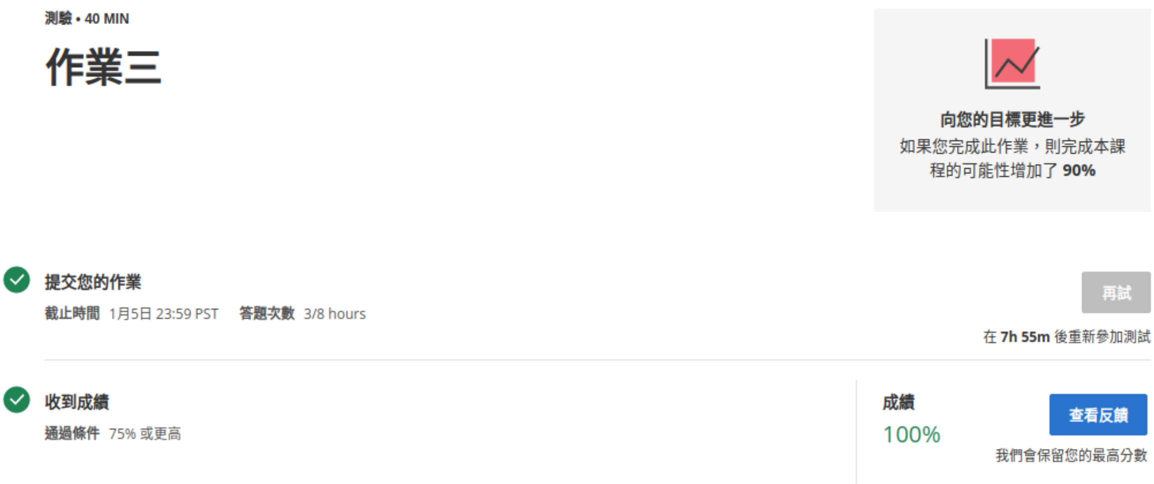
Machine Learning Foundation HW3

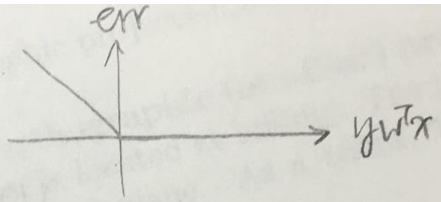
R08942062 林志皓

Problem 1



Problem 2

(2)



Considering 2 case for one x_n

(1) $\text{sign}(w^T x_n) = y_n \Rightarrow y_n w^T x_n > 0$

$err(w) = \max(0, -y_n w^T x_n) = 0 \quad \frac{\partial err}{\partial w} = 0$

$\Rightarrow$ w is not updated

(2) $\text{sign}(w^T x_n) \neq y_n \Rightarrow y_n w^T x_n < 0$, (x_n is classified incorrectly)

$err(w) = -y_n w^T x_n \quad \nabla err = -y_n x_n$

by SGD, $w(t+1) = w(t) - \eta \nabla err$, set $\eta = 1$

$w(t+1) = w(t) + y_n x_n$

by (1) & (2), it's the same as typical PLA.

Problem 3

③ Define varibles as below

$\frac{\partial E}{\partial u}(a.b) = e_u \quad \frac{\partial E}{\partial v}(a.b) = e_v \quad \frac{\partial^2 E}{\partial u^2}(a.b) = e_{uu}$

$\frac{\partial^2 E}{\partial v^2}(a.b) = e_{vv} . \quad \frac{\partial^2 E}{\partial u \partial v}(a.b) = e_{uv} \quad E(a,b) = e$

$\therefore$ Hessian matrix $H = \begin{bmatrix} e_{uu} & e_{uv} \\ e_{uv} & e_{vv} \end{bmatrix}$

$\nabla E = \begin{bmatrix} e_u \\ e_v \end{bmatrix} \qquad$ Let $x = \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix}$.

