< 한산한 머신러닝 >

Date. No.

Chapter 3. 선형모델

*3.1 기본 형식

$x = (x_1; x_2; \dots; x_d)$

선형모델 ; 속성들의 linear combination 으로 예측하는 함수.

$$f(x) = w_1x_1 + w_2x_2 + \cdots + w_dx_d + b$$

$$= w^Tx + b \quad \text{where } w = (w_1; w_2; \dots; w_d)$$

* non-linear model → 선형모델을 기반으로 층을 쌓거나 고차원으로 수행하여 얻을 수 있음.

* 해석능력 comprehensibility (이해가능성 understandability)
→ w은 예측에 있어 각 속성의 중요성을 직관적으로 드러냄

*3.2 linear Regression.

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\} = \{(x_i, y_i)\}_{i=1}^{m}$$

$x_i = (x_{i1}; x_{i2}; \dots; x_{id})$, $y_i \in \mathbb{R}$

first, $d=1$.

이 discrete feature,

case 1. ordered ⇒ Label Encoding

case 2. else ⇒ One Hot Encoding

↳ 이를 순서에 해당하면 정확히 같은 순서 관계를 가지고 거리(계산) 등에 영향을 미침!

(○ 평균 제곱오차 (square loss)
→ Euclidean Distance 사용.

∨ $E(w,b) = \sum_{i=1}^{m} (y_i - (xw_i - b))^2$; w와 b에 대한 convex function.

$[a, b]$에서 연속된 함수 f에 대해, 만약 구간 내 두 점 x_1, x_2가

$$f\left(\frac{x_1+x_2}{2}\right) \le \frac{f(x_1) + f(x_2)}{2}$$

을 만족하면 f는 구간 $[a, b]$ 상의 convex function 이라고 함.

Date. No.

Second, $d \geq 2$, Multivariate Linear Regression

목표: $f(\vec{x}_i) = \vec{w}^T \vec{x}_i + b \rightarrow f(\vec{x}_i) \simeq y_i$

Let $\hat{w} = (w; b)$, 즉 X를 set. $X \in \mathbb{R}^{m \times (d+1)}$

$$\hat{w}^* = \arg\min_{\hat{w}} \underbrace{(y - X\hat{w})^T (y - X\hat{w})}_{E_{\hat{w}}}$$

$$\frac{\partial E_{\hat{w}}}{\partial \hat{w}} = 2X^T (X\hat{w} - y)$$

*Caution! $X^T X$: full-rank or positive definite,

$$\hat{w}^* = (X^T X)^{-1} X^T y.$$

→ 현실에서 full-rank 인 경우는 드물!!

(ex) 생물 정보학, 유전자 데이터 (천만개씩 되는 특성 / 몇십 ~ 몇백의 샘플)

-※ Reference, Connection between rank and positive-definite / math-stackexchange

✓ 한번 $X^T X \in M_n(\mathbb{R})$, full-rank가 아니면, $\lambda_i = 0$ 인 고유치 존재, 즉, PD X.

✓ 위 조건에서, nontrivial kernel 을 가짐.

$\Rightarrow \exists\, v \neq \vec{0}$ s.t. $Mv = 0$.

$v^T M v = 0$. Hence, PD X.

✓

$\ln y = w^T x + b$ log-linear function

$y = g^{-1}(w^T x + b)$ where g (→ link function) monotone differentiable function

generalize linear model

Date. No.

※ 3.3 Logistic Regression.

분류문제 → "GLM", 단조미분함수 h를 찾아 분류 레벨 y와 선형 회귀 모델의 예측값을 연결하면 OK!

- Heaviside function (단위-계단 함수 unit-step function)

$$y = \begin{cases} 0, & z<0 \\ 0.5, & z=0 \\ 1, & z>0 \end{cases}$$ → 불연속!!

cf, heaviside function과 근사한 sigmoid function (연속함수)
monotone differentiable

- sigmoid function → $x \to -\infty,\ f(x)=0,\ x \to \infty,\ f(x)=1$.
  - v logistic function

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

GLM에 넣기! $y = \dfrac{1}{1+e^{-(w^T x+b)}}$

$$\ln \underbrace{\frac{y}{1-y}}_{\text{odds}} = w^T x + b$$

log odds (logit) → 선형 예측값을 확률로 근사!

logistic regression! 분류학습법!

