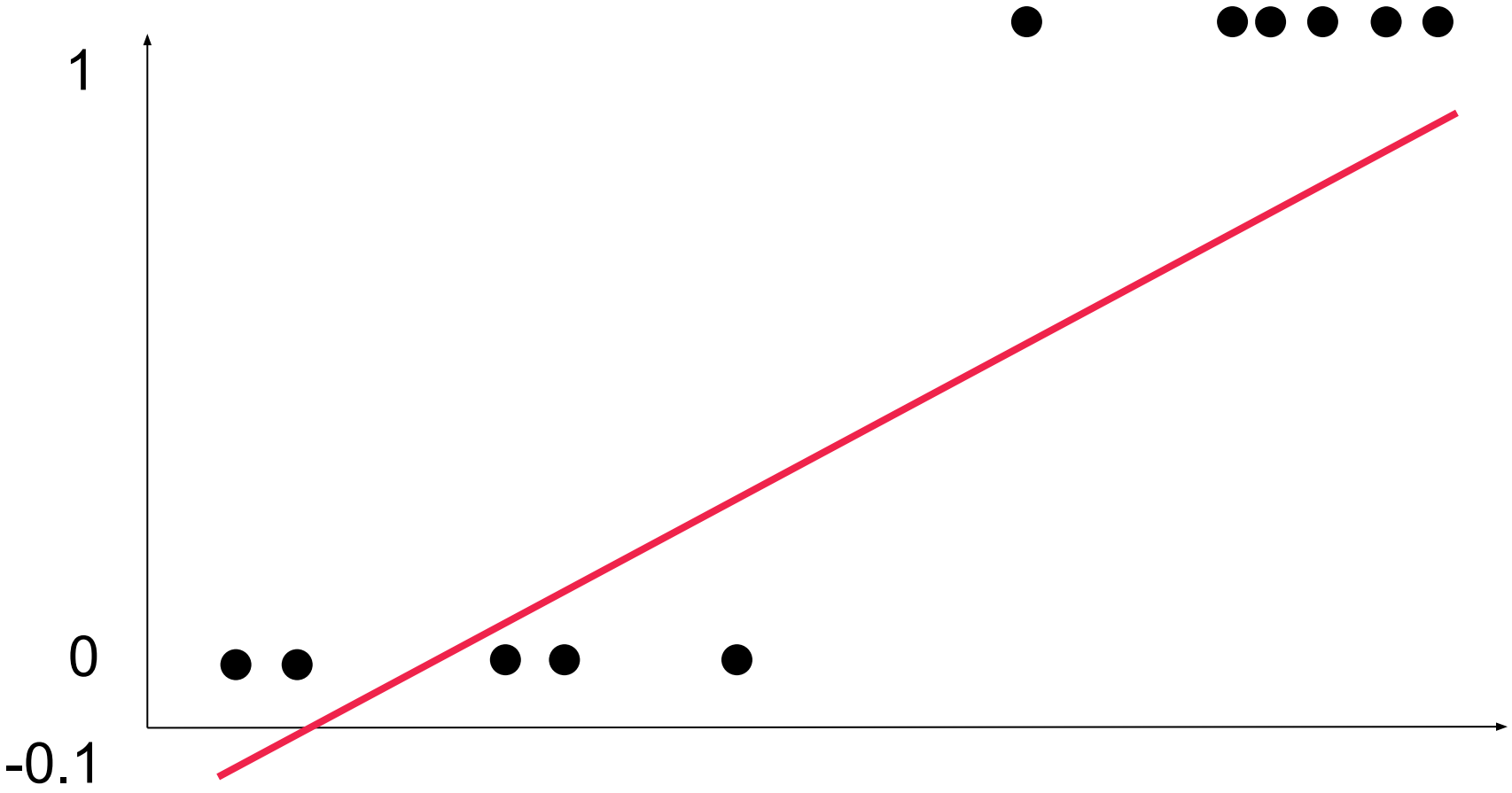
FAST CAMPUS
ONLINE
윤기태 강사.

회귀 분석의 확률 추정



FAST CAMPUS
ONLINE
윤기태 강사.

I 회귀 분석의 확률 추정



FAST CAMPUS
ONLINE
윤기태 강사.

