Machine learning

Logistic Regression

Logo

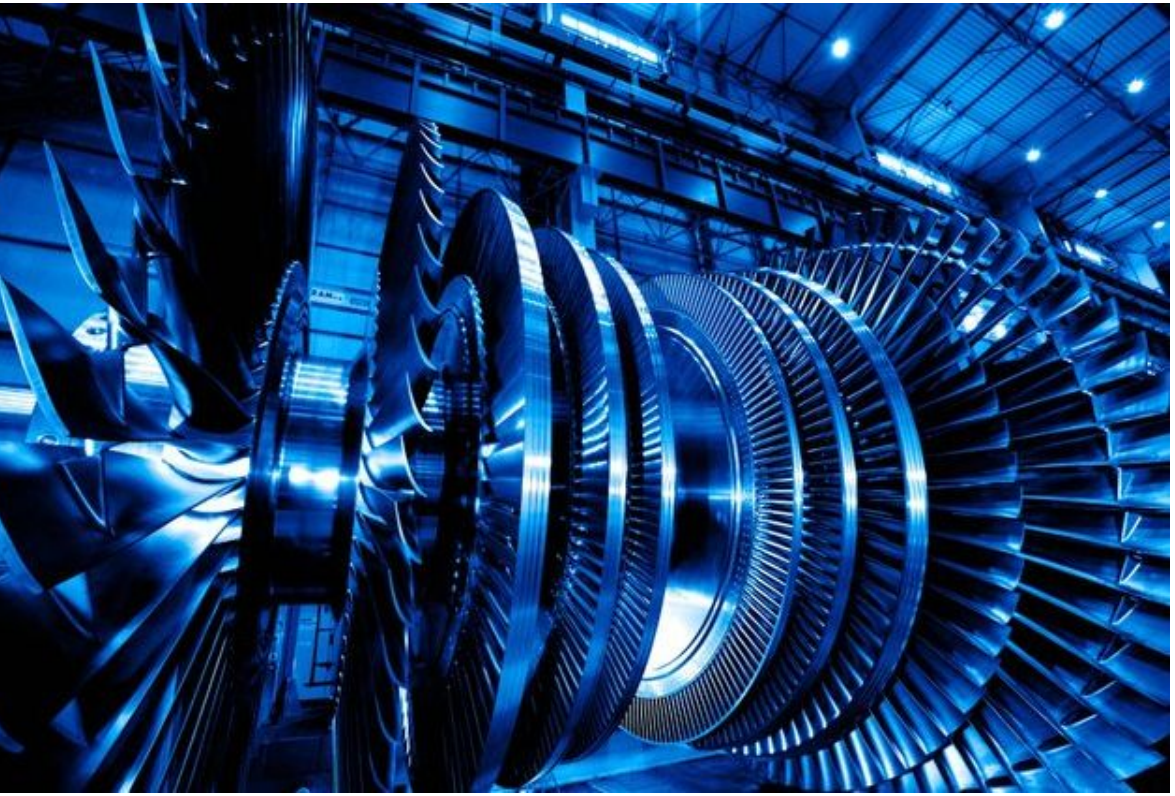
기호적인 회귀


Regression 회귀 Mission
Classification 범주 분류 Mission

회귀
Regression

feat1 - RPM: Revolutions (Revolve) Per Minute
target - Vibration: 특정 부품 진동 값

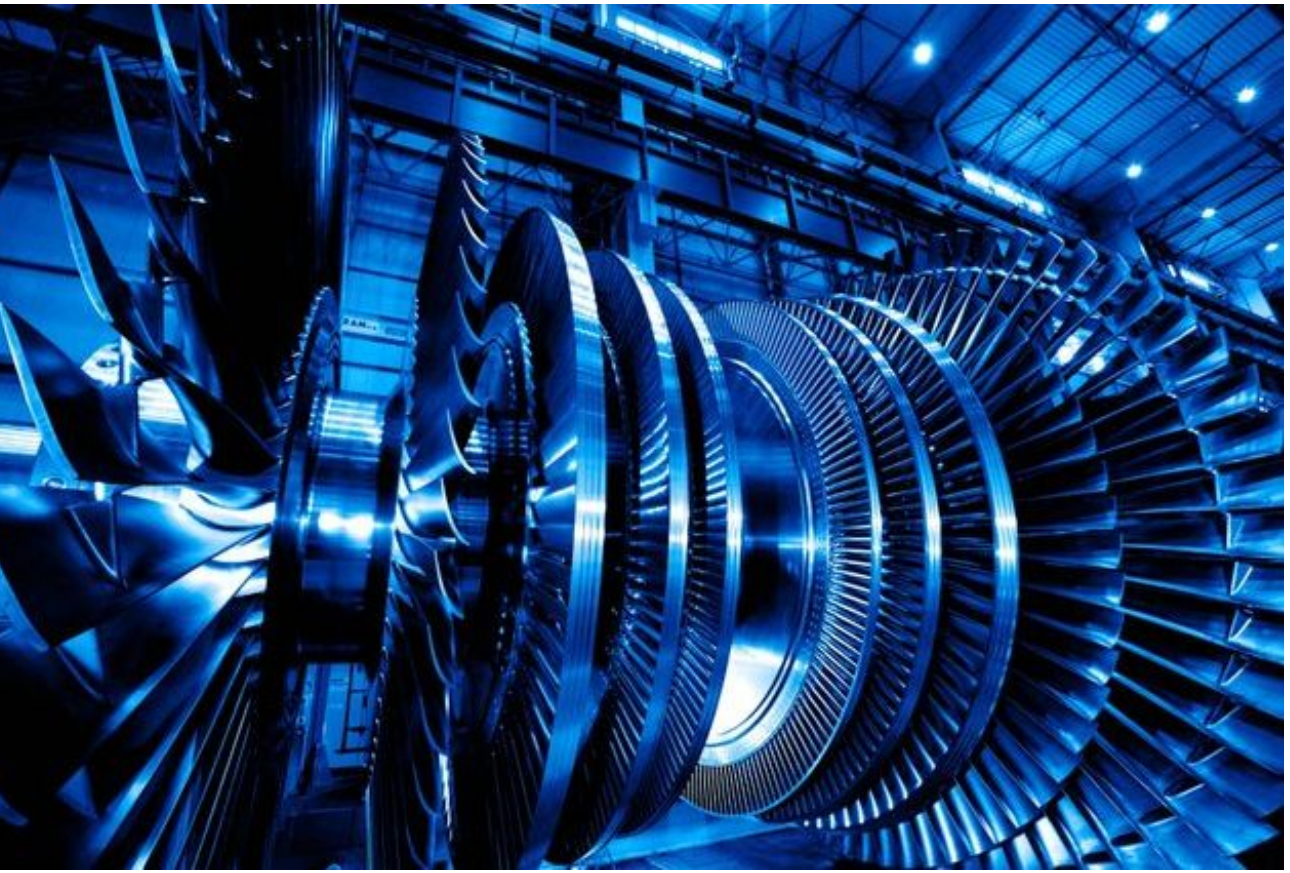
- 미션: 특정 부품 진동 값을 추정
(너무 진동수가 높을 것 예상 되면 RPM을 낮출 수 있죠)



ID	RPM	VIBRATION
1	568	585
2	586	565
3	609	536
4	616	492
5	632	465
6	652	528
7	655	496
8	660	471
9	688	408
10	696	399
11	708	387
12	701	434
13	715	506
14	732	485
15	731	395
16	749	398
17	759	512
18	773	431
19	782	456
20	797	476
21	794	421
22	824	452
23	835	441
24	862	372
25	879	340
26	892	370
27	913	373
28	933	330

범주 분류 문제
Classification Problem

- 미션: 발전기의 상태를 알아야 한다!

