분류모형
- 확률
  - 생성
  - 판별 ⇒ 로지스틱 회귀
- 판별

[확률적 판별 모형]

$\hat{y} = 0$ or 1 분류하고자 할때, $P(y|x)$ 값을 "베이즈 정리가 아닌" x의 함수값으로 가정. 판별함수를 찾아 데이터를 0 or 1로 분류하자는 방법론

(데이터) $f(x)$, 판별함수 x_new

ex) $f(x)$ = 판별식

$f(x) = 0 \iff$ 인 x 집합은 경계선 (데이터)

$f(x) > 0 \iff$ 인 x 집합은 클래스 1 (데이터) $\hat{y} = 1$

$f(x) < 0 \iff$ 인 x 집합은 클래스 0 (데이터) $\hat{y} = 0$

[로지스틱 회귀모형]

① 일단, 이진분류 ($\hat{y} = 0$ or 1 만 고려)

② y는 이진클래스 값. 베르누이 확률분포를 따른다.
로지스틱 회귀모형에선, 이러한 y를 다음처럼 가정.

$$y \sim \mathrm{Bern}(y; \mu(x))$$
$$y \sim \mathrm{Bern}(y; w^T x) = \mathrm{Bern}(y; \mu(x; w))$$

$\hat{y}$ (추정값) $\fallingdotseq E[y] = \mu(x; w)$ = x에 대한 함수

* Regression 에선, $\hat{y} = \mu = E[y] = w^T x$
(지금은 $\hat{y}$는 0 or 1로 이진 분류!?)

* 기대값 = 모수 = $\mu(x; w) = w^T x$ = 독립변수 x의 선형결합
(데이터)
$\fallingdotseq$ 확률론적 선형회귀

* x (데이터)는 주어졌고, y 분포 (Bern)의 모수는 'w' 이다.

③ $y \sim \mathrm{Bern}(y; \mu(x;w)) = \mu(x;w)^y (1-\mu)^{1-y}$

$\hat{y} \doteq E[y] = \mu(x;w)$ > 0.5 라면, $\hat{y} = 1$
〃 < 0.5, $\hat{y} = 0$
〃 $= 0.5$, 경계선

"$W^Tx = 0.5$" 일 때의
x 집합
(직선, 선형결합)

④ 로지스틱 회귀모형의 분류 "판별식"
$= \mu(x;w) = w^Tx$
$\therefore \mu(x;w) = w^Tx$ 를 찾아야 한다.

시그모이드!
로 위아래
막기!

⑤ $\mu(x;w) = w^Tx \Longrightarrow \mu(x;w) = \sigma(w^Tx)$
시그모이드 함수

로지스틱
하이퍼 볼릭 탄젠트
탄자

* $\mu(x;w)$는 베르누이분포의 기대값이자 "확률값"
w^Tx 는 $-\infty < \; < \infty$ $\neq$ $0 \le \; \le 1$
$\therefore \sigma(w^Tx)$는 $0 < \; < 1$

(o)

$\mu(x;w) \neq w^Tx$
$\mu(x;w) = \sigma(w^Tx) = \sigma(z)$

* σ = 시그모이드 함수 (대부분이 로지스틱), $\sigma(z) = \dfrac{1}{1+e^{-z}}$
(= S-like)

⑥ 로지스틱 회귀모형의 분류 "판별식"
$\Rightarrow$ x값에 따라 판별식 값 결정

$w^Tx = z$ $= 0 \Longrightarrow \sigma(z) = \mu = \frac{1}{2}$ 경계선
$> 0 \Longrightarrow$ $> \frac{1}{2}$ $\hat{y} = 1$
$< 0 \Longrightarrow$ $< \frac{1}{2}$ $\hat{y} = 0$

이제
모델(w) 찾아야함

입력되는 데이터

[로지스틱 회귀분석 모형]

⑦ 문제는 $x \xrightarrow{W} z$ W가 결정되지 않음!
(판별 모델 자체가 아직...)

$W^T x$ 의 모수(W)를 MLE로 추정.

* MLE : y의 가능도를 maximize 하는 μ값(모수, 분포) 찾기
$L(\mu; y)$
= 확률분포 찾기

< MLE >

$L(\theta; x_1, \ldots x_n)$
$= p(x_1, \ldots x_n; \theta)$
변수 모수(상수로 가정) = 분포

$p(y|x) = pdf = p(y; \mu) = \mu^y (1-\mu)^{1-y}$
$\Longleftrightarrow$ 의미상으로 μ가 정해졌을 때, y값 = 확률값
$\mu_0^y (1-\mu_0)^{1-y}$

$L(\mu; y) = \mu^y (1-\mu)^{1-y}$
$\Longleftrightarrow$ y가 정해졌을 때 (결과가 정해짐) μ값 = 가능도
$\mu^{y_0} (1-\mu)^{1-y_0}$

어느분포에서 (μ) 가능도↑ ?
(= 이 결과 값일 때, (y) 가장 많은 μ는 무엇인가? $= \arg\max_{\mu} L(\mu; y)$)
= y가 나왔는데, 어느 분포에서 (μ) 나온건가?