$$\Rightarrow \hat{E}_2(\nabla u, \nabla v) = E(a,b) + \frac{\partial E}{\partial u}(a,b)\Delta u + \frac{\partial E}{\partial v}(a,b)\Delta v + \frac{\partial^2 E}{2\partial u^2}(a,b)\Delta u^2 + \frac{\partial^2 E}{2\partial v^2}(a,b)\Delta v^2 + \frac{\partial^2 E}{\partial u \partial v}(a,b)\Delta u \Delta v$$

$$= \frac{1}{2}[\Delta u, \Delta v]\begin{bmatrix} e_{uu} & e_{uv} \\ e_{uv} & e_{vv} \end{bmatrix}\begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} + [e_u, e_v]\begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} + e$$

$$= \frac{1}{2}x^T H x + \nabla E^T x + e$$

when optimum. $\frac{\partial \hat{E}_2}{\partial x} = 0 = Hx + \nabla E$.

$$\therefore X = -H^{-1}\nabla E$$

$$= -(\nabla^2 E(u,v))^{-1}\nabla E(u,v) \quad \#$$

Problem 4

(16)

In multiclass logistic regression,

the goal is to maximize probability:

$$\prod_{n=1}^{N} h_y(x_n) = \prod_{n=1}^{N} \frac{\exp(w_{y_n}^T x_n)}{\sum_{k=1}^{K} \exp(w_k^T x_n)}$$

which can be considered as minimizing loss:

$$-\frac{1}{N}\sum_{n=1}^{N} \ln h_y(x_n) = \frac{1}{N}\sum_{n=1}^{N} -\ln\left(\frac{\exp(w_{y_n}^T x_n)}{\sum_{k=1}^{K} \exp(w_k^T x_n)}\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left[\ln\left(\sum_{k=1}^{K} \exp(w_k^T x_n)\right) - w_{y_n}^T x_n\right]$$

Problem 5

(5)

$$\min_{w} \frac{1}{N+K}\left(\sum_{n=1}^{N}(y_n - w^T x_n)^2 + \sum_{k=1}^{K}(y_k - w^T x_k)^2\right)$$

$$\equiv \min_{w}\left[(Xw - y)^2 + (\tilde{X}w - \tilde{y})^2\right]$$

let $E = (Xw - y)^2 + (\tilde{X}w - \tilde{y})^2$

when optimum, $\frac{\partial E}{\partial w} = 2(X^T X w - X^T y) + 2(\tilde{X}^T \tilde{X} w - \tilde{X}^T \tilde{y}) = 0$

$$\Rightarrow (X^T X + \tilde{X}^T \tilde{X}) w = X^T y + \tilde{X}^T \tilde{y}$$

$$\therefore w = (X^T X + \tilde{X}^T \tilde{X})^{-1} (X^T y + \tilde{X}^T \tilde{y}) \quad \#$$

Problem 6

⑥

Let $E_{aug} = \lambda \|w\|^2 + \|Xw - y\|^2$

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

when minimum E_{aug}, $(\lambda I + X^T X) w - X^T y = 0$.

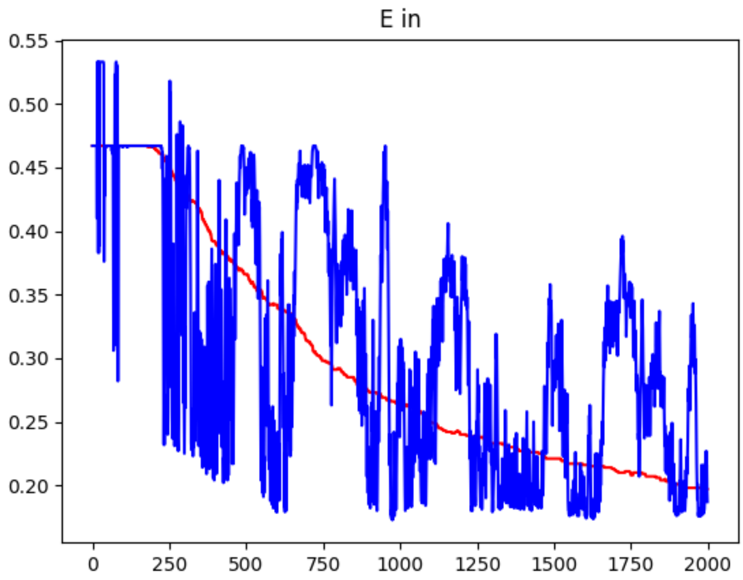
$$w = (X^T X + \lambda I)^{-1} X^T y$$

(compared to ⑤) $= (X^T X + \tilde{X}^T \tilde{X})^{-1} (X^T y + \tilde{X}^T \tilde{y})$