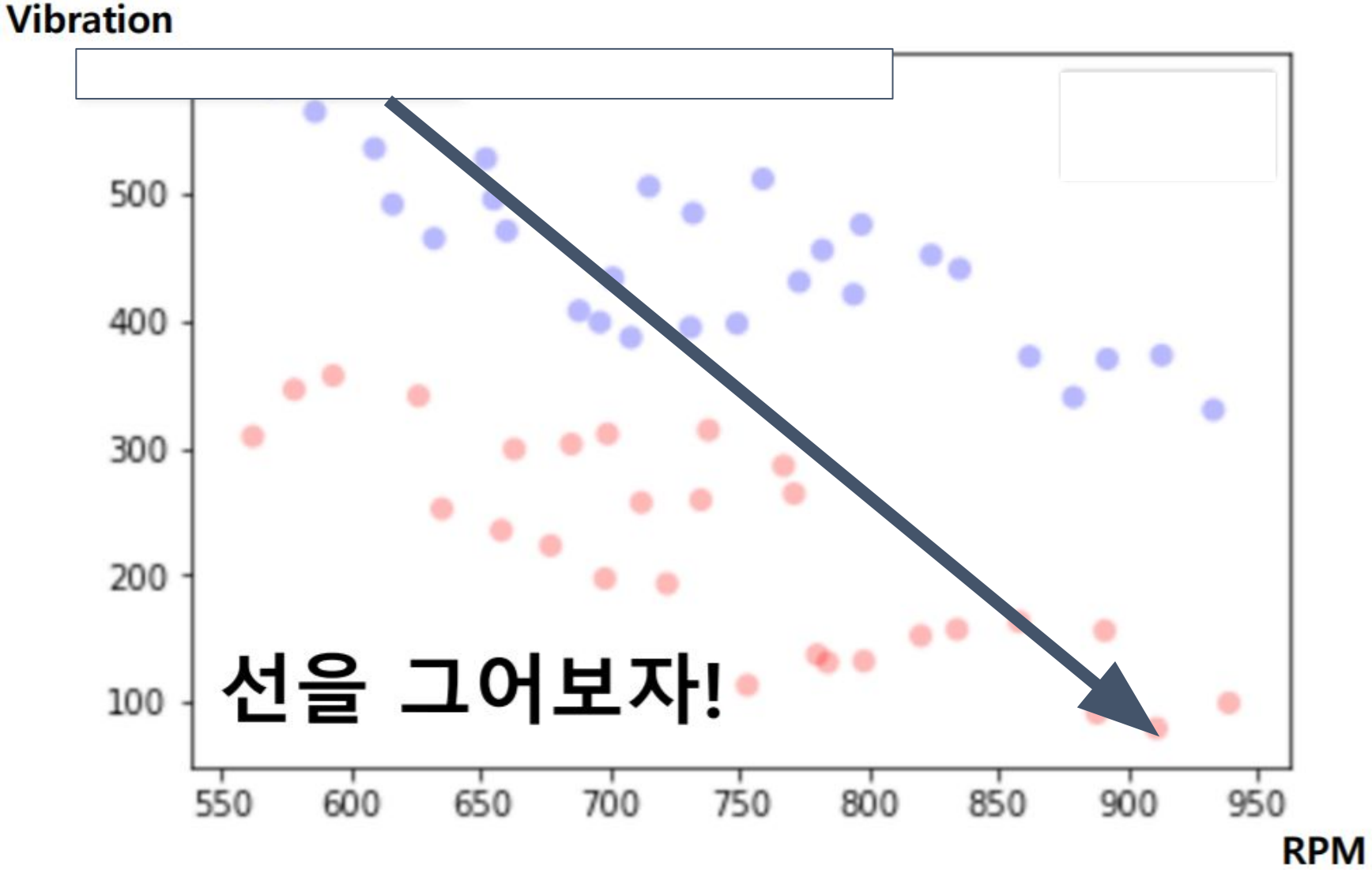


feat1 - RPM: Revolutions (Revolve) Per Minute
feat2 - Vibration: 특정 부품 동 상태
Label = Status: good
vs faulty

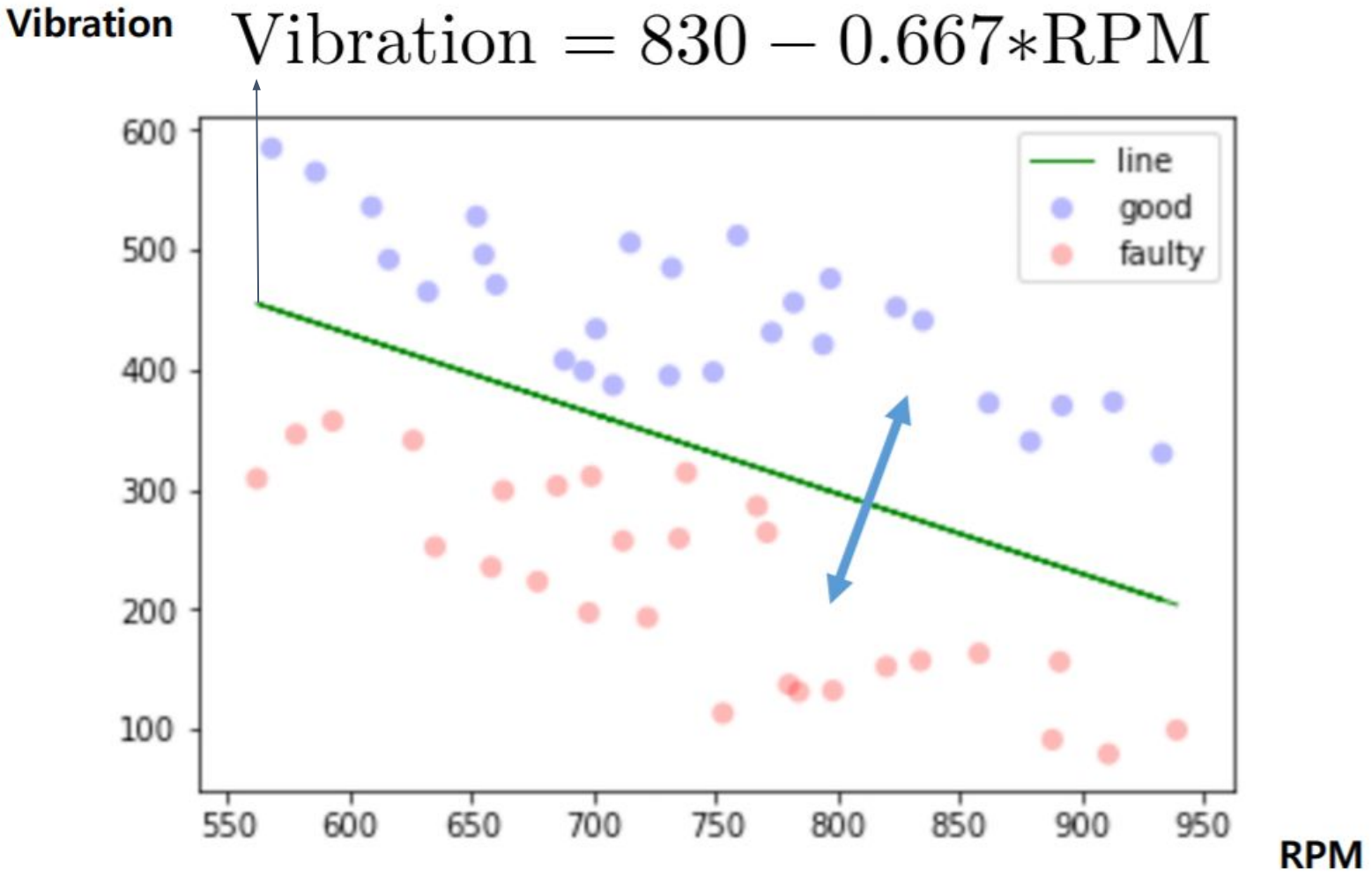
ID	RPM	VIBRATION	STATUS
1	568	585	good
2	586	565	good
3	609	536	good
4	616	492	good
5	632	465	good
6	652	528	good
7	655	496	good
8	660	471	good
9	688	408	good
10	696	399	good
11	708	387	good
12	701	434	good
13	715	506	good
14	732	485	good
15	731	395	good
16	749	398	good
17	759	512	good
18	773	431	good
19	782	456	good
20	797	476	good
21	794	421	good
22	824	452	good
23	835	441	good
24	862	372	good
25	879	340	good
26	892	370	good
27	913	373	good
28	933	330	good

ID	RPM	VIBRATION	STATUS
29	562	309	faulty
30	578	346	faulty
31	593	357	faulty
32	626	341	faulty
33	635	252	faulty
34	658	235	faulty
35	663	299	faulty
36	677	223	faulty
37	685	303	faulty
38	698	197	faulty
39	699	311	faulty
40	712	257	faulty
41	722	193	faulty
42	735	259	faulty
43	738	314	faulty
44	753	113	faulty
45	767	286	faulty
46	771	264	faulty
47	780	137	faulty
48	784	131	faulty
49	798	132	faulty
50	820	152	faulty
51	834	157	faulty
52	858	163	faulty
53	888	91	faulty
54	891	156	faulty
55	911	79	faulty
56	939	99	faulty

