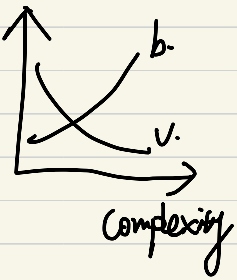
L5

Recap: — Bias vs Variance

— Thumb rule

low bias ⟶ higher complexity

low variance —— low complexity



How to define Complexity ?

Regularization:

—— Method to control complexity of a ML model.

Components: — Representation

— Loss function

— Optimization

Instead of MSE, we define an new loss function

$$L(w) = MSE(w) + \alpha\, \phi(w)$$

Scalar ≥ 0 ⟵ control level of regularization

"Regularizer"
promote simpler w
penalize "unlikely" w

加了 regularizer 之后, Loss function 求GD取最小时, 一同对 regularizer 取小, 以此达到 regularize 的作用, 比如 L2 norm 取小, W↓

Example:

1) $\phi(w) = \|w\|_2^2$ squared L2 norm

2) $\phi(w) = \|w\|_1$ L1 norm

3) $\phi(w) = \frac{1}{2}\|w\|_2^2 + \frac{1}{2}\|w\|_1$ 两种特性都兼具

"Elastic net"

4) $\phi(w) = \sum_{i=1}^{d} \left(w_i - \frac{1}{d}\sum_{j=1}^{d} w_j\right)^2$ 相邻相等

$\phi(w) = \sum_{i=1}^{d-1} (w_i - w_{i+1})^2$

Interpretation in terms of Linear Regression.

① $\boxed{L(w) = \frac{1}{2}\|y - Xw\|_2^2 + \alpha\|w\|_2^2}$

Set $\nabla L(w) = 0$ and solve for

$w^* = (X^TX + \alpha I)^{-1}X^Ty$ α do:

☆ "Ridge regression" larger α, W smaller, hence less variance

② $\boxed{L(w) = \frac{1}{2}\|y - Xw\|_2^2 + \alpha\|w\|_1}$

$\|w\|_1 = \sum_{i=1}^{d} |w_i|$ not differentiable

LASSO regression

Cannot optimize using GD but can use other algorithm

[sub-gradient descent, LARS, Fixed point

$\phi(w) = \|w\|_2^2 \longrightarrow$ shrinks value of w

$\phi(w) = \|w\|_1 \longrightarrow$ sparsifies w

most of coefficients are "0"

Logistic regression

Thus far : regression

Training dataset : (x, y)

Data $\rightarrow x$, $y \leftarrow$ label ( )

How to model classification ?

$y=0 \rightarrow$ NO

$y=1 \rightarrow$ YES

pros: can use linear regression directly.

Cons: $y = \langle w, x \rangle$

- Not scale invariant

[100 x give 100 y]

Solution: Instead of finding model of $y_i = f(x_i)$

Instead we predict $Prob(y_i = 1 \mid x_i) = f(x_i)$

$\hookrightarrow$ conditional probability of predicting lable YES given data point x_i

$\Rightarrow Prob(y_i = 0 \mid x_i) = 1 - f(x_i)$

[assume 2 classes]

||

⇓

Combine into one equation

$$\text{Prob}(y_i|x_i) = f(x_i)^{y_i}(1-f(x_i))^{1-y_i}$$

$y_i=1$ 时

$y_i=0$ 时

Dataset (x_1, y_1)
(x_2, y_2)
$\vdots$
(x_n, y_n)

With independent samples

$$X = [x_1, \dots, x_n] \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$\text{Prob}(y|X) = p(y_1|x_1)\,p(y_2|x_2) \cdots p(y_n|x_n)$$

$$= \prod_{i=1}^{n} \text{prob}(y_i|x_i)$$

$$= \prod_{i=1}^{n} f(x_i)^{y_i}(1-f(x_i))^{1-y_i}$$

"Likelihood" of training dataset

More convenient to take log 这是 Logistic Loss $\boxed{\text{新的 cost func.}}$

$$-\underbrace{\log(\text{prob}(y|X))}_{\text{negative log likelihood}} = -\left[\sum_{i=1}^{n} y_i \log f(x_i) + (1-y_i)\log(1-f(x_i))\right]$$

交叉熵

$$y_i \log f(x_i) + (1-y_i)\log(1-f(x_i))$$

"Cross - entropy" 交叉熵

$$\to L(f) = \sum_{i=1}^{n} l_i(f)$$

☆ 且 $f(x_i) = \underbrace{\frac{1}{1+e^{-\langle w, x_i\rangle}}}$

定义的函数 有特性

Sigmod

因为f(x)是概率，需满足0~1，则用sigmoid，$\frac{1}{1+e^{-z}}$，实际情况由Logist Reg. 训练出的模型将$\langle w,x_i\rangle$ map到z上且符合情况

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

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

$$\to L(w) = \sum_{i=1}^{n} l_i(w)$$

$$l_i(w) = -\left(y_i \log \frac{1}{1+e^{-\langle w,x_i\rangle}} + (1-y_i)\log \frac{e^{-\langle w,x_i\rangle}}{1+e^{-\langle w,x_i\rangle}}\right)$$