I 로지스틱 함수

Odds(승산) :

임의의 사건이 발생하지 않을 확률 대비 일어날 확률의 비율

$$Odds = (p(X))/(1-p(X))$$

I 로지스틱 함수

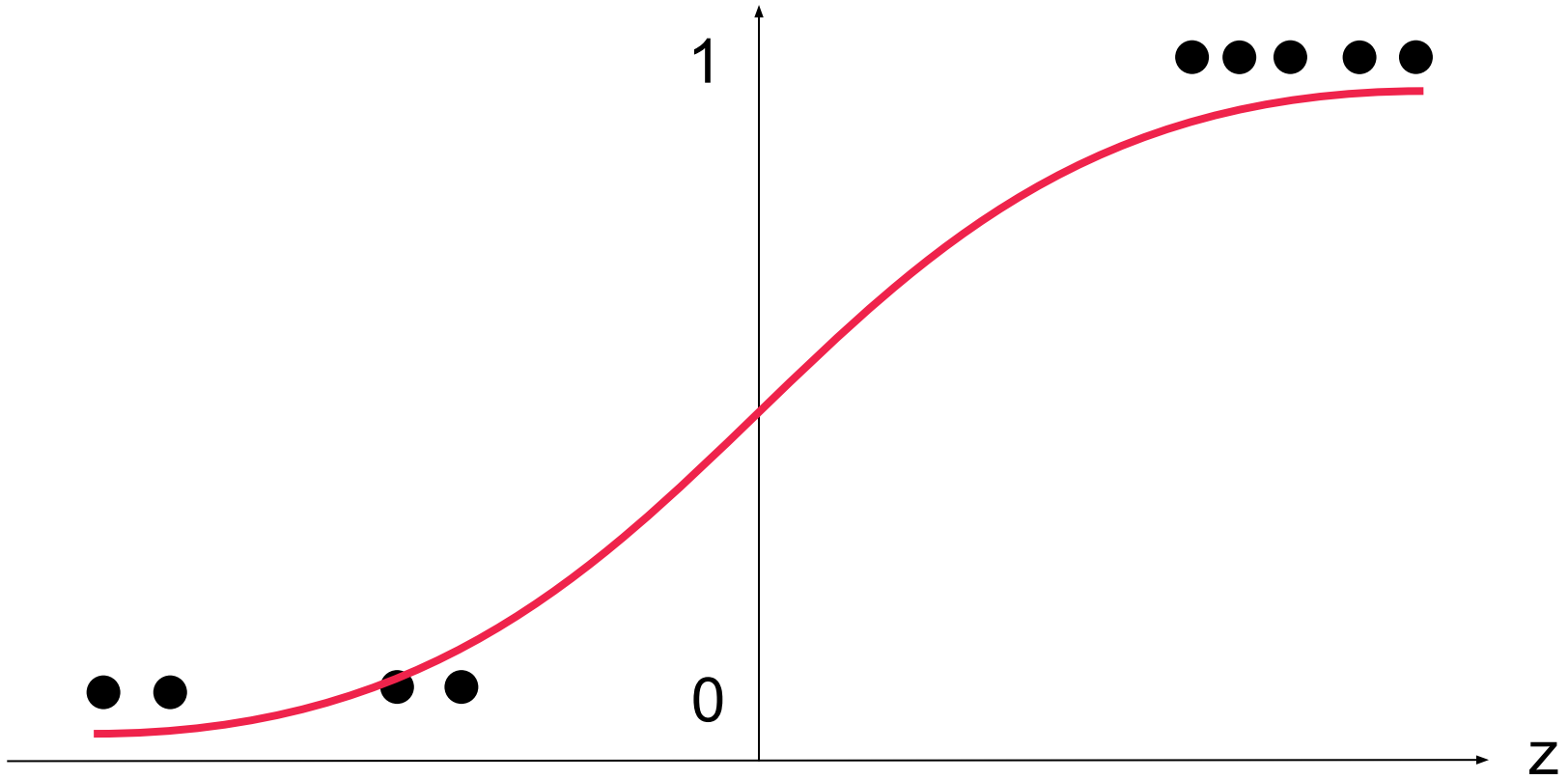
로짓 변환과 로지스틱 함수

$$\text{logit}(p) = \log \frac{p}{1-p}$$