[1] 회귀 미션
Regression Mission



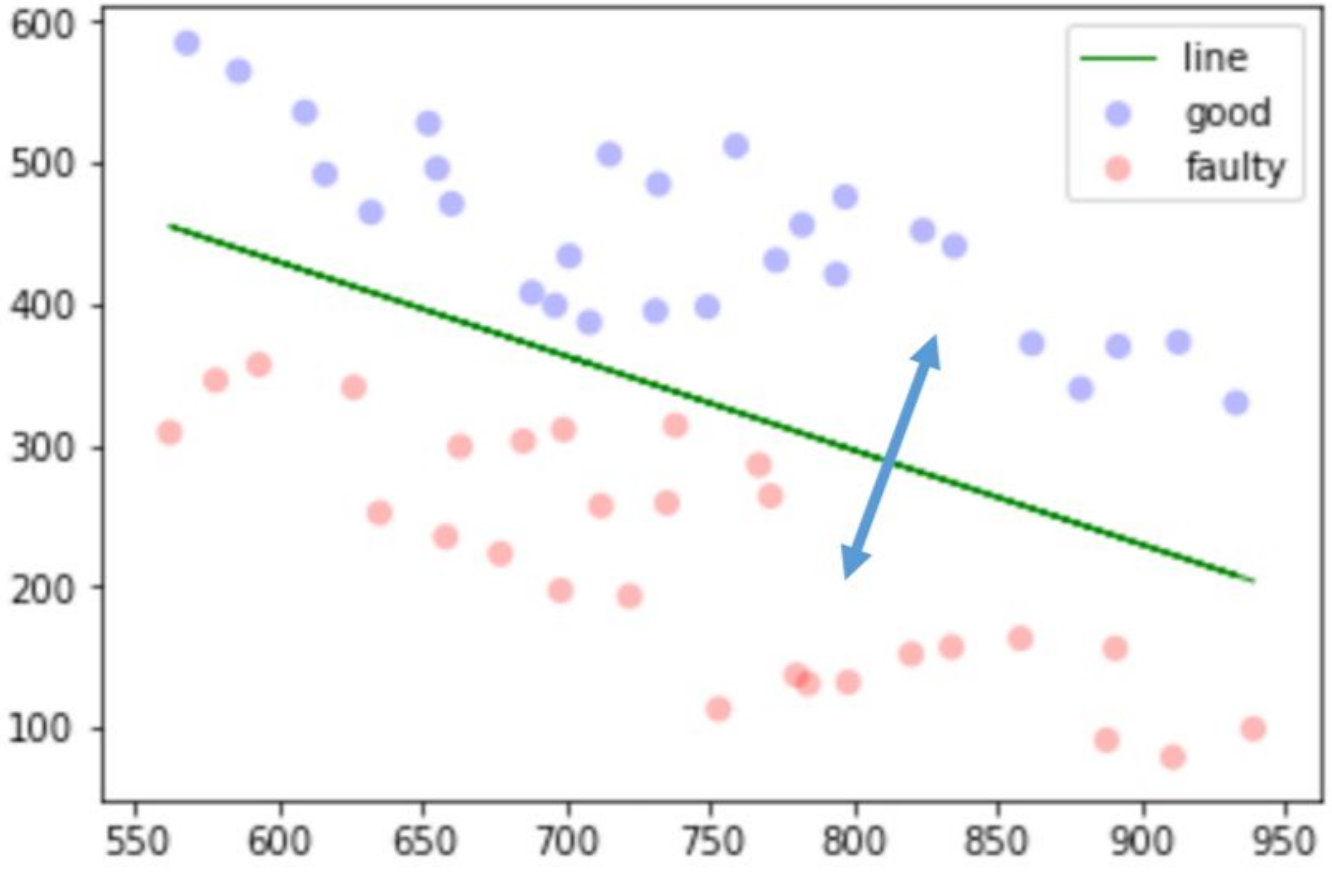
[2] 범주 분류
Classification Mission



[2] 범주 분류
Classification Mission

Vibration

$$\text{Vibration} = 830 - 0.667 * \text{RPM}$$



새 측정 Vibration, RPM 조합이
'회귀선' 기준으로
아래 범위 좌표에 있으면

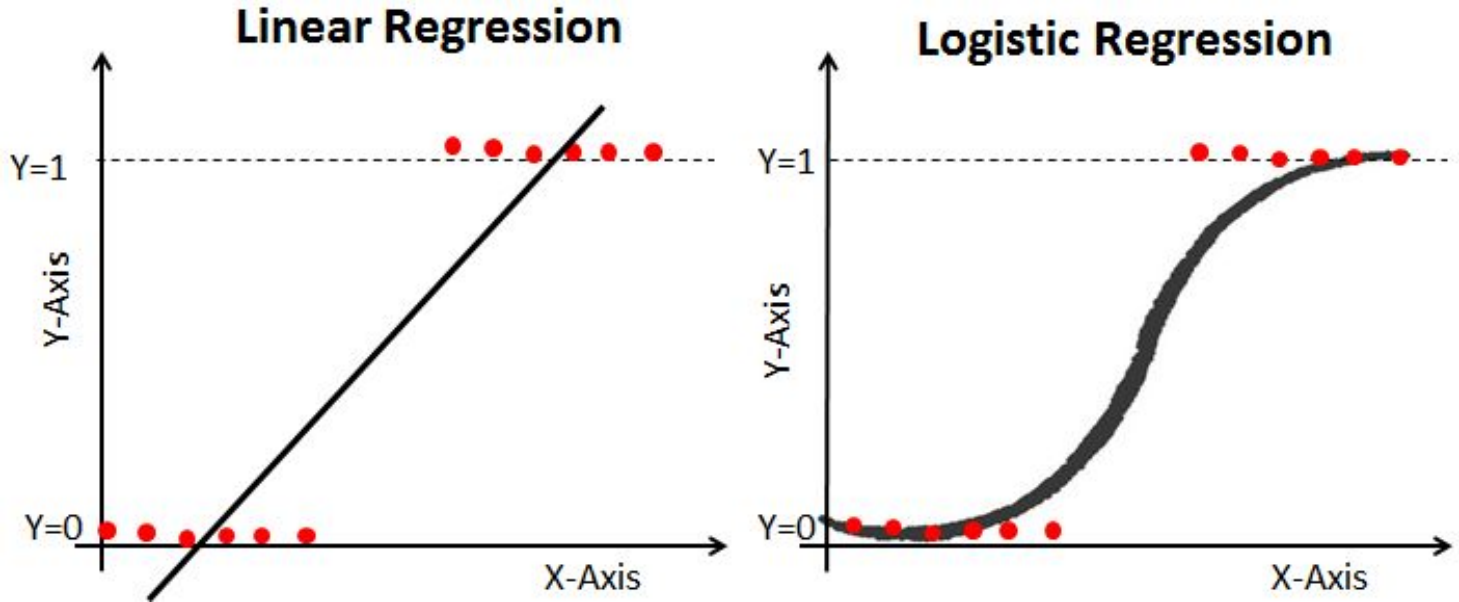
label = faulty일 것
이라는 분류 추정이 가능

그런데, 범주형(Categorical) 분류 문제점

- 우리는 label을 색상으로 보고 알았음.
- 컴퓨터에겐 0 또는 1 숫자로 알려줘야하는 것이 포인트

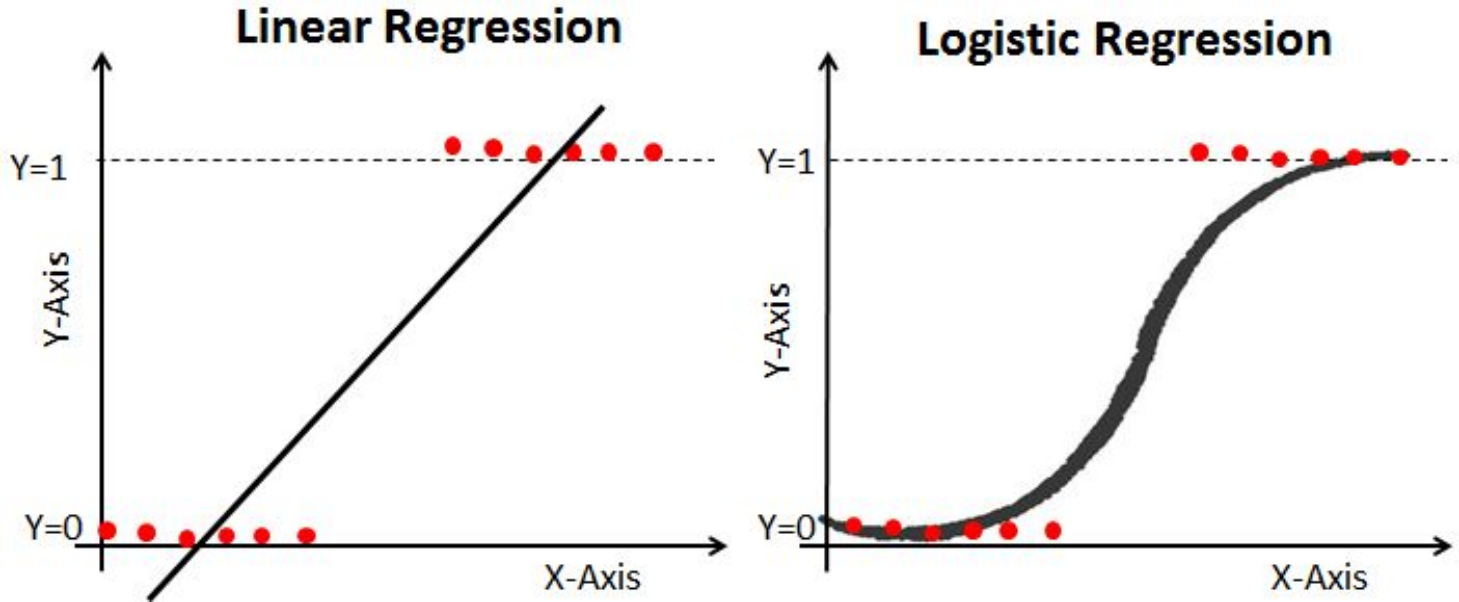
그런데, 범주형(Categorical) 분류 문제점

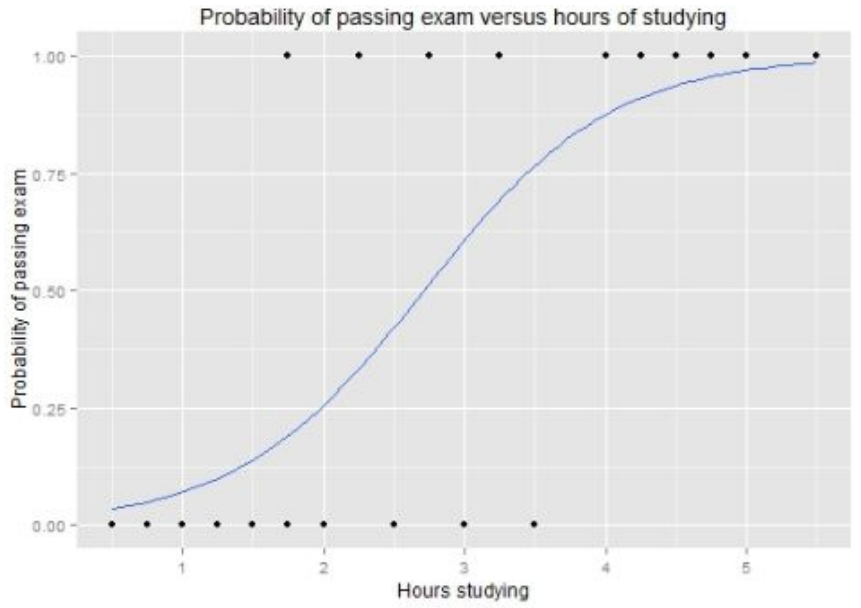
- 우리는 label을 색상으로 보고 알았음,
- 컴퓨터에겐 0 또는 1 숫자로 알려줘야하는 것이 포인트
- 선형 회귀와 연결하느라 feature 2개로 label 분류 예제
- feature가 x-axis (RPM) 1개 뿐이고, label을 Y = 1 or 0으로!



