PART 1-2

Three Kingdoms of Classification

Optimization Approach

logistic surrogate

logistic reg. / Softmax

$\Sigma_c = \Sigma \quad \forall c$

GDA

三国归晋

Ⓐ Softmax

$P(Y=c|\bar{x}) \propto e^{\bar{W}_c^T \bar{x}}$

Ⓑ GDA

$P(Y), \; P(\bar{x}|\bar{Y}=c)$

$= \mathcal{N}(\bar{x}|\bar{\mu}_c, \Sigma_c)$

$\xLeftarrow{\Sigma_c = \Sigma \; \forall c}$

$\cancel{\Longrightarrow}$

inductive bias

Weaker assumptions	Stronger Assumptions
More Data	less Data
More Robust	less robust

Data + Assumptions ⟹ results

FNN, ViT, CNN,

	Discriminative	Generative
	$P(Y \mid \bar{X})$	$P(Y)$, $P(\bar{X} \mid Y)$
Missing Value	No principled approach (heuristic, fill in) impute	EM (latent variable) (expectation maxization)
		intelligence
Feature transformation $\bar{X} \Rightarrow \phi(\bar{X})$	Easy $\searrow$ Superised Deep learning	hard Unsupervised DL:

Discrete Data

$\{\bar{x}_i, y_i\}_{i=1}^{N} \Rightarrow p(y), p(\bar{x}|y)$

$$\bar{x}_i = \begin{bmatrix} x_{i,0} \\ x_{i,1} \\ \vdots \\ x_{i,D} \end{bmatrix} \qquad \bar{x} = \begin{bmatrix} x_0 \\ \vdots \\ x_D \end{bmatrix} \qquad x_i \in \{0,1\}$$

$p(\bar{x}|y) = P(x_0, x_1, \dots x_D | y)$

$p(\bar{x}\|y)$	x_0	x_1	$\dots$	x_D
0.3	0	0	-	0
0.01	0	0	$\dots$	1
$\vdots$	-	-	-	
0.0001	1	1		1

# of para's $= 2^{D+1} - 1$

$$p(x_0, x_1, \ldots, x_D \mid Y=c)$$

$$= p(x_0 \mid Y=c)\, p(x_1 \mid Y=c) \cdots p(x_D \mid Y=c)$$

independence Assumption

$\#$ of para's $= D+1$

$$\pi_c = P(Y=c), \qquad \mu_{jck} = P(x_j = k \mid Y=c)$$

$$\hat{\pi}_c = \frac{n_c}{N} \qquad \hat{\mu}_{jck} = \frac{n_{jck} + \alpha}{n_c + K\alpha}$$

$0 < \alpha$

Laplace smoothing

L05

So far, Training $D = \{\bar{x}_i, y_i\}_{i=1}^{N}$ $\sim P(\bar{x}, y)$, $y_i \sim R^1$

Model: $y = f(\bar{x}) = \bar{w}^T \phi(\bar{x})$

Learning: Determine $\bar{w}$.

This part: Training Set $S = \{\bar{x}_i, y_i\}_{i=1}^{m} \sim$ population D

Hypothesis space: $H = \{h(x) = \bar{w}^T \phi(\bar{x}), \bar{w} \in R^D\}$

Learning: pick one h from H

$h \in H$ $\qquad S = \{\bar{x}_i, y_i\}_{i=1}^{m}$

$$\hat{\varepsilon}(h) = \frac{1}{m} \sum_{i=1}^{m} (y_i - h(x_i))^2 \quad : \text{empirical error}$$

Training: $h_S = \arg\min_{h \in H} \hat{\varepsilon}(h)$

Generalization error : $\varepsilon(h_S) = E_{(\bar{x}, y) \sim D}[(y - h_S(\bar{x}))^2]$