f(X) : Logistic Function

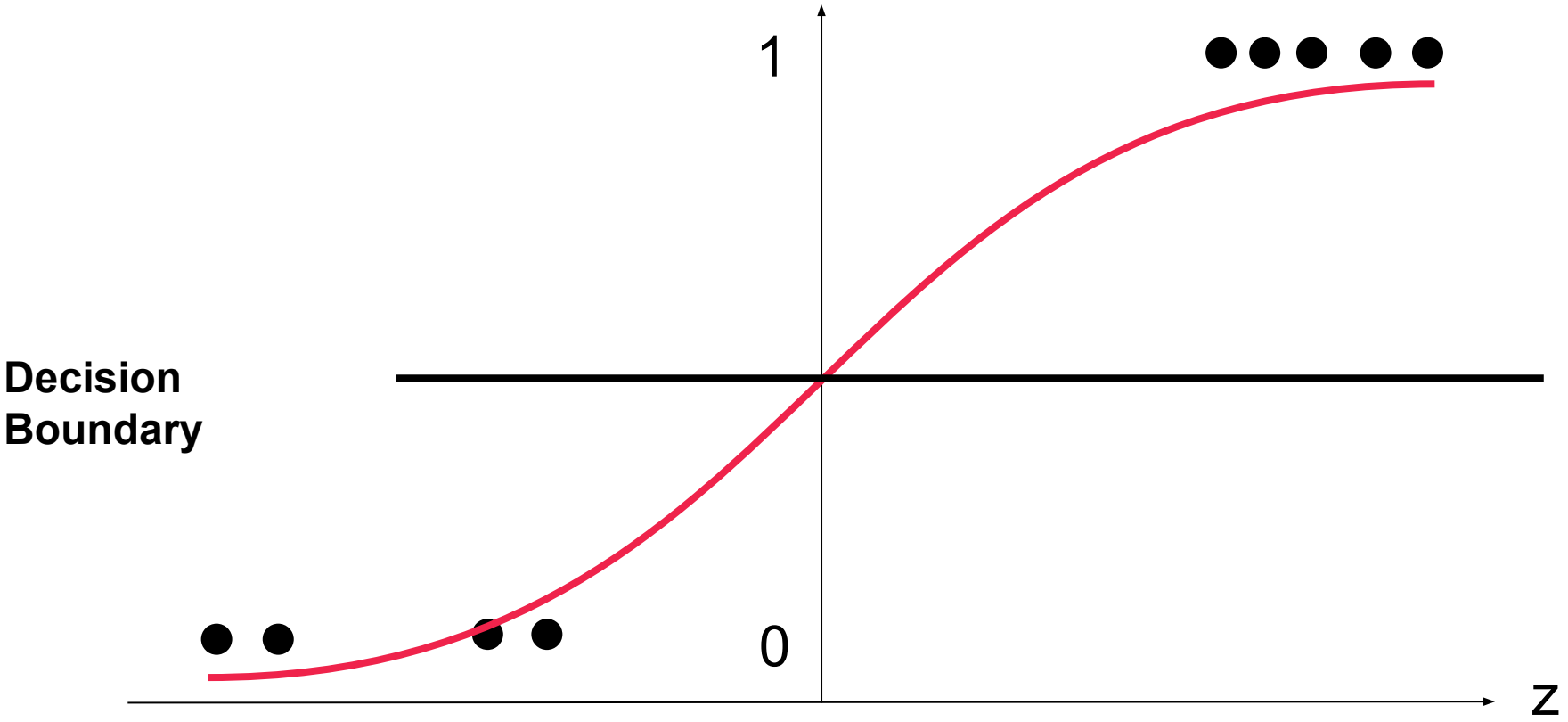
$$f(x) = \frac{1}{1+e^{-x}}$$

I 로지스틱 회귀 모델



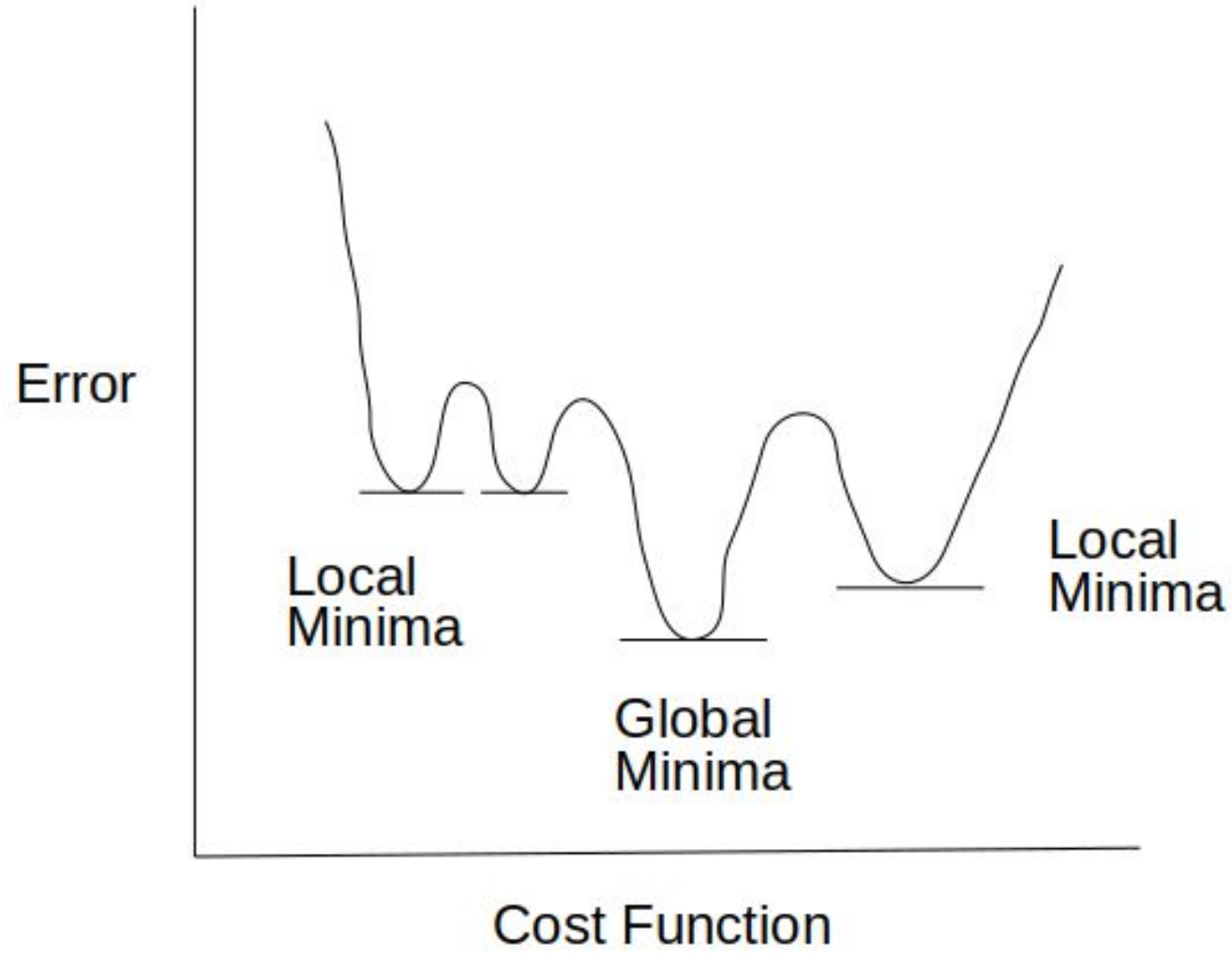
FAST CAMPUS
ONLINE
윤기태 강사.

