Module 1 - Elements of machine learning

① Regression error function : $E(w) = \frac{1}{2} \sum_{n=1}^{N} [y(x_n, w) - t_n]^2$

$y(x_n, w)$: predicted, t_n : true

Find w that can minimize $E(w)$

② Underfitting (high train and test error)
Overfitting (Low train error but high test error)

Larger dataset can afford a more complex model

③ Regularization : penalty term to training objectives

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} [y(x, w) - t_n]^2 + \frac{1}{2} \|w\|^2$$

④ Model selection : 1. Validation set
2. K-fold - computational expensive
3. leave one out - computational expensive

Probabilistic theory concepts

① max likelihood = max log likelihood = min negative log likelihood

② Bootstrap : sampling with replacement, compute maximum likelihood of model parameters
Higher variance = more uncertain

③ Bayes theorem : $P(w|D) = \frac{P(D|w) P(w)}{P(D)}$

KNN

Module 2 - Linear model for regression

$y(x, w) = w \cdot \phi(x)$ $\quad w = [0\ 0\ 0]$ $\quad \phi(x) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

1×3 $\quad 3 \times 1$ $\quad 1 \times 1$

logistic sigmoid $= \frac{1}{1 - e^{-a}}$ $\quad \tanh = \frac{1 - e^{-2a}}{1 + e^{-2a}}$

$E(w) = \frac{1}{2} \sum_{n=1}^{N} [t_n - w \phi(x_n)]^2$

Gradient descent

Batch

$w_t = w_{t-1} \times (-\eta \times \nabla E(w_{t-1}))$

$\downarrow$

$- \sum_{n=1}^{N} [t_n - w \phi(x_n)] \phi(x_n)$

Stochastic

$w^{\tau} = w^{\tau-1} - \eta^{\tau} \nabla E(w^{\tau-1})$

$= w^{\tau-1} - \eta^{\tau} [t_n - w \phi(x_n)] \phi(x_n)$

L2 regression

$\frac{1}{2} \sum_{n=1}^{N} (t_n - w \phi(x))^2 + \frac{\lambda}{2} \|w\|^2$

L1 regression

$\frac{1}{2} \sum_{n=1}^{N} (t_n - w \phi(x))^2 + \frac{\lambda}{2} |w|$

if λ large : some weights $\to 0$ $\quad \therefore$ fast prediction

more interpretable

Average prediction over all data sets differs from the desired regression function

Bias: consistently learn wrong thing

Variance: learn random things irrespective to real signals

Solution for individual datasets vary around their average.

Complex model	high variance	low bias
Simple	low	high

Stochastic gradient descent

x_0	x_1	x_2	x_3	x_4		w			
1	0.46	0.33	0.56	0.49	0.3	0.2	0.1	0.1	0.2

$$w_1 \text{ new} = 0.2 + 0.01 \times (\text{True} - \text{pred}) \times 0.46$$

Module 3 - Models for classifications

Discriminative models : perceptrons

Probabilistic dis model : logistic regression

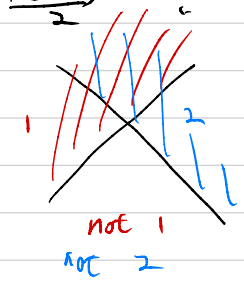
generative model : more parameters, higher complexity → overfitting

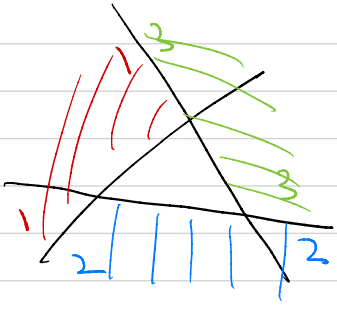
$y(x) = w_0 + w^T x$ Assign to class 1 if $y(x) \geq 0$ else class 2

