Linear Classification

Monday, 23 May 2022 10:47 AM

Classification

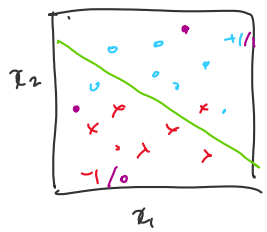
Supervised learning
$D = \{(x_n, t_n)\}_{n=1}^{N}$

- Regression: $t_n \in \mathbb{R}$
- classification: t_n is discrete
  - $\in \{$"dog", "cat"$\}$
  - $\in \{$"real", "fake"$\}$

2-classes

$D = \{(x_n, t_n)\}_{n=1}^{N}$

x_n: features
$t_n \in \{-1, +1\}$ (or $\{0, 1\}$ or "one hot" $\{[1, 0], [0, 1]\}$)



$t_n = [1 \quad 0]$ (class 0, class 1)

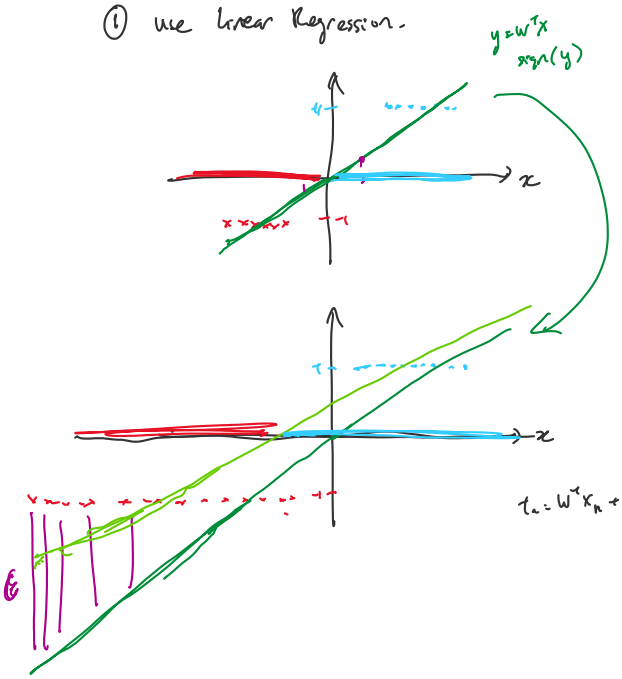
$[0 \quad 1]$ (class 0, class 1)

$[0 \quad 1 \quad 0]$ (class 0, class 1, class 2)

$\vdots$

(SLA – Super Lazy Approach)

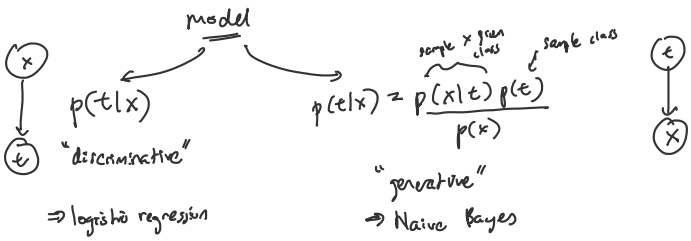
① use linear Regression.



$t_n = w^T x_n + \epsilon$, $\epsilon \sim \mathcal{N}(\epsilon \mid 0, \beta^{-1})$

---

① Design linear model for classification.
(clever lazy approach)



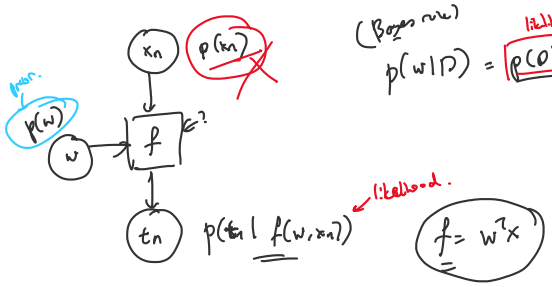
⇒ logistic regression

⇒ Naive Bayes

Discriminative Model

Logistic Regression (Sec 4.3 of PRML)

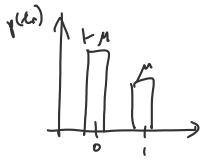
$D = \{(x_n, t_n)\}$. $t_n \in \{0, 1\}$. 2-class (Binary classification)



(Bayes rule) $p(w|D) = \frac{p(D|w)\,p(w)}{p(D)}$ (likelihood, prior)

prior: $p(w) = N(0, \alpha^{-1} I)$

likelihood. $p(D|w) = \prod_{n=1}^{N} p(t_n | x_n, w)$ (i.i.d)



$p(t_n | x_n, w) = \text{Bern}(\mu_n)$

$= \begin{cases} \mu_n & \text{if } t_n = 1 \\ 1-\mu_n & \text{if } t_n = 0 \end{cases}$

$= \mu_n^{t_n} (1-\mu_n)^{(1-t_n)}$

$\mathbb{E}[t_n] = \mu$

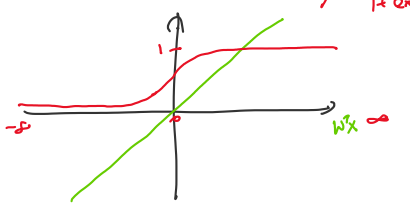
$\mathbb{V}[t_n] = \mu(1-\mu)$

$p(t_n | x_n, w) = \text{Bern}(\mu_n)$ μ_n? $\in [0, 1]$

$\mu_n \stackrel{?}{=} w^T x_n$ ✗

logistic function.