I 로지스틱 회귀 모델



FAST CAMPUS
ONLINE
윤기태 강사.

로지스틱 회귀 모델의 비용 함수



FAST CAMPUS
ONLINE
윤기태 강사.

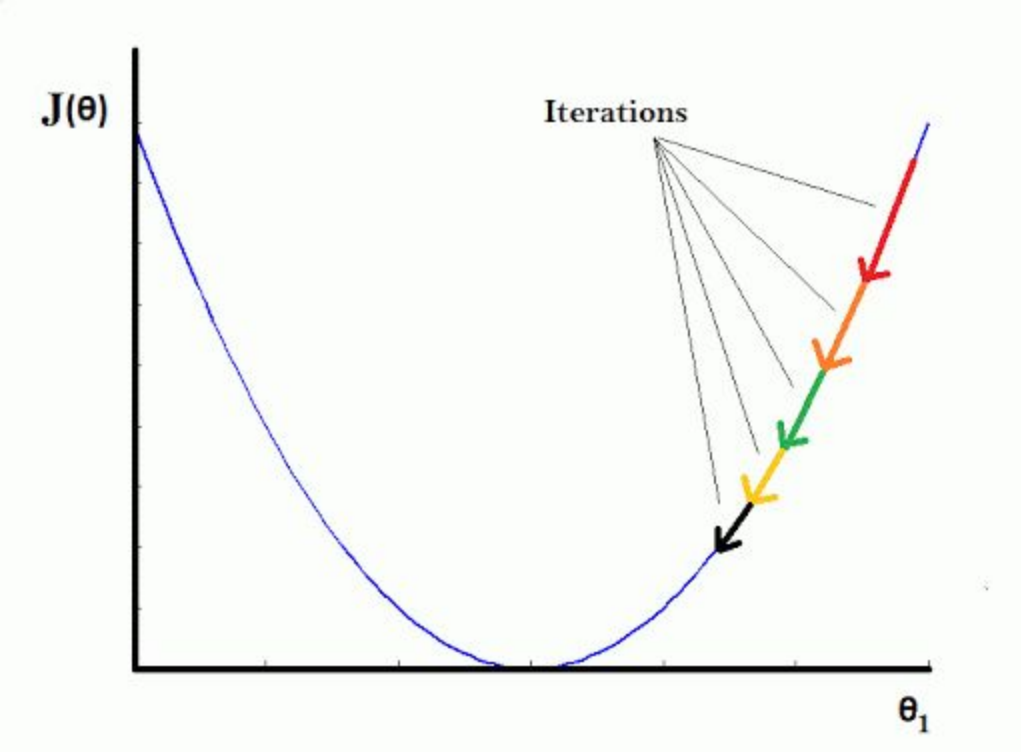
로지스틱 회귀 모델의 비용 함수

$$J(\boldsymbol{\Phi}) = \frac{1}{m}\sum_{i=1}^{m} Cost\left(h_{\boldsymbol{\Phi}}\left(x^{(i)}\right), y^{(i)}\right)$$

$$Cost(h_{\boldsymbol{\Phi}}(x), y) = -\log\left(h_{\boldsymbol{\Phi}}(x)\right) \;\; if \; y = 1$$

$$Cost(h_{\boldsymbol{\Phi}}(x), y\,) = -\log\left(1 - h_{\boldsymbol{\Phi}}(x)\right) \;\; if \; y = 0$$

I 로지스틱 회귀 모델



FAST CAMPUS
ONLINE
윤기태 강사.