One versus rest $(K-1)$ classifier

One vs One $\frac{K(K-1)}{2}$

Ambigious region





Perceptron :

misclassified when $w \phi(x_n) \times t_n < 0$ (pred × true)

Training objective: minimize $-\sum_{n \in M} w \phi(x_n) t_n$

↑ negative for misclassified data point

$$w^{\tau+1} = w^{\tau} - \eta \nabla E(w) = w^{\tau} + \eta \phi(x_n) t_n$$ ← 1 / -1

Prob Gen Model

$$p(C_1|x) > p(C_2|x) = \frac{p(x|C_1)\,p(C_1)}{p(x)} > \frac{p(x|C_2)\,p(C_2)}{p(x)}$$

$$p(x|C_1)\,p(C_1) > p(x|C_2)\,p(C_2)$$

$$a = \ln \frac{p(x|C_1)\,p(C_1)}{p(x|C_2)\,p(C_2)} \qquad a > 0 \rightarrow C_1 \quad \text{else } C_2$$

$w\,x \qquad w_0$

Max Likelihood

$$\text{Likelihood} = \prod \left[\rho\, N(x_n|\mu_1, \Sigma)\right]^{t_n} \left[(1-\rho)\, N(x_n|\mu_2, \Sigma)\right]^{(1-t_n)}$$

$$\text{Log Likelihood} = \sum t_n \ln\left[\rho N(x_n|\mu_1 \Sigma)\right] + (1-t_n)\left[(1-\rho)\, N(x_n|\mu_2, \Sigma)\right]$$

$$\frac{dL}{d\rho} = \rho = \frac{N_1}{N_1+N_2}$$

$$\mu_1 = \frac{1}{N_1} \sum_{n=1}^{N} t_n x_n$$

$$\mu_2 \quad \frac{1}{N_2} \sum_{n=1}^{N} (1-t_n) x_n$$

$$\Sigma = \frac{N_1}{N} S_1 + \frac{N_2}{N} S_2 \qquad S_1 = \frac{1}{N_1} \sum_{n \in C_1} (x_n - \mu_1)(x_n - \mu_1)^T$$

Logistic regression

$$P(C_1|x) = \frac{1}{1-e^{-w \cdot x}}$$

$$L(w) = \log \prod_{n=1}^{N} y_n^{t_n} (1-y_n)^{1-t_n}$$

t_n = True class

$y_n = P(C_1|x)$

$$w^{\tau+1} = w^{\tau} - \eta^{\tau} (y_n - t_n) x_n \quad (1 \quad x_1 \quad x_2)$$

Perceptron

Phi			w			T
1	0.5	0.4	0.2	0.3	0.5	1

$$W_1 = W \times \eta \times Phi \times T$$

Bayes Classifier

weighted avg of class cov

step 1: Calculate prior prob $p(C_k)$ for each class

2: calculate μ_k, S_k class cov, Σ share cov

3: use PDR calculate $P(x|C_k)$ for each pt.

4: calculate $P(C_k)\, p(x|C_k)$ ← find max

Module 4 — Latent Variable and EM

$$\gamma(z_{nk}) = p(z_n = k \mid x_n) = \frac{\rho_k \, N(x_n \mid \mu_k, \Sigma_k)}{\sum_j \rho_j \, N(x_n \mid \mu_j, \Sigma_j)}$$

$$\text{Log-likelihood} = \sum_{n=1}^{N} \ln p(x_n) = \sum_{n=1}^{N} \ln \underbrace{\sum_{k=1}^{K} \rho_k \, N(x_n \mid \mu_k, \Sigma_k)}_{p(x_n)}$$

(1) $\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \, x_n$ (N_k: sum of γ)

(2) $\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (x_n - \mu_k)(x_n - \mu_k)^T$