그런데, 범주형(Categorical) 분류 문제점

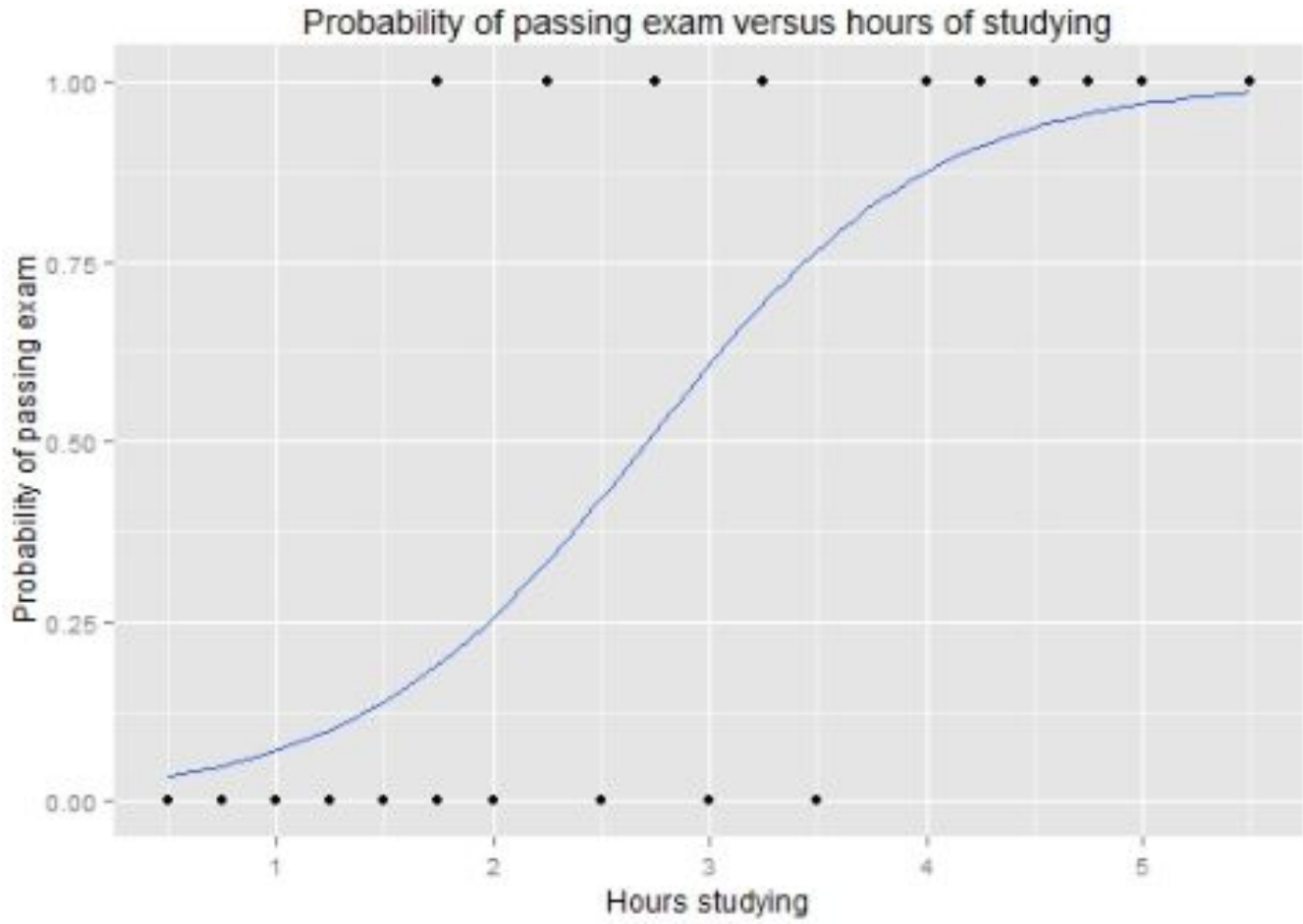
- 우리는 label을 색상으로 보고 알았음,
- 컴퓨터에겐 0 또는 1 숫자로 알려줘야하는 것이 포인트
- 선형 회귀와 연결하느라 feature 2개로 label 분류 예제
- feature가 x-axis (RPM) 1개 뿐이고, label을 Y = 1 or 0으로!





그런데, 범주형(Categorical) 분류 문제점

- 혹시 어색하셔도 당연
- 왜냐, 범주 나누는 것
  자체가 컴퓨터에 어색.



70점이 Pass 기준이면,
30점 fail과
60점 fail을 동일하게
보는 것이 사회 기준,
이것을 컴퓨터에 이해시켜야 함

Confusing words in Categorical 분류

Logit

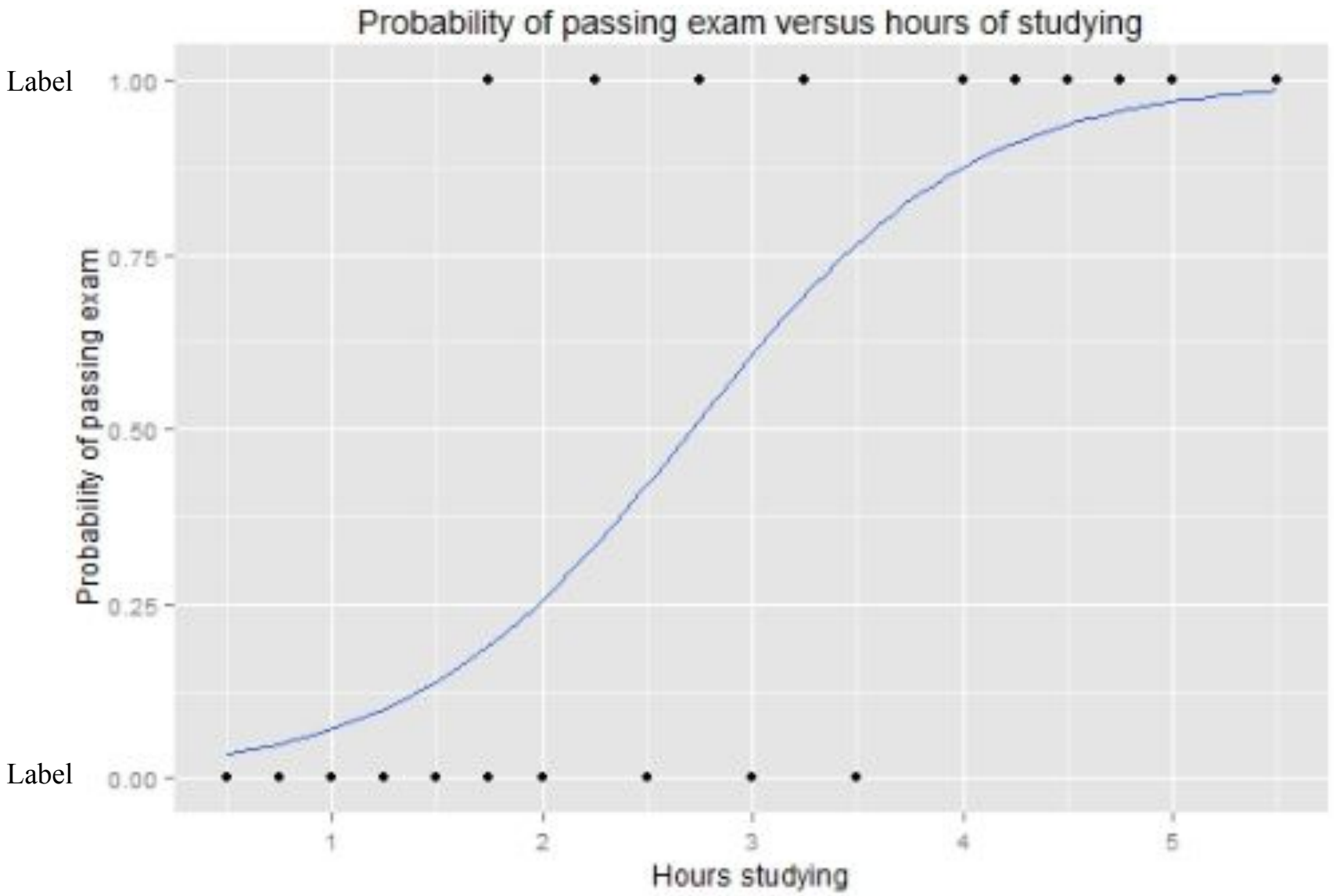
Logistic Unit

Logistic Regression

Sigmoid

Odds

범주를 분류하기 위한 함수



어떤 사건이 일어날 확률 (0과 1사이로)

$$P(X) \qquad 1 - P(X)$$

일어날 확률 **일어나지 않을 확률**

시험 Pass
질병 확진

$$0 \leq P(X) \leq 1$$

우리는 안 쓰던 Odds란 말을 영어권은 이렇게 쓰더군요

Odds

From Wikipedia, the free encyclopedia

This article is about the gambling and statistical term.

Odds To Win 2018 FIFA World Cup

COUNTRY	ODDS TO WIN
Germany	9-2
Brazil	5-1
France	11-2
Spain	7-1

https://www.espn.com/chalk/story/_/id/20991480/soccer-odds-win-2018-fifa-world-cup

Odds Ratio = 오즈비

- 오즈비 = Odds / Odds

- Odds = 불균형 빈도비 = 불균형 확률비

  = 고양이3 / 개7 (한 마리씩 나타나는데, 두가지 펫만 있다면)

  = a / b = 3 / 7

  = { a/(a+b) } / { b/(a+b) } = 0.3 / 0.7

고양이 $P(X)$ 일어날 확률 $1 - P(X)$ 일어나지 않을 확률 고양이 제외 $\frac{P(X)}{1 - P(X)}$

$0 \leq P(X) \leq 1$ = 개

Odds Ratio = 오즈비 = 승산?비
(승리의 발견율 비)

- 특정 조건의 승산(승리발견율) / 다른 조건에서 승산(승리발견율)**의 비율**

예) 클럽o, 확진자 확률 Odds 0.01/0.99

$$\frac{\frac{P(X)}{1-P(X)}\ \text{Odds (클럽o 조건)}}{\frac{P(X)}{1-P(X)}\ \text{Odds (클럽x 조건)}}$$

클럽x, 확진자 확률 Odds 0.04/0.96

$$\text{Odds Ratio= OR} = \frac{\text{P(disease|exposed)/ [1 – P(disease|exposed)]}}{\text{P(disease|unexposed)/ [1 – P(disease|unexposed)]}}$$

Odds Ratio = 오즈비 = 승산?비
(승리의 발견율 비)

- 특정 조건의 승산(승리발견율) / 다른 조건에서 승산(승리발견율)**의 비율**

[1] 환자-대조군 연구

이미 질환이 발생한 환자군과 질환이 발생하지 않은 대조군을 모집 후,
위험인자 노출 여부(특정 시점에서의 결과)를 후향적으로 조사하여 위험인자와 질환 발생 간의 연관성 추정.