Chapter. 02

회귀 분석 기반 모델의 전처리 방법

FAST CAMPUS
ONLINE
윤기태 강사.

I 원-핫 인코딩(one-hot encoding) 처리 방법

이름	키	몸무게	좋아하는 과일
기택	183	80	바나나
기우	177	66	키위
다송	123	35	사과
기정	162	54	석류

I 원-핫 인코딩(one-hot encoding) 처리 방법

y x1 ? ?

키	몸무게	좋아하는 과일
183	80	바나나
177	66	키위
123	35	사과
162	54	석류

FAST CAMPUS
ONLINE
윤기태 강사.

I 원-핫 인코딩(one-hot encoding) 처리 방법

	바나나	키위	사과	석류
바나나 →	1	0	0	0
키위 →	0	1	0	0
사과 →	0	0	1	0
석류 →	0	0	0	1

I 원-핫 인코딩(one-hot encoding) 처리 방법

y	x1	x2	x3	x4	x5
키	몸무게	바나나	키위	사과	석류
183	80	1	0	0	0
177	66	0	1	0	0
123	35	0	0	1	0
162	54	0	0	0	1

I 멀티 레이블 인코딩(multi-label encoding) 처리 방법

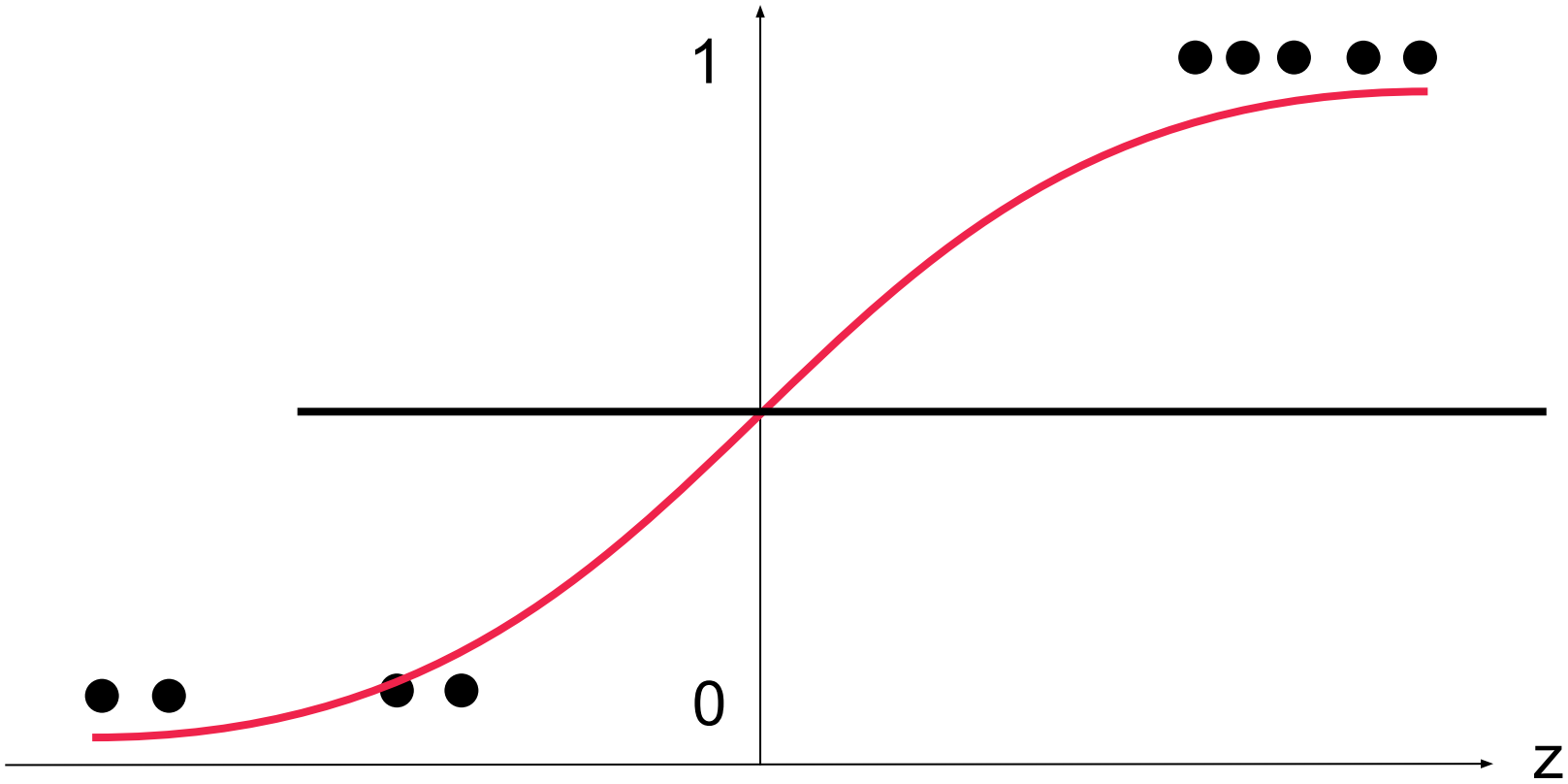
y	x1	x2	x3	x4	x5
키	몸무게	바나나	키위	사과	석류
183	80	1	0	**1**	0
177	66	0	1	0	0
123	35	0	**1**	1	0
162	54	0	0	0	1