(3) $\rho_k = \frac{N_k}{N}$

Document Clustering

$$p(k \mid d) = \rho_k \prod \mu_{k,w}^{c(w,d)}$$

(μ_{kw}: $c(w,d)$; ρ)

$$\gamma = P(z \mid d) = \frac{P(d \mid z) \, P(z)}{P(d)}$$

$$\rho_k = \frac{N_k}{N} \qquad N_k = \sum \gamma(z_{n,k})$$

$$\mu_{k,w} = \frac{\sum \gamma(z_{n,k}) \, c(w,d)}{\sum_{w' \in A} \sum_{n=1}^{N} \gamma(z_{n,k}) \, c(w', d_n)}$$

Module 5

Neural Network

$$\frac{\partial E}{\partial w_{jk}} = \frac{\partial E}{\partial o} \times \frac{\partial o}{\partial w_{jk}}$$

$$= \frac{\partial (t_k - o_k)^2}{\partial o_k} \times \frac{\partial o_k}{\partial w_{jk}}$$

$$= -2(t_k - o_k) \times \frac{\partial o_k}{\partial w_{jk}}$$

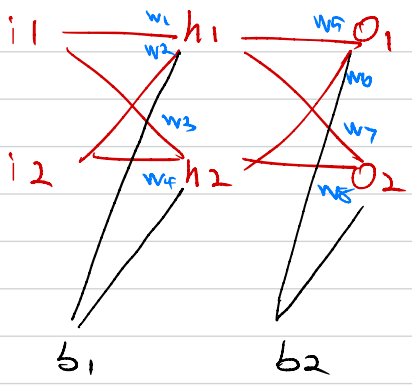
$$= -2(t_k - o_k) \times \frac{\partial \sigma(\sum_j w_{jk} \cdot o_j)}{\partial w_{jk}}$$ derivative of sigmoid

$$= -2(t_k - o_k) \times \left[\sigma(\sum_j w_{jk} \cdot o_j)(1 - \sigma(\sum_j w_{jk} \cdot o_j)) \times \frac{\partial}{\partial w_{jk}}(\sum_j w_{jk} \cdot o_j)\right]$$

$$= -2(t_k - o_k) \times \sigma(\sum_j w_{jk} \cdot o_j)(1 - \sigma(\sum_j w_{jk} \cdot o_j)) \times o_j$$

$d_3 =$ $-$ (True $-$ pred) $*$ derivative of sigmoid (z_3)

$d_2 =$



(target 1) $\frac{\partial E_{total}}{\partial w_5} = \underset{1}{\frac{\partial E_{total}}{\partial out_{o1}}} \times \underset{2}{\frac{\partial out_{o1}}{\partial net_{o1}}} \times \underset{3}{\frac{\partial net_{o1}}{\partial w_5}}$

(target 2)

$$E_{total} = \frac{1}{2}(target_{o1} - out_{o1})^2 + \frac{1}{2}(target_{o2} - out_{o2})^2$$

1 $$\frac{\partial E_{total}}{\partial out_{o1}} = -(target_{o1} - out_{o1})$$

2 $$\frac{\partial out_{o1}}{\partial net_{o1}} = \frac{\partial \sigma(net_{o1})}{\partial (net_{o1})} = \sigma(net_{o1}) \times (1 - \sigma(net_{o1})) = out_{o1} \times (1 - out_{o1})$$

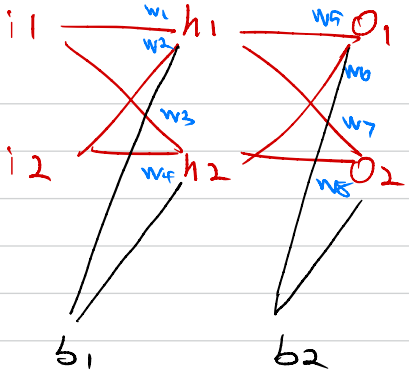
3 $$\frac{\partial net_{o1}}{\partial w_5} = \frac{\partial (w_5 \times out_{h1} + w_6 \times out_{h2} + b_2 \times 1)}{\partial w_5} = out_{h1}$$