Algorithm to minimize $L(w)$

gradient descent

Fact: $\frac{df(z)}{dz} = f(z)(1-f(z))$

GD $\Rightarrow$ $w_{k+1} = w_k + \alpha_k \sum_{i=1}^{n}\left(y_i - \frac{1}{1+e^{-\langle w,x_i\rangle}}\right) x_i$

Multicase classification

k classes $\quad 1, \cdots, k$

$$\text{prob}(y_i = k \mid x_i) = \frac{\exp(\langle w, x_i \rangle)}{Z}$$

Z : normalization or "partition function"

$$Z = \sum_{k=1}^{k} \exp(\langle w_k, x_i \rangle) \qquad k \in 0/1 \text{ 问题}$$

$$L(w) = \sum_{i=1}^{n} l_i(w)$$

$$l_i(w) = \sum_{k=1}^{K} \boxed{\mathbb{1}(y_i = k)} \log \frac{\exp(\langle w_k, x_i \rangle)}{Z}$$

"softmax"

属于这一类为1，其余为0

indicator

logistic Regression

进一步理解

review logistic Regression

$y = 1$ 或 -1 ?：

① $P(y_i=1 \mid x_i) = f(x_i)$

$P(y_i=-1 \mid x_i) = 1-f(x_i)$

② $P(y_i \mid x_i) = f(x_i)^{a_1 y_i + b_1} \cdot (1-f(x_i))^{a_2 y_i + b_2}$

$$\begin{cases} P(y_i=1 \mid x_i) = f(x_i)^{a_1+b_1} \cdot (1-f(x_i))^{a_2+b_2} = f(x_i) \\ P(y_i=-1 \mid x_i) = f(x_i)^{b_1-a_1} \cdot (1-f(x_i))^{b_2-a_2} = (1-f(x_i)) \end{cases}$$

$\Rightarrow \begin{cases} a_1+b_1=1 \\ b_1-a_1=0 \end{cases}$ & $\begin{cases} a_2+b_2=0 \\ b_2-a_2=1 \end{cases}$

$\Rightarrow \begin{cases} b_1=\frac{1}{2} \\ a_1=\frac{1}{2} \end{cases}$ & $\begin{cases} b_2=\frac{1}{2} \\ a_2=-\frac{1}{2} \end{cases}$

③ $P(y_i \mid x_i) = f(x)^{\frac{1+y_i}{2}} (1-f(x_i))^{\frac{1-y_i}{2}}$

$\Downarrow$

$P(y \mid X) = \prod_{i=1}^{n} f(x)^{\frac{1+y_i}{2}} (1-f(x_i))^{\frac{1-y_i}{2}}$

$\Downarrow$ negative log-likelihood

$L(f) = -\sum_{i=1}^{n} \left(\frac{1+y_i}{2}\right) \log f(x_i) + \left(\frac{1-y_i}{2}\right) \log(1-f(x_i))$

④ $f(x) \Rightarrow$ sigmoid : $\frac{1}{1+e^{-z}} = \frac{1}{1+e^{-\langle w, x_i \rangle}}$

$$L(w) = -\sum_{i=1}^{n}\left[\left(\frac{1+y_i}{2}\right)\log\frac{1}{1+e^{-\langle w,x_i\rangle}} + \left(\frac{1-y_i}{2}\right)\log\frac{e^{-\langle w,x_i\rangle}}{1+e^{-\langle w,x_i\rangle}}\right]$$

cross entropy (cost)

⑤ $\sigma(z)$ represent sigmoid

性质
$$\begin{cases} \sigma'(z) = \sigma(z)(1-\sigma(z)) \\ 1-\sigma(z) = \sigma(-z) \end{cases}$$

thus:

$$\frac{1+y}{2}\log(\sigma(wx)) + \frac{1-y}{2}\log(1-\sigma(wx))$$

$$= \frac{1+y}{2}\log\sigma(wx) + \frac{1-y}{2}\log\sigma(-wx)$$

$$\frac{\partial}{\partial w} = \frac{1+y}{2}\frac{1}{\sigma(wx)}\cdot\sigma(wx)[1-\sigma(wx)]\cdot X +$$

$$\frac{1-y}{2}\frac{1}{\sigma(-wx)}\cdot\sigma(-wx)[1-\sigma(-wx)]\cdot(-X)$$

$$= \frac{1+y}{2}[1-\sigma(wx)]X - \frac{1-y}{2}\sigma(wx)X$$

$$= \frac{1+y}{2}\cdot X - \sigma(wx)X\left[\frac{1+y}{2}+\frac{1-y}{2}\right]$$

$$= \frac{1+y}{2}X - \sigma(wx)\cdot X = \left[\frac{1+y}{2} - \sigma(wx)\right]X$$

gradient:

$$\frac{\partial L(w)}{\partial w} = -\sum_{i=1}^{n}\left[\frac{1+y}{2} - f(wx)\right]X$$

$$W_{k+1} = W_k - \alpha \cdot \text{gradient}$$

$$= W_k + \alpha \sum_{i=1}^{n}\left[\frac{1+y}{2} - f(wX)\right]X$$