Chapter. 02

분류 모델의 평가 방법

FAST CAMPUS
ONLINE
윤기태 강사.

참고 : 로지스틱 회귀 모델의 학습 결과 평가 방법



FAST CAMPUS
ONLINE
윤기태 강사.

I Counfusion Matrix를 활용한 분류 분석 평가 방법

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

FAST CAMPUS
ONLINE
윤기태 강사.

I Counfusion Matrix를 활용한 분류 분석 평가 방법

		Predicted class: P	Predicted class: N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

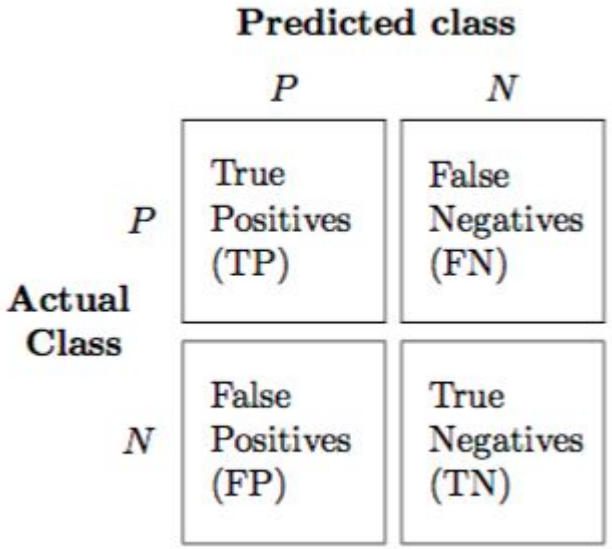
- 정확도(Accuracy): $\frac{TP+TN}{TP+TN+FP+FN}$
- 정밀도(Precision): $\frac{TP}{TP+FP}$
- 재현도(Recall): $\frac{TP}{TP+FN}$
- 특이도(Specificity): $\frac{TN}{TN+FP}$

I Counfusion Matrix를 활용한 분류 분석 평가 방법

		Predicted class	
		P	N
Actual Class	P	8	2
	N	2	188

- 정확도(Accuracy): $\frac{TP+TN}{TP+TN+FP+FN}$
- 정밀도(Precision): $\frac{TP}{TP+FP}$
- 재현도(Recall): $\frac{TP}{TP+FN}$
- 특이도(Specificity): $\frac{TN}{TN+FP}$