$$\frac{\partial E_{total}}{\partial w_5} = \overbrace{-(target_{o1} - out_{o1}) \times (out_{o1} \times (1 - out_{o1}))}^{d_{o1}} \times out_{h1}$$

$$\frac{\partial E_{total}}{\partial w_5} = d_{o1} \times out_{h1}$$

SAME FOR w_6 w_7 w_8

$$w_5 = w_5 - \eta \times (d_{o1} \times out_{h1})$$



$$\frac{\partial E_{total}}{\partial w_1} = \frac{\partial E_{total}}{\partial out_{h_1}} \times \frac{\partial out_{h_1}}{\partial net_{h_1}} \times \frac{\partial net_{h_1}}{\partial w_1}$$

$$\frac{\partial E_{total}}{\partial out_{h_1}} = \frac{\partial E_{o1}}{\partial out_{h_1}} + \frac{\partial E_{o2}}{\partial out_{h_1}}$$

$$\frac{\partial E_{o1}}{\partial out_{h_1}} = \frac{\partial E_{o1}}{\partial net_{o1}} \times \frac{\partial net_{o1}}{\partial out_{h_1}}$$ ← w_5

$$\left(\frac{\partial E_{o1}}{\partial out_{o1}} \times \frac{\partial out_{o1}}{\partial net_{o1}}\right)$$ from previous step

$\frac{\partial \frac{1}{2}(t-o)^2}{\partial o}$ ← $\frac{\partial E_{o1}}{\partial out_{o1}}$

$out_{o1} \times (1 - out_{o1})$ ← $\frac{\partial out_{o1}}{\partial net_{o1}}$

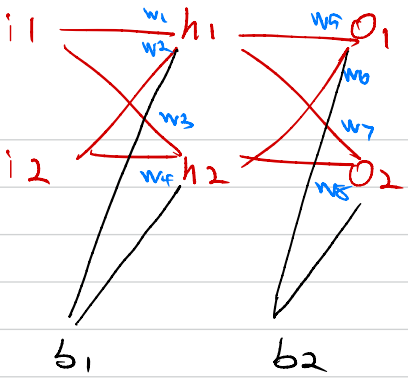
δ_1

$$out_{h_1} = \sigma(net_{h_1})$$

$$\frac{\partial out_{h_1}}{\partial net_{h_1}} = out_{h_1} \times (1 - out_{h_1})$$

$$net_{h_1} = i_1 \times w_1 + i_2 \times w_2 + b_1$$

$$\frac{\partial net_{h_1}}{\partial w_1} = i_1$$



$$\frac{\partial E_{TOTAL}}{\partial w_5} = \frac{\partial E_{TOTAL}}{\partial Out_{o1}} \times \frac{\partial Out_{o1}}{\partial Net_{o1}} \times \frac{\partial Net_{o1}}{\partial w_5}$$

$$\frac{\partial E_{TOTAL}}{\partial Out_{o1}} = \frac{\partial E_{o1}}{\partial Out_{o1}} = \frac{\partial (t_1 - Out_{o1})^2}{\partial Out_{o1}}$$

$$= -2(t_1 - Out_{o1})$$

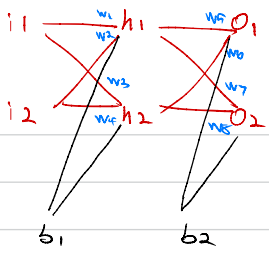
$$\frac{\partial Out_{o1}}{\partial net_{o1}} = Out_{o1} \times (1 - Out_{o1})$$

$$\frac{\partial Net_{o1}}{\partial w_5} = \frac{\partial (b2 + w_5 Out_{h1} + w_6 Out_{h2})}{\partial w_5}$$

$$= Out_{h1}$$

$$\frac{\partial E_{TOTAL}}{\partial w_5} = -2(t_1 - Out_{o1}) \times Out_{o1} \times (1 - Out_{o1}) \times Out_{h1}$$