이러한 경우에는 위험인자에 노출된 전체 모집단과 노출되지 않은 전체 모집단을 파악할 수가 없으므로
(특정 시점에서의 집단 수만 파악할 수 있기 때문에) 승산비를 사용할 수밖에 없다.

$$\text{Odds Ratio= OR} = \frac{P(\text{disease}|\text{exposed})/\,[1 - P(\text{disease}|\text{exposed})]}{P(\text{disease}|\text{unexposed})/\,[1 - P(\text{disease}|\text{unexposed})]}$$

Odds Ratio V. Relative Risk(Risk Ratio)

[1] 환자-대조군 연구

이미 질환이 발생한 환자군과 질환이 발생하지 않은 대조군을 모집한 후
위험인자 노출 여부(특정 시점에서의 결과)를 후향적으로 조사하여 위험인자와 질환 발생 간의 연관성 추정.
이러한 경우에는 위험인자에 노출된 전체 모집단과 노출되지 않은 전체 모집단을 파악할 수가 없으므로
(특정 시점에서의 집단 수만 파악할 수 있기 때문에) 승산비를 사용할 수밖에 없다.

[2] 코호트(Cohort) 연구

아직 질환이 발생되지 않은 모집단을 위험인자에 노출된 집단과
노출되지 않은 집단으로 구분하여 추적 관찰(시간적 개념 포함).
모집단을 파악 가능 → 위험인자(클럽 노출)와 질병 발생 간의 연관성을 추정.

상대적 위험도(Relative Risk, RR)

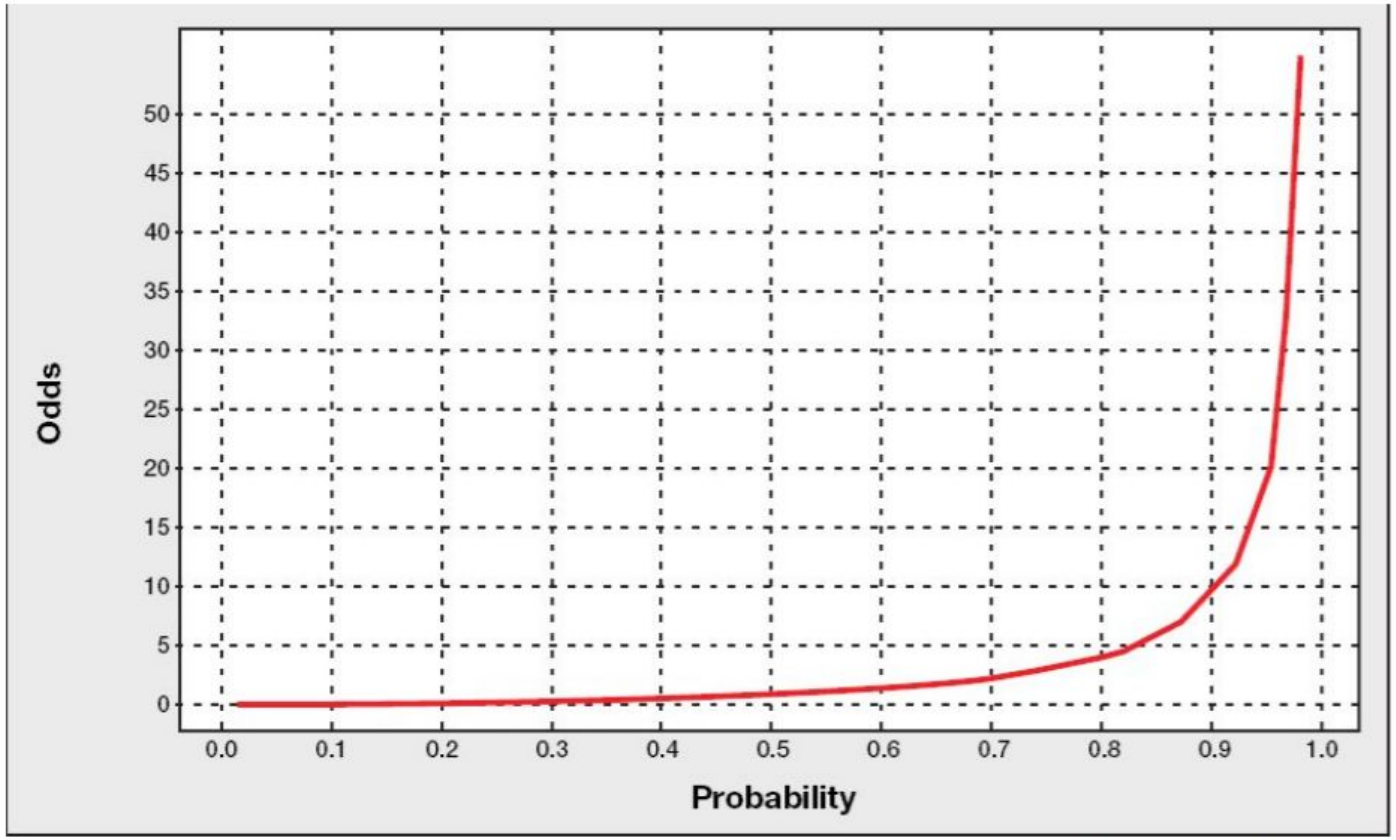
위험인자에 노출o 질병 발생 확률 / 위험인자에 노출x 질병 발생 확률로 나눈 값

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{\text{위험인자에 노출되었을 때 질병이 발생할 확률}}{\text{위험인자에 노출되지 않았을 때 질병이 발생할 확률}}$$

Odds Ratio = 오즈비 = 승산비

만약, 어떤 확률이 계속 증가한다면
Odds 오즈는 어떻게 될까요?

$$\frac{P(X)}{1-P(X)}$$



https://goo.gl/9a6zkS

In 1944, Joseph Berkson

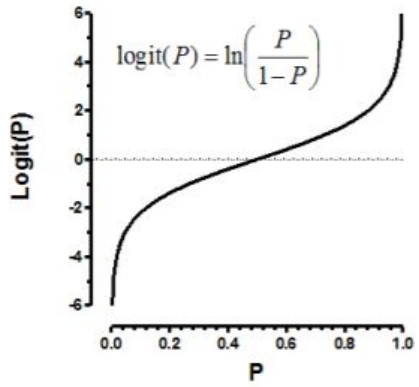
기호 논리학의

Logit link F. = Logistic unit link F.

= Log-odds link F. = Logit is a transformation

- Q: 뭐가 기호적이라는걸까요? A: 바로 그래프 모양 -무한, +무한

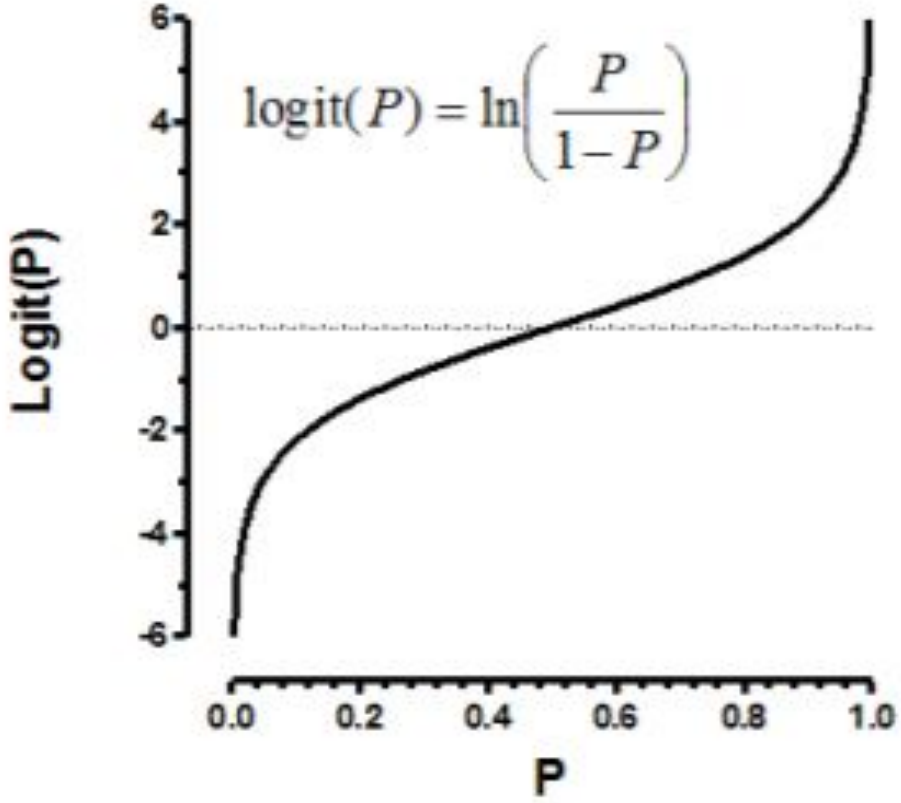
$$logit(p(y=1|x)) = log_e\left(\frac{p}{1-p}\right)$$