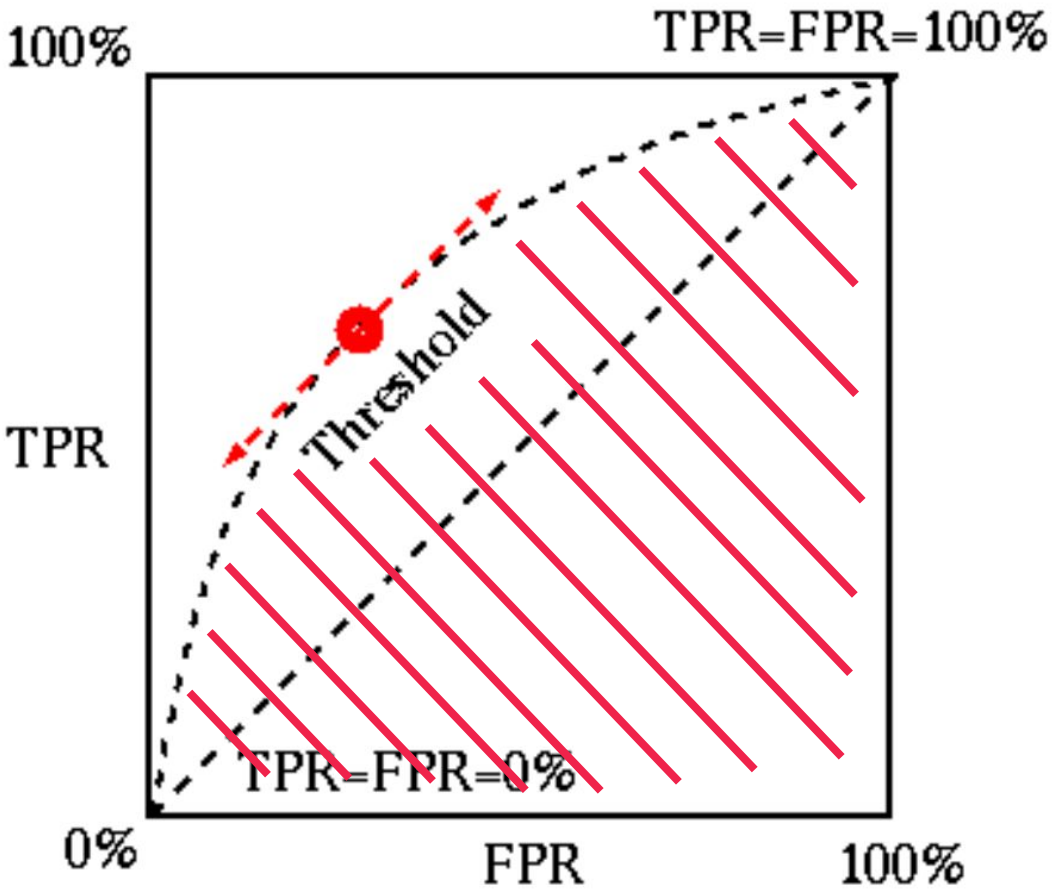
I Counfusion Matrix를 활용한 분류 분석 평가 방법



$$F1 = \frac{2 * precision * recall}{precision + recall}$$

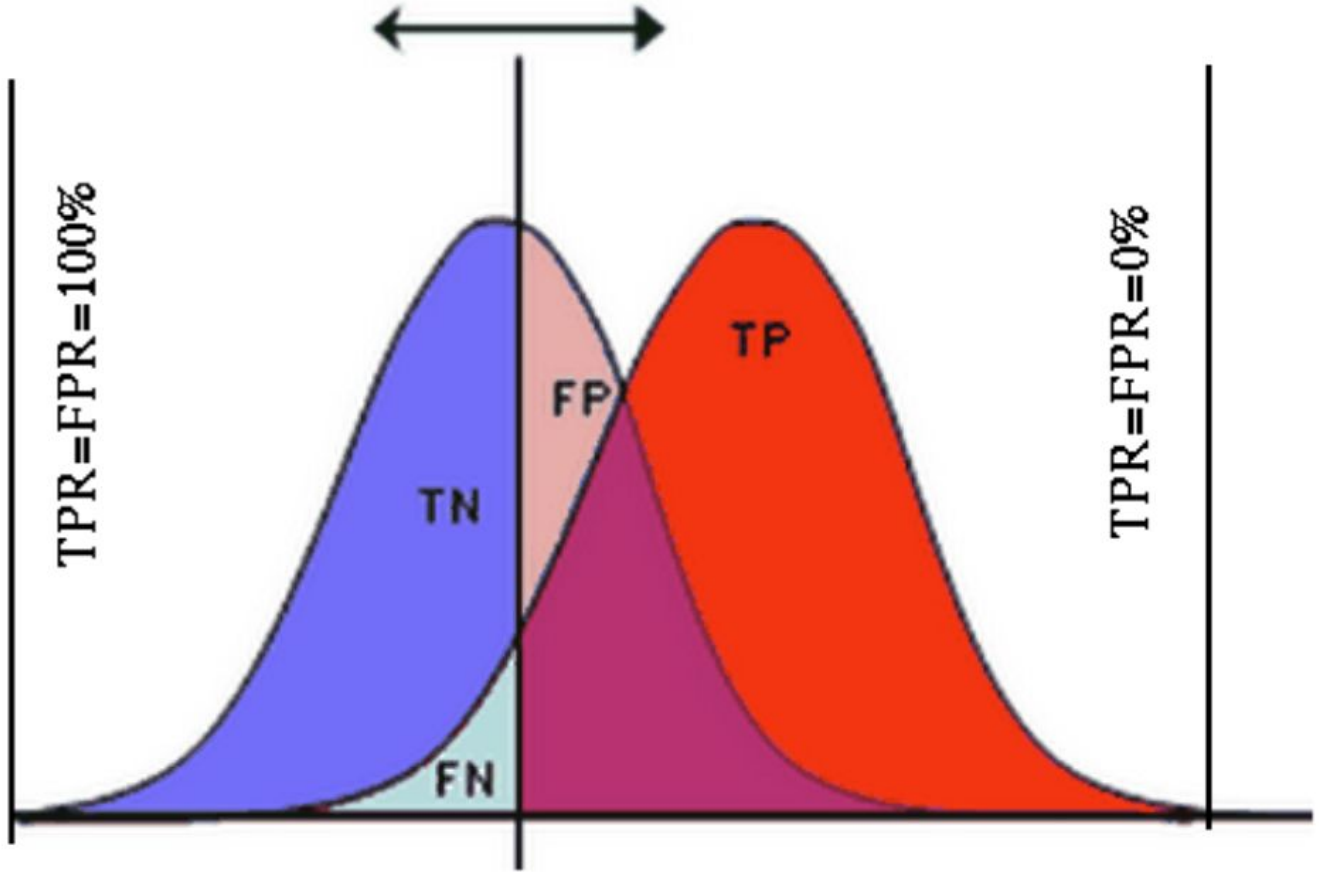
AUC (Area Under Curve)