$$\frac{\partial E_{TOTAL}}{\partial b2} = -2(t_1 - Out_{o1}) \times Out_{o1} \times (1 - Out_{o1}) \times 1$$



$$\frac{\partial E_{TOTAL}}{\partial w_1} = \frac{\partial E_{TOTAL}}{\partial out_{h_1}} \times \frac{\partial out_{h_1}}{\partial net_{h_1}} \times \frac{\partial net_{h_1}}{\partial w_1}$$

$$\frac{\partial E_{TOTAL}}{\partial out_{h_1}} = \frac{\partial E_{O1}}{\partial out_{h_1}} + \frac{\partial E_{O2}}{\partial out_{h_1}}$$

$$\frac{\partial E_{O1}}{\partial out_{h_1}} = \frac{\partial E_{O1}}{\partial out_{O1}} \times \frac{\partial out_{O1}}{\partial net_{O1}} \times \frac{\partial net_{O1}}{\partial out_{h_1}}$$

$$= -2(t_1 - out_{O1}) \times out_{O1} \times (1 - out_{O1}) \times w_5$$

$$\frac{\partial E_{O2}}{\partial out_{h_1}} = \frac{\partial E_{O2}}{\partial out_{O2}} \times \frac{\partial out_{O2}}{\partial net_{O2}} \times \frac{\partial net_{O2}}{\partial out_{h_1}}$$

$$= \underbrace{-2(t_2 - out_{O2}) \times out_{O2} \times (1 - out_{O2})}_{d_3} \times w_7$$

$w_5, w_7 \rightarrow d_2$

$$\frac{\partial out_{h_1}}{\partial net_{h_1}} = out_{h_1} \times (1 - out_{h_1})$$ h.d (≥2)

$$\frac{\partial net_{h_1}}{\partial w_1} = \frac{\partial (w_1 \times i_1 + w_2 \times i_2 + b_1)}{\partial w_1} = i_1$$ → d_1

Linear Regression Walk through

W

0.25 0.82 0.25 0.11 0.41

Observations

	X_0	X_1	X_2	X_3	X_4	T
①	1	2	3	4	5	1
②	2	3	4	5	6	1

Stochastic ← gradient, if batch, then sum all

$$W_0 = W_0 + \eta\left[(T_1 - \hat{y}_1) * Phi[i,j]\right]$$
$$= 0.25 + \eta(1 - \hat{y}_1) * 1$$

Error func $= \frac{1}{2}(T - \hat{y})^2$

Gradient $= -(T - \hat{y}) * Phi[i,j]$

Regularisation – Ridge

Error func $= \frac{1}{2}(T - \hat{y})^2 + \frac{\lambda}{2}\sum_{i=0}^{n} B_i^2$

Perceptron Walk through

Error Func $= -\sum \hat{y} \times t_n$

$= -\sum w \cdot \phi(x) \cdot t_n$

For a single pt : $-w \cdot \phi(x) \cdot t_n$

$\downarrow$

$-(w_0 x_0 + w_1 x_1 + w_2 x_2) \cdot t_n$

Gradient $= -t_n x_1$ for w_1

update : $w_1 = w_1 - \eta \nabla$

$= w_1 + \eta t_n x_1$

W		
w_0	w_1	w_2
0.13	0.49	0.18

	Observation			T
①	1	-0.28	3.46	-1

pred $= 0.13 - 0.28 \times 0.49 + 0.18 \times 3.46$

$= 0.61 \Rightarrow$ classified as $+1$

Wrong classification : need update

Bayes Classifier

1. Calculate $P(C)$
2. Calculate μ_c Σ_c
3. Use result in step 2 to calculate $P(x|C)$
4. Calculate $P(C|x)$ using the result from above