① 사전 데이터 분포에 대한 가정 필요 X.
② 근사확률에 대한 예측 가능.
③ Solution의 목적함수가 convex function!
→ 수학적으로 완전히 Good!!

Date. No.

let $y = p(y=1 \mid x)$. then,

$$\ln \underbrace{\frac{p(y=1 \mid x)}{p(y=0 \mid x)}}_{\text{odds}} = w^T x + b$$

$$p(y=1 \mid x) = \frac{e^{w^T x + b}}{1 + e^{w^T x + b}}$$

$$p(y=0 \mid x) = \frac{1}{1 + e^{w^T x + b}}$$

w와 b를 추정하기 위해, 최대 우도법 maximum likelihood method 사용!

$$\ell(w, b) = \sum_{i=1}^{m} \underbrace{\ln p(y_i \mid x_i; w, b)}_{\text{log-likelihood}}$$

log-likelihood 최대화하자!

why?

$$p(y_i \mid x_i; w, b) = y_i p_1(\hat{x}_i; \beta) + (1 - y_i) p_0(\hat{x}_i; \beta)$$

$$\ell(\beta) = \sum_{i=1}^{m} \left(-y_i \beta^T \hat{x}_i + \ln\left(1 + e^{\beta^T \hat{x}_i}\right) \right)$$ minimize!

where $\beta = (w; b)$, $\hat{x} = (x; 1)$

$$\left(\because y_i \cdot \frac{e^{\beta^T \hat{x}}}{1 + e^{\beta^T \hat{x}}} + (1 - y_i) \cdot \frac{1}{1 + e^{\beta^T \hat{x}}}\right.$$

$$\left. = \frac{y_i}{1 + e^{\beta^T \hat{x}}} \left(e^{\beta^T \hat{x}} - 1\right) + \frac{1}{1 + e^{\beta^T \hat{x}}} = \frac{1}{1 + e^{\beta^T \hat{x}}} \left(y e^{\beta^T \hat{x}} - y_i + 1\right)\right)$$

Date. No.

※ 3.4 Linear Discriminant Analysis

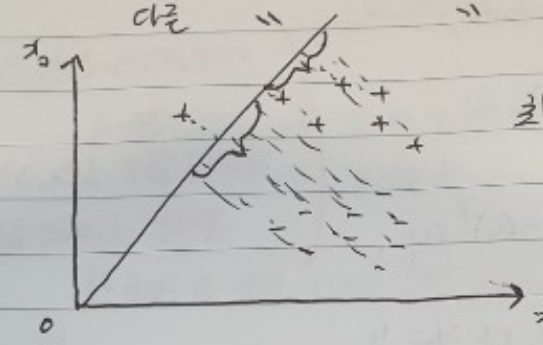
↳ 각 클래스 샘플의 공분산 행렬이 같고 full rank.

↳ Fisher 판별분석이라고도 함 (이는 위의 가정을 따르지 X)

Idea : 훈련 데이터 샘플을 하나의 직선에 투영.

같은 class에 속하는 sample을 가능한 가까이 투영점에

다른 " " 최대한 먼 위치에!



최대한 가깝게 vs 최대한 멀게!

Let $D = \{(x_i, y_i)\}_{i=1}^{m}$, $y_i \in \{0, 1\}$

X_i, μ_i, Σ_i ; i번째 클래스 집합, 평균벡터, 공분산 행렬

① 같은 클래스 샘플 최대한 투영시 가깝게

$$\text{minimize } w^T\Sigma_0 w + w^T\Sigma_1 w$$

(∵ 직선은 1-d, 모두 실수!)