I Counfusion Matrix를 활용한 분류 분석 평가 방법



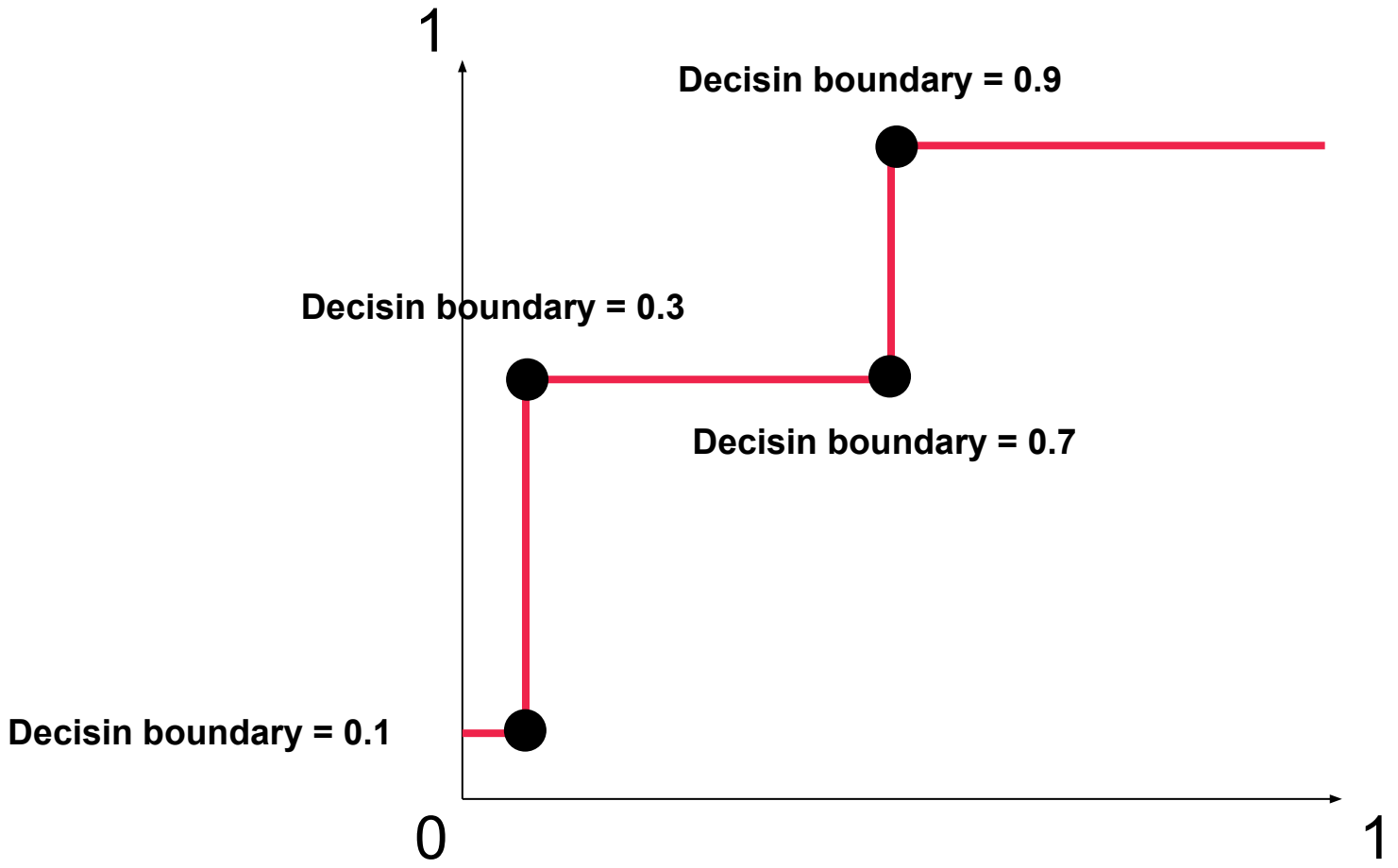
FAST CAMPUS
ONLINE
윤기태 강사.

I Counfusion Matrix를 활용한 분류 분석 평가 방법



FAST CAMPUS
ONLINE
윤기태 강사.

I Counfusion Matrix를 활용한 분류 분석 평가 방법



FAST CAMPUS
ONLINE
윤기태 강사.