Random because of data collection

Expected Generalization error $\varepsilon = E_S[\varepsilon(h_S)]$

Want to minimize ε because we hope that our algorithm work well for everyone overall

$$\underline{\mathcal{E}} = E_S \, E_{\bar{x},y} \left[(y - h_S(\bar{x}))^2 \right]$$

$$= E_S E_{\bar{x},y} \left[(\underbrace{y - E_S[h_S]}_{a} + \underbrace{E_S[h_S] - h_S}_{b})^2 \right]$$

$$= \underbrace{E_S E_{\bar{x},y} \left[(y - E_S[h_S])^2 \right]}_{\text{bias}^2} + \underbrace{E_S E_{\bar{x},y} \left[(E_S[h_S] - h_S)^2 \right]}_{\text{Variance}}$$

$$\cancel{- 2 E_S E_{\bar{x},y} \left[(y - E_S[h_S]) (E_S[h_S] - h_S) \right]} = 0$$

$$E_S \left[E_S[h_S] - h_S \right]$$

$$= E_S[E_S[h_S]] - E_S[h_S]$$

$$= E_S[h_S] - E_S[h_S]$$

$$= 0$$

$$E_x \, E_x[f(x)]$$

$$= E_x[f(x)]$$

$$\epsilon = E_{(\mathbf{x},y)}[(y - E_S(h_S(\mathbf{x})))^2] + E_S E_{(\mathbf{x})}[(E_S(h_S(\mathbf{x})) - h_S(\mathbf{x}))^2]$$

First term: $E_{(\mathbf{x},y)}[(y - E_S(h_S(\mathbf{x})))^2]$

predicted value for $\bar{x}$: $h_S(\bar{x})$ depends on S

$y \neq E_S[h_S(\bar{x})]$ not depend on S

Reason: Bias in model

Example: Income of HKers

method: check the bank account of 100,000 randomly selected people

Bias: because some people don't receive income through bank

Second term: $E_S E_{(\mathbf{x})}[(E_S(h_S(\mathbf{x})) - h_S(\mathbf{x}))^2]$

$$E_S[h_S(x)] \approx \frac{1}{100} \sum_{i=1}^{100} h_{S_i}(\bar{x}) \qquad S_1, \dots \quad S_{100}$$

$$= h_a(\bar{x})$$

$$\approx E_x \frac{1}{100} \sum_{i=1}^{100} (h_{S_i}(\bar{x}) - h_a(\bar{x}))^2$$

Variance

$h_{S_1}(\bar{x})$

$h_a(\bar{x})$

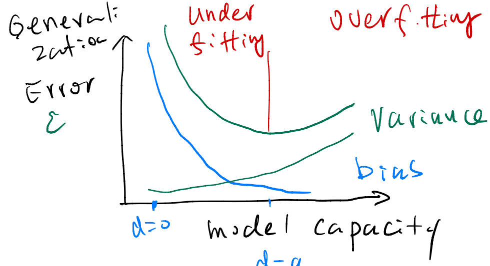
$h_{S_2}(\bar{x})$

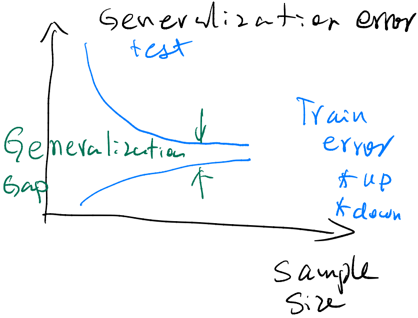
Example:

Method: Ask 5 randomly selected people

High Variance: Team 1: 30K

Team 2: 10K

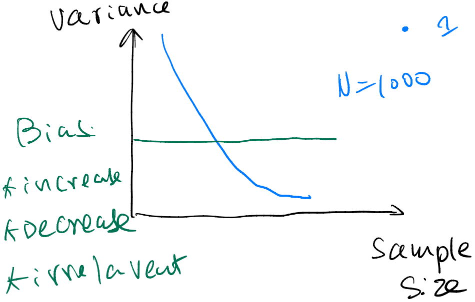




N=2 . 0

. 1

N=1000



Method 1: Ask 5 pple

Method 2: Ask 1000 pple

↓ less Variance

Method: Ask 5 people

Variance: high

Soln: Team 1: result 1

Team 2: result 2

⋮

Team 100: result 100

average

$$D = \{\bar{x}_i, y_i\}_{i=1}^{N}$$

~~sampling~~ with replacement

D_1 ... D_{1000}

$f_1(\bar{x})$ $f_{1000}(\bar{x})$

$$y = f(x) = \begin{cases} \frac{1}{1000}\sum_{i=1}^{1000} f_i(\bar{x}) & \text{regression} \\ \text{Majority vote:} & \\ & \text{Classification} \end{cases}$$