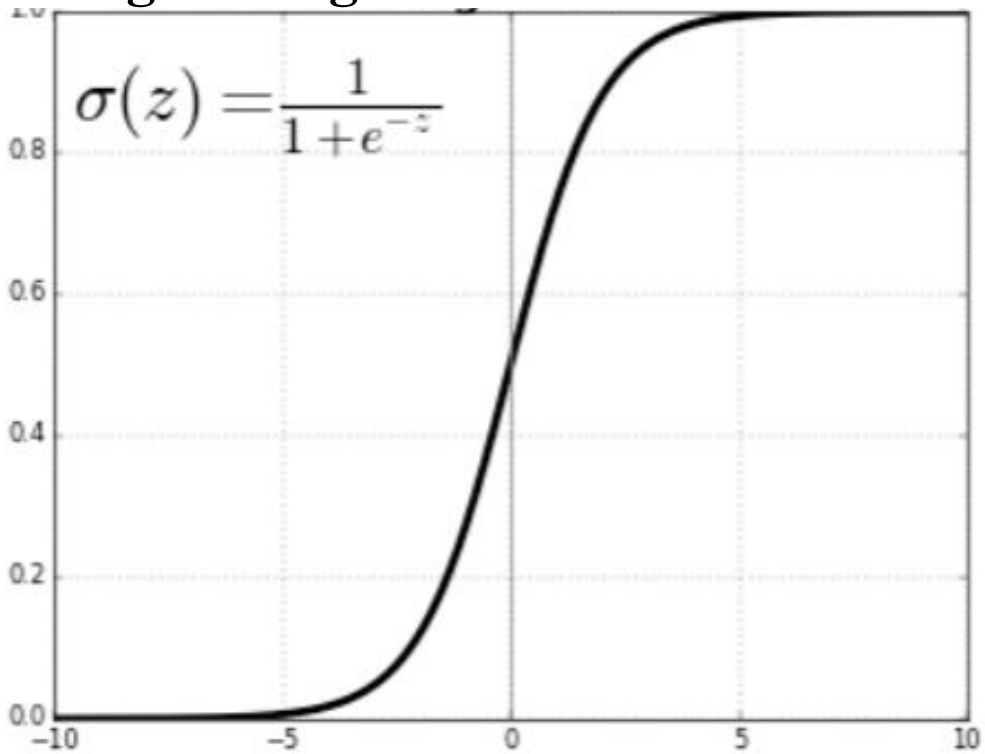
$$logit(P(Y=1|x_1,\ldots,x_k)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Inverse Logit
= Logistic Sigmoid F.

- Logit의 역함수 = Logistic Sigmoid vs 다른 S 모양은?



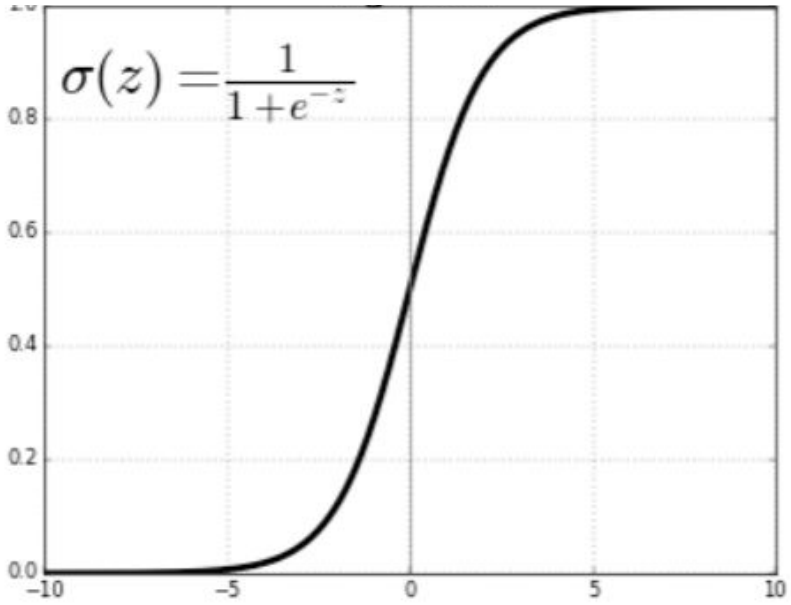
Logistic Sigmoid



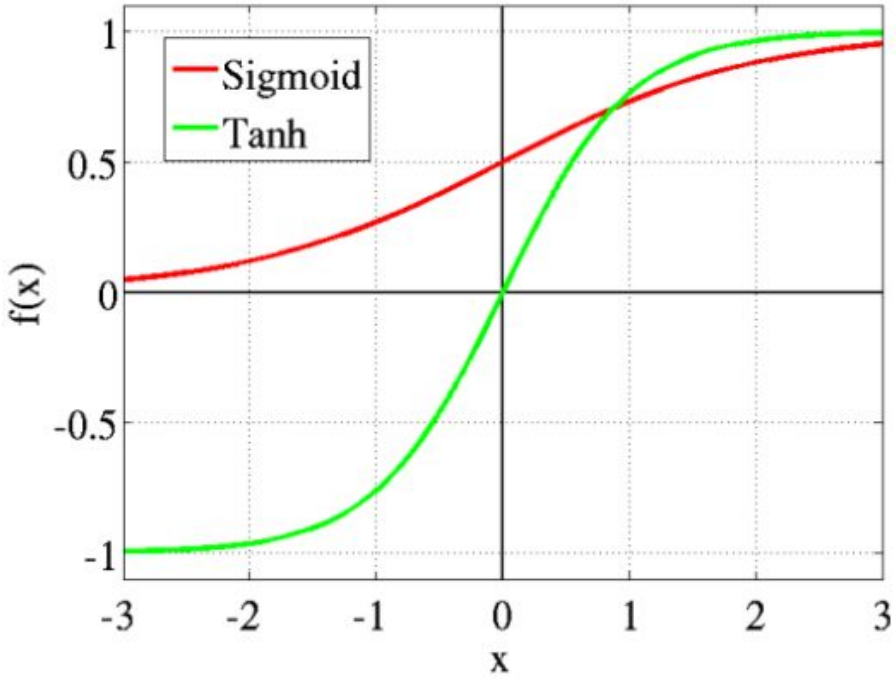
https://goo.gl/38SsHw

Inverse Logit
= Logistic Sigmoid F.

- **Logit의 역함수 = Logistic Sigmoid = S 모양의 로지스틱 함수**



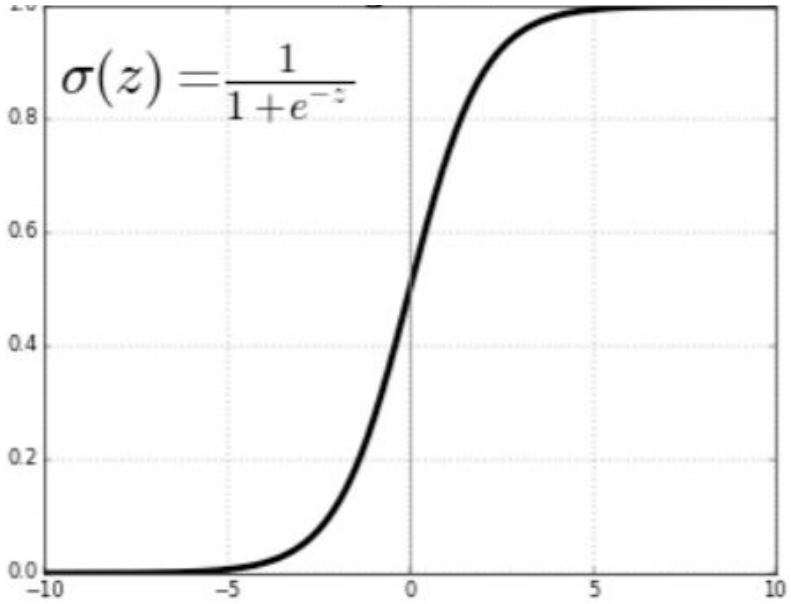
https://goo.gl/38SsHw



vs *tanh*

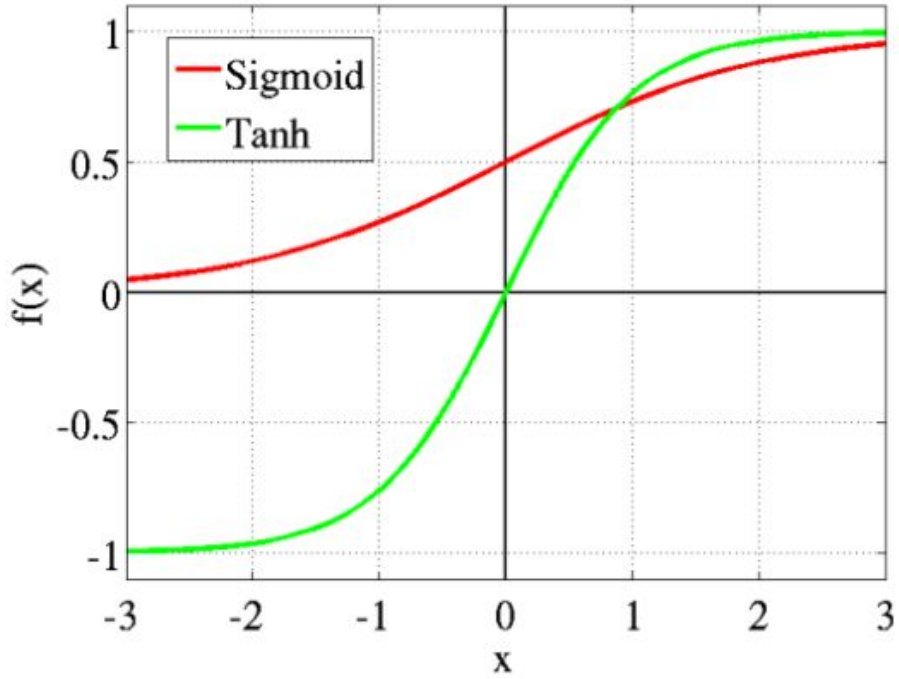
Inverse Logit
= Logistic Sigmoid F.