Logistic regression

Log likelihood = $\log \sum y_n^{t_n} (1-y_n)^{1-t_n}$

Gradient = $(y_n - t_n) x_n$

update = $-\eta (y_n - t_n) x_n$

	A	B	C	D	
Doc 1	1	3	7	8	← count
Doc 2	2	5	1	2	

r	1	0.2	0.3	0.4	0.1
r	2	0.3	0.5	0.1	0.1

2 cluster : 1, 2

$\mu_{1,A}$ = $0.2 \times 1 + 0.3 \times 2$

$0.2 \times 1 + 0.2 \times 3 + 0.2 \times 7 + 0.2 \times 8 + 0.3 \times 2 + 0.3 \times 5 + 0.3 \times 1 + 0.3 \times 2$

$\pi_k = \frac{N_k}{N}$ $N_k = \sum_1^n r$

$- W \phi(x) t_n$

$+ \eta \phi(x) t_n$

For 1 to N:

  For 1 to K:

$$\gamma_{nk} = \frac{\rho_k \, N(x_n \mid \mu_k, \Sigma_k)}{\sum_{k=1}^{K} \rho_k \, N(x_n \mid \mu_k, \Sigma_k)}$$

$$N_k = N_k + \gamma_{nk}$$

$$\mu_k = \mu_k + \gamma_{nk} x_n$$

$$\Sigma_k = \Sigma_k + \gamma_{nk} x_n x_n^T$$

For 1 to K:

$$\rho_k = N_k / N$$

$$\mu_k = \mu_k / N_k$$

$$\Sigma_k = \Sigma_k / N_k - \mu \mu^T$$

$$\sum_{m=1}^{M} n_{k,m} \Big/ \sum_{k=1}^{K} \sum_{m=1}^{M} n_{k,m}$$

$$\sum \mu_{k,m} \Big/ \sum_{m=1}^{M} \mu_{k,m}$$

$$\sum_{m=1}^{M} \Sigma_{k,m} \Big/ \sum_{m=1}^{M} n_{k,m}$$

For 1 to N

$$k^* = \arg\min \, dist(x_n, \mu_k)$$

$$\tilde{\mu}_k = \tilde{\mu}_k + x_n$$

$$n_k = n_k + 1$$

$$\sum_{m=1}^{M} \tilde{\mu}_{k,m} \Big/ \sum_{m=1}^{M} n_{km}$$

For 1 to N:

For 1 to K:

$$N_K = N_K + \gamma_{nK}$$
$$\mu_K = \mu_K + \gamma_{nK} x_n$$
$$\Sigma_K = \Sigma_K + \gamma_{nK} x_n x_n^T$$

For 1 to K:

$$\pi_K = N_K / N$$
$$\mu_K = \mu_K / N_K$$
$$\Sigma_K = \Sigma_K / N_K - \mu \mu^T$$

$$\pi_K = \sum_{m=1}^{m} n_{K,m} \Big/ \sum_{k=1}^{K} \sum_{m=1}^{m} n_{k,m}$$
$$\mu_K = \sum_{m=1}^{M} \mu_{K,m} \Big/ \sum_{m=1}^{m} n_{K,m}$$
$$\Sigma_K = \sum_{m=1}^{M} \Sigma_{K,m} \Big/ \sum_{m=1}^{m} n_{K,m} - \mu_K \mu_K^T$$

$$\gamma_{nK} = \frac{\rho_K \, N(x_n \mid \mu_K, \Sigma_K)}{\sum_{k=1}^{K} \rho_K \, N(x_n \mid \mu_K, \Sigma_K)} \qquad N_K = \sum_{n=1}^{N} \gamma_{n,K}$$

$$\mu_K = \frac{1}{N_K} \left(\sum_{n=1}^{N} \gamma_{n,K} x_n \right)$$

$$\rho_K = \frac{N_K}{N}$$

$$\Sigma_K = \frac{1}{N_K} \sum_{n=1}^{N} \gamma_{n,K} (x_n - \mu_K)(x_n - \mu_K)^T$$