$\mu_n = \frac{1}{1+\exp(-w^T x_n)} = \sigma(w^T x_n)$



"squashing function"

$\sigma(z)$

$p(t_n = 1 | w, x_n) = \sigma(w^T x_n)$

$p(t_n = 0 | w, x_n) = 1 - \sigma(w^T x_n)$

$p(t_n | w, x_n) = \sigma(w^T x_n)^{t_n} (1 - \sigma(w^T x_n))^{(1-t_n)}$

---

② Est / inference.

- MLE $\max_w \lg p(D|w)$
- MAP $\max_w \lg p(w|D) = \lg p(D|w) + \lg p(w)$
- Bayes. $p(w|D)$

MAP

$$\max_w \lg \left[\prod_{n=1}^{N} p(t_n | x_n, w) \right] + \lg p(w)$$

$$= \min_w - \left[\sum_{n=1}^{N} \lg p(t_n | x_n, w) \right] - \lg p(w)$$

$\lg p(t_n | x_n, w)$

$\lg p(w)$

$\lg N(w | 0, \alpha^{-1} I)$

$= -\frac{\alpha}{2} w^T w$

$\lg \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(w-0)^2}{2\alpha^{-2}} \right) \right]$

$\lg [t_n (\quad)^{1-t_n}]$

$$= \lg\left[\underbrace{\sigma_n}_{\sigma_n=\sigma(w^T x_n)} \cdot (1-\sigma_n)\right]$$

$$= \lg \sigma_n^{t_n} + \lg(1-\sigma_n)^{1-t_n}$$

$$= t_n \lg \sigma_n + (1-t_n)\lg(1-\sigma_n) \quad \text{("cross entropy loss/error")}$$

What I want

$$\min_w \mathcal{L}(w)$$

$$\mathcal{L}(w) = -\sum_{n=1}^{N}\left[t_n \lg \sigma(w^T x_n) + (1-t_n)\lg(1-\sigma(w^T x_n))\right] + \frac{\lambda}{2} w^T w$$

regularizer

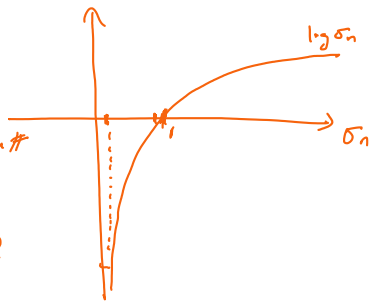
$$-\left[t_n \lg \sigma_n + (1-t_n)\lg(1-\sigma_n)\right]$$

case $t_n = 1$, would like $\sigma_n = ?$

if $\sigma_n \approx 1$ then $\lg \sigma_n \approx 0$

$\sigma_n \approx 0$ then $\ln \sigma_n$ = [negative #]

$\Rightarrow$ ↑ loss.



case $t_n = 0$, would like $\sigma_n = ?$

Solve!

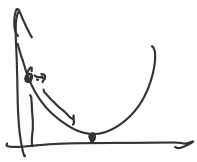
Attempt ①: $\frac{\partial \mathcal{L}}{\partial w} = 0$, Solve!

$\Rightarrow$ does not work (no closed form solution)

$\frac{\partial \mathcal{L}}{\partial w} = 0$

$w = (X^T X + \lambda \mathbb{I})^{-1} X^T t$

Attempt ②: gradient descent



$$\nabla_w \mathcal{L}(w)$$

$$= \nabla_w \left[-\sum_{n=1}^{N} t_n \log \sigma_n + (1-t_n)\lg(1-\sigma_n)\right] + \underbrace{\nabla_w\left(\frac{\lambda}{2} w^T w\right)}_{\text{exercise}}$$

$$= -\sum_{n=1}^{N} t_n \nabla_w \log \sigma_n + (1-t_n)\nabla_w(1-\sigma_n)$$

FACT:

$\sigma(z)$

$\frac{\partial \sigma}{\partial z} = \sigma(z)(1-\sigma(z))$

$$\nabla_w \lg \sigma_n = \nabla_w \lg \sigma(w^T x_n)$$

$$= \frac{1}{\sigma(w^T x_n)} \cdot \sigma(w^T x_n)(1-\sigma(w^T x_n)) \cdot x_n$$

$$= \underbrace{(1-\sigma(w^T x_n))}_{t_n=1} \cdot x_n$$

$$\nabla_w \lg(1-\sigma_n) = \nabla_w \lg(1-\sigma(w^T x_n))$$

$$= \frac{1}{1-\sigma(w^T x_n)}\left(-\sigma(w^T x_n)(1-\sigma(w^T x_n))\right) \cdot x_n$$

$$= \underbrace{(0 - \sigma(w^T x_n))}_{t_n=0} x_n$$

$$\Rightarrow (t_n - \sigma(w^T x_n)) x_n$$

$$\nabla_w \mathcal{L}(w) = \sum_{n=1}^{N} (\sigma(w^T x_n) - t_n) \cdot x_n + (\text{homework})$$

error.

alg:

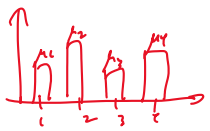
$$w_{t+1} = w_t - \frac{\eta}{N} \sum_{n=1}^{N} (t_n - \sigma_n) \cdot x_n$$

Extensions: 1) nonlinear ϕ_n

2) k-classes.

$p(w|D) \propto p(D|w)\, p(w)$

Categorical Distribution.



$\sum_k \mu_k = 1$

$\mu_k \geq 0$

t_n: $t_{n,1}$ $t_{n,2}$ $t_{n,3}$ $t_{n,4}$

$t_n = [0, 0, 1, 0]$

"one hot" rep / encoding

$$p(t_n | x) = \text{Cat}[\mu] = \prod_{k=1}^{K} \mu_k^{t_{nk}}$$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_K \end{bmatrix}$$

$$\mu_k = \frac{\exp(w_k^T x_n)}{\sum_k \exp(w_k^T x_n)}$$

"softmax"