- **Logit의 역함수 = Logistic Sigmoid = S 모양의 로지스틱 함수**



https://goo.gl/38SsHw

vs ***tanh***



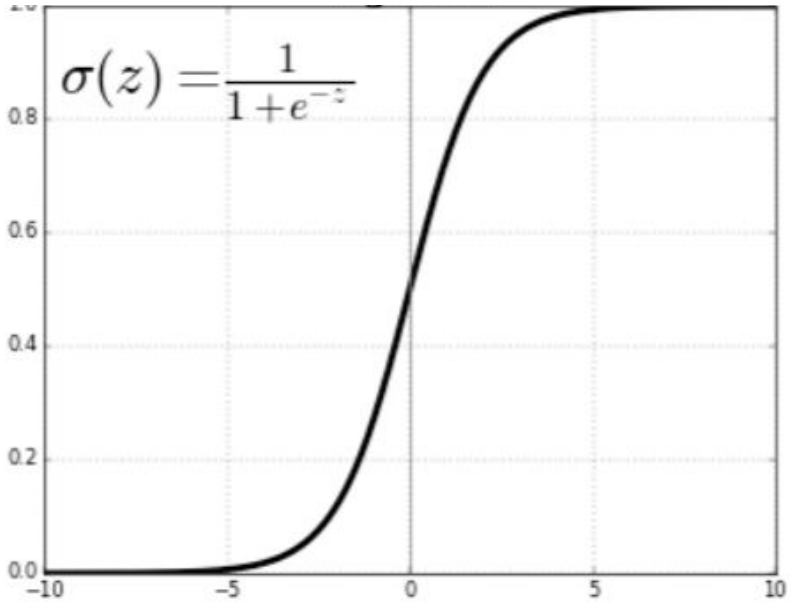
www.datamar

?

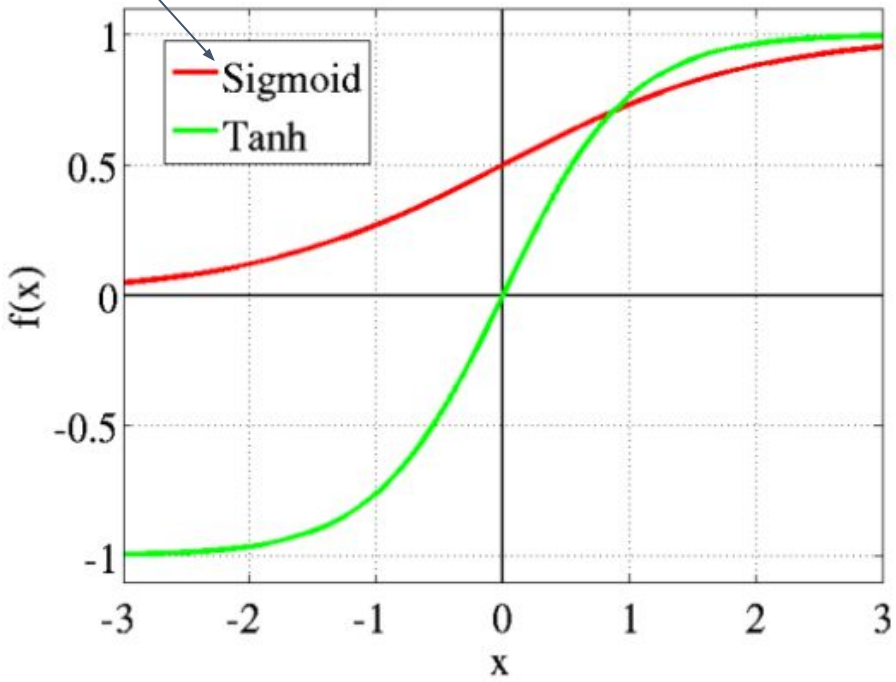
2019. 3. 12. - **tanh : sigmoid function**의 가중치 학습시 역전파된 gradient의 방향에 제약이 가해져 학습속도가 늦거나 수렴이 어렵게 되는 문제를 해결한 함수.

Inverse Logit
= Logistic Sigmoid F.

- **Logit의 역함수 = Logistic Sigmoid = S 모양의 로지스틱 함수**



https://goo.gl/38SsHw



vs *tanh*

Sigmoid

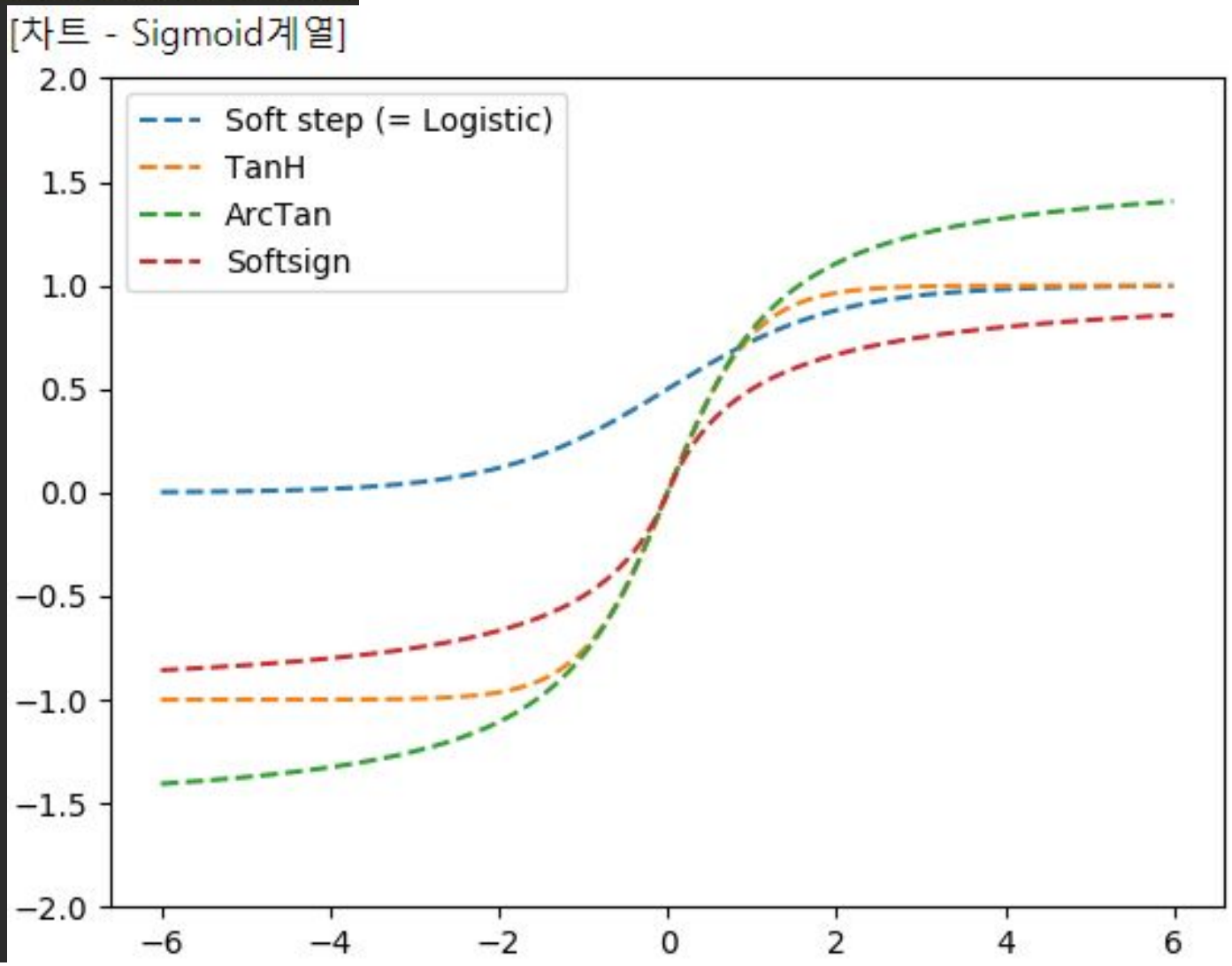
S 모양의 하이퍼볼릭 탄젠트 함수

www.datamar

2019. 3. 12. - **tanh : sigmoid function**의 가중치 학습시 역전파된 gradient의 방향에 제약이 가해져 학습속도가 늦거나 수렴이 어렵게 되는 문제를 해결한 함수.