$\therefore$ $\tilde{X}^T \tilde{X} = \lambda I$. $\tilde{X} = \sqrt{\lambda}\, I$

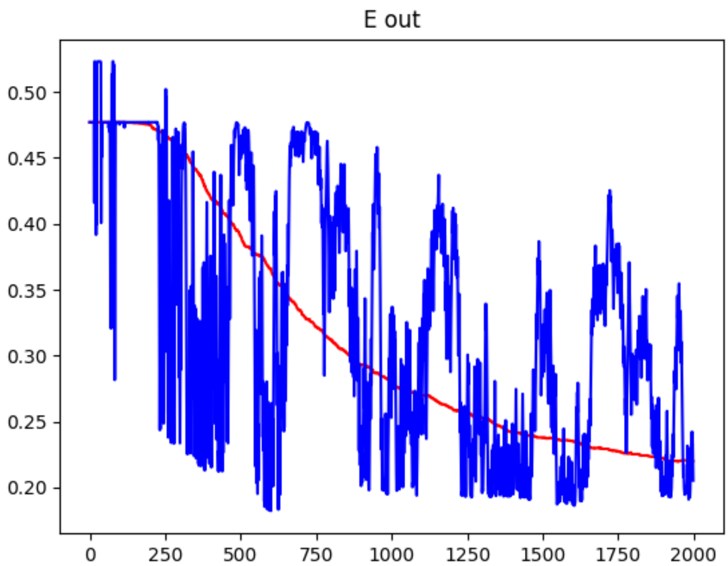
$\tilde{X}^T \tilde{y} = 0$. $\tilde{y} = 0$. #

Problem 7



The red line represents the curve for gradient decent, and the blue one is for stochastic gradient decent. We can see that the red one is much more stable than blue one. On the other hand, blue line might reach lowest E_in in some iteration, but it's very unstable.

Problem 8



The red line represents the curve for gradient decent, and the blue one is for stochastic gradient decent. We can find similar observation with previous question. Fortunately, the model generalize well on testing data with both GD and SGD.

Problem 9

⑨

(a) $X^T X W_{lin} = X^T X (V \Gamma^{-1} U^T y)$

$= X^T (U \Gamma V^T)(V \Gamma^{-1} U^T) y$

$= X^T U U^T y \quad (\because V^T V = I,\ \Gamma \Gamma^{-1} = I)$

$= (V \Gamma U^T) U U^T y = V \Gamma U^T y \quad (\because U^T U = I)$

$= (U \Gamma V^T)^T y = X^T y$

(b) $X^T X w = X^T y$, by SVD decomposition. $X = U \Gamma V^T$.

$(V \Gamma U^T)(U \Gamma V^T) w = (V \Gamma U^T) y$

$V \Gamma^2 V^T w = V \Gamma U^T y$ multiply $\Gamma^{-2} V^T$ on both sides

$V^T w = \Gamma^{-1} U^T y.\ (\because V^T V = I)$ let $\Gamma^{-1} U^T y = a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$

then it's equivalent to solve $V^T w = a$. $V \in \mathbb{R}^{(d+1) \times p}$.

V is composed of p othogonal basis in $(d+1)$-dimension space. that is,

$V = [v_1, v_2, v_3, \cdots v_p]$, $\|v_i\| = 1\ \forall i$, $v_i^T v_j = 0\ \forall i \neq j$

let $w = c_1 v_1 + c_2 v_2 + \cdots c_p v_p + c_k v_k$, where $v_k^T v_i = 0$, $i = 1, 2, \ldots p$, $c_i \in \mathbb{R}\ \forall i$

$V^T w = \begin{bmatrix} v_1^T w \\ v_2^T w \\ \vdots \\ v_p^T w \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_p \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \quad \therefore\ w = a_1 v_1 + a_2 v_2 + \cdots a_p v_p + c_k v_k$

$\|w\|^2 = \|a_1 v_1\|^2 + \|a_2 v_2\|^2 + \cdots \|c_k v_k\|^2$ reach min when $c_k = 0$.

$\therefore \underset{w}{\operatorname{argmin}} \|w\| = a_1 v_1 + a_2 v_2 + \cdots + a_p v_p = V a = V \Gamma^{-1} U^T y = W_{lin}$

which is shortest w that satisfy $X^T X w = X^T y$

$\|W_{lin}\| \leq \|w\| \quad \forall w$.