* $\mu = \mu(x; w)$

⑧ $L(\mu(x; w); y) = \mu^{y_0} (1-\mu)^{1-y_0}$

⑨ $L \longrightarrow LL \longrightarrow \dfrac{\partial LL}{\partial w}$ $\Longleftrightarrow$ y를 배출한 가장 가능성 높은 분포 찾기 (모수 μ, w는 일종의 하이퍼 모수)

x, y가 모두 복수데이터
x는 한 확률변수에서 나온 데이터 (반복시행)이므로 독립
∴ L → 단변수함수로 변환 → LL (로그가능도 함수화)

⑩ $\frac{\partial LL}{\partial w} = 0$ 으로 만드는 w 찾기

$= J$ = 그레디언트 벡터

= 비선형... why?

$a(x; w) = \sigma(w^T x)$

= 비선형...

∴ 수치적 최적화로 w값 찾기.

"최적화 ⟺ $J = 0$ 인 x 찾기"

- 수도인버스
- 수치적 최적화

⑪ 수치적 최적화

< SGD >

$$w_{k+1} = w_k - h_i \times J$$

$$= w_k - h_i \times \frac{\partial LL}{\partial w}$$

결국은...

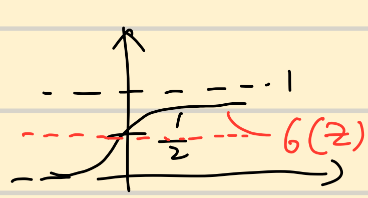
⑫ w값 찾기 ⟺ 모델찾기 (판별식!)

⇓

찾은모델로 x_{test} 받아서

$w^T x = z$ vs 0

↓



$\sigma(z)$ vs $\frac{1}{2}$

↓

$\hat{y} = ?$

[로지스틱 회귀 성능측정]

R^2_{suedo} = log loss 이용한 맥파든 R^2 (worst = "$R^2_{suedo}=0$" best = "$R^2_{suedo}=1$")

로지스틱 회귀 성능은 맥파든 의사결정계수(McFadden pseudo R square)값으로 측정한다.

R^2_{suedo} = log loss를 이용한 맥파든 R^2

$$R^2_{\text{pseudo}} = 1 - \frac{G^2}{G_0^2}$$

worst : $R^2_{suedo} = 0$
best : $R^2_{suedo} = 1$

G^2는 이탈도(deviance)라고 하는 양으로 다음과 같이 정의된다.

$$G^2 = 2\sum_{i=1}^{N}\left(y_i \log\frac{y_i}{\hat{y}_i} + (1-y_i)\log\frac{1-y_i}{1-\hat{y}_i}\right) \Longrightarrow$$

log loss cross entropy 와 부호만 다름

여기에서 $\hat{y}$는 $y=1$일 확률 μ를 뜻한다.

$\hat{y} = y = \mu = 1$

$$\hat{y}_i = \mu(x_i)$$

이탈도는 모형이 100% 정확한 경우에는 0이 되고 모형의 성능이 나빠질수록 값이 커진다. 부정확할 수록 '0'보다 커진다.

또한 이탈도는 로그 가능도에 음수를 취한 값과 같다.

$$G^2 = -LL$$

* null model = 가장 쓸모없는 모델 = 귀무모델 = random model = x이 상관없이 y예측하는 모델

G^2는 현재 이탈도이고 G_0^2는 귀무모형(null model)으로 측정한 이탈도다. 귀무이탈도

귀무모형이란 모든 x가 y를 예측하는데 전혀 영향을 미치지 않는 모형을 말한다. 즉, 무조건부 확률 $p(y)$에 따라 x에 상관없이 동일하게 y를 예측하는 모형을 말한다. 결국 우리가 만들 수 있는 가장 성능이 나쁜 모형이 된다. (결과값)

$$\mu_{\text{null}} = \frac{\text{number of } Y = 1 \text{ data}}{\text{number of all data}}$$

= 단순히 1의 갯수에 따라 μ 측정
x가 뭐가 들어오든 0 or 1 (데이터 갯수 많은 쪽)

따라서 맥파든 의사결정계수는 가장 성능이 좋을 때는 1이 되고 가장 성능이 나쁠 때는 0이 된다.

G^2이 가질수 있는 가장 나쁜 성능모델의 값 = G_0^2 ∴ 가장나쁜모델의 $R^2_{suedo} = 1-1 = 0$