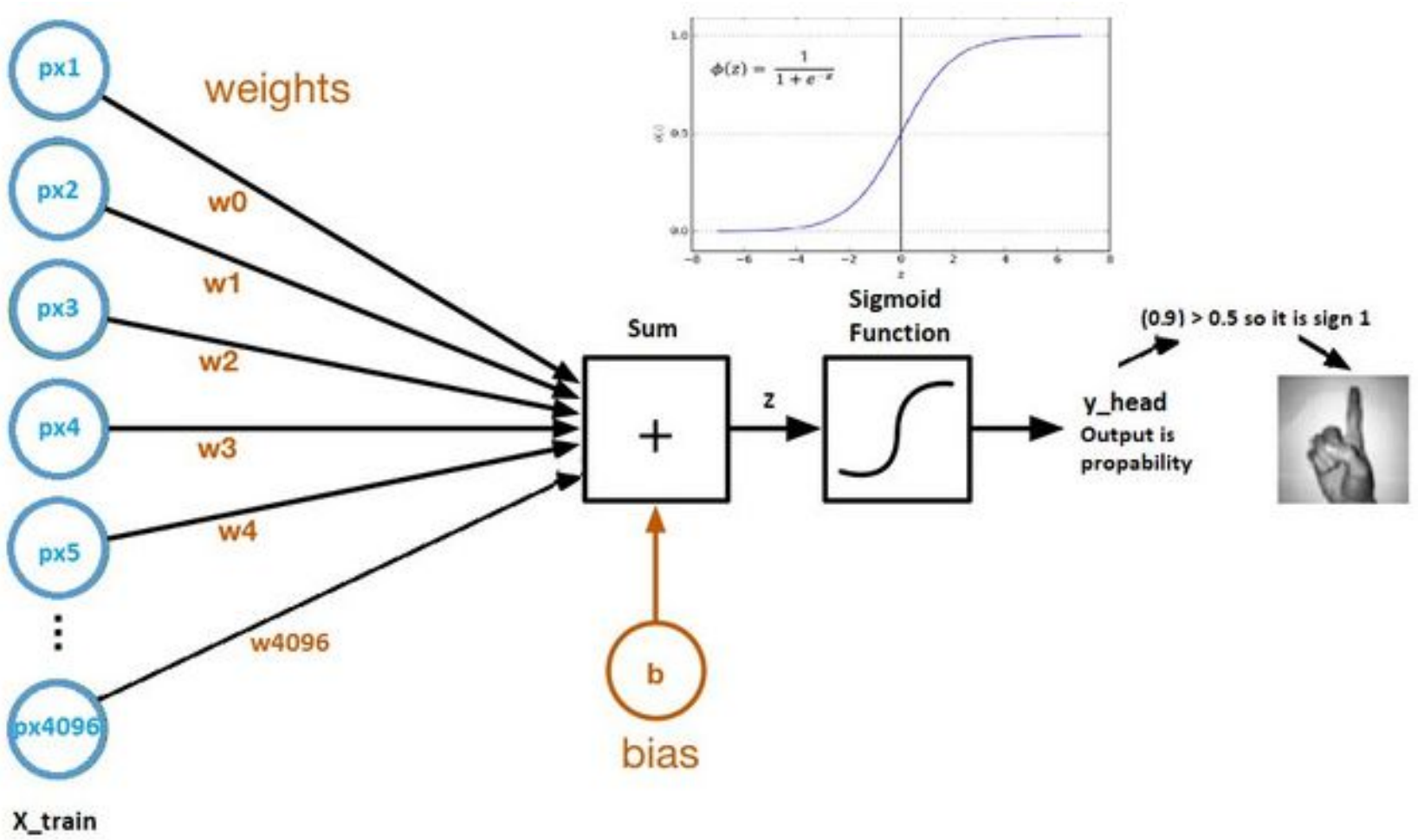
np 지원 그래프 plt.plot해보기

```
def softstep_func(x):  # Soft step (= Logistic), 시그모이드(Sigmoid, S자모양) 대표적인 함수
    return 1 / (1 + np.exp(-x))


# 그래프 출력
plt.plot(x, softstep_func(x), linestyle='--', label="Soft step (=


def tanh_func(x):  # TanH 함수
    return np.tanh(x)
    # return 2 / (1 + np.exp(-2*x)) - 1 # same


# 그래프 출력
plt.plot(x, tanh_func(x), linestyle='--', label="TanH")


def arctan_func(x):  # ArcTan 함수
    return np.arctan(x)


# 그래프 출력
plt.plot(x, arctan_func(x), linestyle='--', label="ArcTan")
```



Logistic Regression에서 Weight 학습하기

가설함수



http://preview.ibb.co/cxP63H/5.jpg

Iris 아이리스, 붓꽃

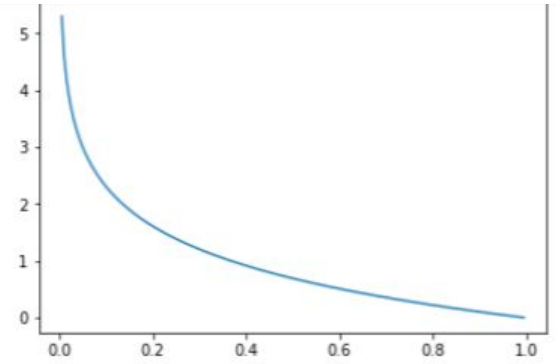
꽃 세가지 종류(Versicolor, Setosa, Virginica)의 꽃을

4가지 숫자 cm: Sepal(꽃받침) 길이, 폭, Petal(꽃잎) 길이, 폭

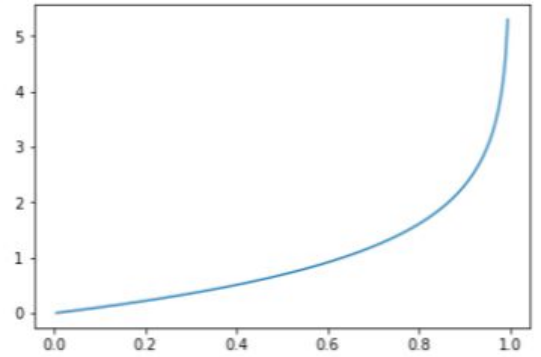
https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data

Sir Ronald Aylmer Fisher (1936)

Cost Function



$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$



Cost Function

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

$$= -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$$

미션!

$$\text{find } \theta, \text{ where } \min_\theta J(\theta) \qquad h_\theta(x) = \frac{1}{1 + e^{-\theta^T \mathbf{x}}}$$

Partial derivation of cost function

$$= -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log h_\theta(x^{(i)}) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))]$$

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[-y^i(\log(1+e^{-\theta x^i})) + (1-y^i)(-\theta x^i - \log(1+e^{-\theta x^i}))\right]$$

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y_i\theta x^i - \theta x^i - \log(1+e^{-\theta x^i})\right]$$

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T \mathbf{x}}}$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left[y_i\theta x^i - \log(1+e^{\theta x^i})\right]$$

$$-\theta x^i - \log(1+e^{-\theta x^i}) = -\left[\log e^{\theta x^i} + \log(1+e^{-\theta x^i})\right]$$

$$= -\log(1+e^{\theta x^i}).$$

Partial derivation of cost function

$$-\frac{1}{m}\sum_{i=1}^{m}\left[y_i\theta x^i - \log(1+e^{\theta x^i})\right]$$

$$z = w_0x_0 + w_1x_1 + \cdots + w_nx_n = \theta^T\mathbf{x}$$

θ 에 관하여 미분하면

$$\frac{\partial}{\partial\theta_j} y_i\theta x^i = y_i x_j^i$$

$$\frac{\partial}{\partial\theta_j}\log(1+e^{\theta x^i}) = \frac{x_j^i e^{\theta x^i}}{1+e^{\theta x^i}} = x_j^i h_\theta(x^i),$$

Partial derivation of cost function

$$-\frac{1}{m}\sum_{i=1}^{m}\left[y_i\theta x^i - \log(1+e^{\theta x^i})\right]$$

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T \mathbf{x}}}$$

$$\frac{\partial}{\partial\theta_j} y_i\theta x^i = y_i x_j^i$$

$$\frac{\partial}{\partial\theta_j}\log(1+e^{\theta x^i}) = \frac{x_j^i e^{\theta x^i}}{1+e^{\theta x^i}} = x_j^i h_\theta(x^i),$$

양성 음성 각각의 이상적인 y값에 가까우면 최적인 것과 통함.

$$\frac{\partial}{\partial\theta_j} J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)x_j^i$$

Weight update

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^i) - y^i) x_j^i$$

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

모든 θ_j 동시에 업데이트

$$:= \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^i) - y^i) x_j^i$$

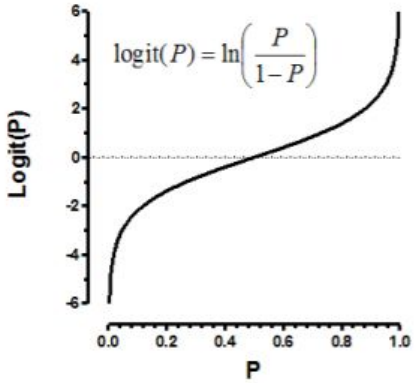
첨부) Logit의 역함수 = Logistic 도출 세부 수식

기호 논리학의 In 1944, Joseph Berkson

Logit link F. = Logistic unit link F. = Log-odds link F.

- Q: 뭐가 기호적이라는걸까요? A: 바로 그래프 모양 -무한, +무한

$$logit(p(y=1|x)) = log_e\left(\frac{p}{1-p}\right)$$



$$= log_e(p) - log_e(1-p)$$

$$= -log_e\left(\frac{1}{p} - 1\right)$$

$$logit(P(Y=1|x_1,\ldots,x_k)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Inverse Logit
= Logistic Sigmoid Function

- **Logit 함수의 역함수로 z에 관한 확률을 산출**

$$f(z) = y = -log_e\left(\frac{1}{z} - 1\right)$$ 역함수로 바꾸면

$$z = -log_e\left(\frac{1}{y} - 1\right)$$ y에 관한 정리

Logistic Sigmoid Function

- **Logit 함수의 역함수로 z에 관한 확률을 산출**

$$z = -log_e\left(\frac{1}{y} - 1\right)$$

y에 관한 정리

$$e^{-z} = \frac{1-y}{y}$$

$$y * e^{-z} + y = 1$$

$$y(e^{-z} + 1) = 1$$

$$y = \frac{1}{1+e^{-z}}$$

Logistic Function =
Inverse of logit function

Logistic Sigmoid Function

- **Sigmoid function 으로 변환**

$$p = \sigma(z) = \frac{1}{1 + e^{-z}}, \quad \frac{p}{1-p} = \frac{\frac{1}{1+e^{-z}}}{\frac{e^{-z}}{1+e^{-z}}} = \frac{1}{e^{-z}} = e^{z}$$

$$log_e \frac{p}{1-p} = z$$

$$log_e \frac{p}{1-p} = z = w_0 x_0 + w_1 x_1 + \cdots + w_n x_n$$