② 다른 클래스 샘플 투영시 최대한 멀게

$$\text{maximize } \|w^T\mu_0 - w^T\mu_1\|_2^2$$

$$\text{maximize } J = \frac{\|w^T\mu_0 - w^T\mu_1\|_2^2}{w^T\Sigma_0 w + w^T\Sigma_1 w} = \frac{w^T(\mu_0-\mu_1)(\mu_0-\mu_1)^T w}{w^T\Sigma_0 w + w^T\Sigma_1 w}$$

집합 내 산포행렬 within-class scatter matrix

$$S_w = \Sigma_0 + \Sigma_1 = \sum_{x \in X_0} (x-\mu_0)(x-\mu_0)^T + \sum_{x \in X_1} (x-\mu_1)(x-\mu_1)^T$$

집합 간 산포행렬 between-class scatter matrix

$$S_b = (\mu_0-\mu_1)(\mu_0-\mu_1)^T$$

S_b, S_w

$$J = \frac{w^T S_b w}{w^T S_w w}$$; S_b, S_w의 일반화된 레일리 몫 generalized Rayleigh quotient

Date. No.

In LDA (Latent Discrimination Analysis),
어떻게 w를 계산하는가?

WLOG, $w^T S_w w = 1$,

$$\min_w \; -w^T S_b w \quad s.t. \quad w^T S_w w = 1$$

by Lagrange multiplier methods, → 제약 최적화 문제 → 라그랑주 함수 $L(x, \lambda)$의 무제약 최적화 문제

$$S_b w = \lambda S_w w$$

Since $S_b w = (\mu_0 - \mu_1)\underbrace{(\mu_0 - \mu_1)^T w}_{\text{scalar}}$,

$S_b w$의 방향은 $\mu_0 - \mu_1$과 같음 !!

$$S_b w = \lambda(\mu_0 - \mu_1), \quad \boxed{w = S_w^{-1}(\mu_0 - \mu_1)}$$

$S_w = U \Sigma V^T \rightarrow S_w^{-1} = V \Sigma^{-1} U^T$

Singular value decomposition !

Date. No.

※ 3.5 다중 분류 학습

WLOG, Let $C_1, C_2, \dots, C_N$ classes.

기본 Idea = (분해법) { OvO, One vs One
OvR, One vs Rest
MvM, Many vs Many

① OvO

N개의 클래스 → 둘씩 묶어서 분류기 구축
→ $N(N-1)/2$ 개의 이진 분류기 생성.

② OvR

한 클래스만 양성, 나머지 음성 → N개의 분류기!

③ MvM

매번 몇 개의 클래스를 양성같이, 나머지 클래스를 음성같이 설정.

음? 어떻게?

v ECOC, 오류 수정 출력 코드 Error Correcting Output Codes

- 코딩 ; N개 클래스를 M개로 나눔 (N < M)
  + 훈련
- 디코딩 ; M개의 분류기는 각각 test sample에 예측을 진행
  해당 예측 레이블이 하나의 코드가 됨!

코딩 matrix → 이진 코드, 상한 코드

	f_1	f_2	f_3	f_4	f_5	해밍	유클리드
C_1	-1	+1	-1	+1	+1	3	$2\sqrt{3}$
C_2	+1	-1	-1	+1	-1	4	4
C_3	-1	+1	+1	-1	+1	1	2
C_4	-1	-1	+1	+1	-1	2	$2\sqrt{2}$
예측	-1	-1	+1	-1	+1		

(DAG Directed Acyclic Graph) 도 있음!

SVM에 대한 연구도!

Date. No.

※ 3.6 클래스 불균형 문제

* 앞선 모든 분류학습기의 ~~공통~~ 공통 기본가정!
(⇒ 서로 다른 클래스의 훈련 샘플들의 수가 같음!
ξ
이와 다르기 때문에 문제가 야기된다!

odds $\left(\frac{y}{1-y}\right)$: 양성값일 가능성과 음성값일 가능성의 비율!

$\frac{y}{1-y} \left(\frac{m^+}{m^-}\right)$ ↳ 관측 오즈 !! (과연 맞는 비율인가?)

$$\frac{y'}{1-y'} = \frac{y}{1-y} \times \frac{m^-}{m^+} \quad \cdots (1)$$

rescaling, rebalancing

* 관측 오즈로 실제 오즈에 대해 확신할 수는 X !
① under sampling
② over sampling (ex SMOTE, Interpolation 사용)
③ (1)을 결정과정에 사용. 임계값 이용 threshold-moving.