學號:T10902303 系級：計資一 姓名：梁高浚

請實作以下兩種不同 feature 的模型，回答第 1 ~ 2 題：

1. 抽全部 9 小時內的污染源 feature 當作一次項(加 bias)
2. 抽全部 9 小時內 pm2.5 的一次項當作 feature(加 bias)

備註：

a. NR 請皆設為 0，其他的非數值(特殊字元)可以自己判斷

b. 所有 advanced 的 gradient descent 技術(如: adam, adagrad 等)都是可以用的

c. 第 1 ~ 2 題請都以題目給訂的兩種 model 來回答

d. 同學可以先把 model 訓練好，kaggle 死線之後便可以無限上傳。

1. (1%)記錄誤差值 (RMSE)(根據 kaggle public+private 分數)，討論兩種 feature 的影響

(1)如果是使用全部 9 小時內的污染源 feature 作為一次項，RMSE 是 8.74189(public), 9.85006(private)

(2)如果是使用全部 9 小時內的 pm2.5 的一次項作為 feature, RMSE 是 8.03427(public), 9.11282(private)

這裡可以看出來第二種方法的誤差值比較小，原因是因為第一種方法可能出現了 overfit 的問題，使用過多的 feature 會令到模型過分貼合訓練集，當上傳到 kaggle 的時候碰見未知的 data 就沒法很好的 generalize 預測的結果。

另外，第二種方法其實算是時間序列的模型，其使用了前九小時的 pm2.5 值預測後面時間的 pm2.5 值，它的準確率比較高，可能說明了 pm2.5 這個變量本來就是比較 time-dependent, 就是說它會隨著時間的變動而變化，合理的解釋可能是上下班通勤時間車輛產生的廢氣、其他人為活動令到 pm2.5 上升。

2. (1%)解釋什麼樣的 data preprocessing 可以 improve 你的 training/testing accuracy，e.g., 你怎麼挑掉你覺得不適合的 data points。請提供數據(RMSE)以佐證你的想法。

如果是使用第一個模型，即使用 X1,X2,...,Xn 等 feature 預測 Y, 第一個可以做的 data preprocessing 是檢查 Xi <-> Y 的線性關係。因為在多元線性回歸裡面，其中一個重要的假設是 predictor variables 與 response variable 存在線性關係。

第二個重要的假設是檢查 Xi <-> Xj, i =/= j，因為在多元線性回歸裡面，另外一個重要的假設就是 predictor variables 之間是獨立不相關的，這樣的話(X'X)^(-1)才有解。

Homework 1

Saturday, 23 October 2021 4:41 PM

ML2021F...
HW1 -...

Leung Ka Tam

HW1 - Handwritten Assignment

(Process of finding the answer should be shown; otherwise, no points will be given)

1. Logistic Regression

1-(a)

Suppose we have a logistic regression model with four features that learns the following bias and weights:
$b = 1, w_1 = -1, w_2 = 2, w_3 = -1, w_4 = 5$

Suppose the following feature values for a given example:

x_1	x_2	x_3	x_4
7	0	3	10

Function set:

$$f_{w,b}(x) = P_{w,b}(C_1|x) = \sigma(\sum_i w_i x_i + b)$$

Please calculate the logistic regression prediction for the above particular example. (The answer should be a scalar indicating the posterior probability of class C_1)

1-(b)

Given training data:

x^1	x^2	x^3	...	x^N
$\hat{y}^1 = 1$ (Class 1)	$\hat{y}^2 = 1$ (Class 1)	$\hat{y}^3 = 0$ (Class 2)	...	$\hat{y}^N = 1$ (Class 1)

Assume the data is generated so that the probability of sample x belonging to C_1 is

$$f_{w,b}(x) = P_{w,b}(C_1|x)$$

Given a set of w and b, the probability of generating the data is as follows(assuming the data is generated independently):

$$L(w,b) = f_{w,b}(x^1)f_{w,b}(x^2)(1 - f_{w,b}(x^3)) \cdots f_{w,b}(x^N)$$

Please write down the loss function $L(w,b)$ defined as the negative of the log likelihood (Hint: Cross entropy)

1-(c)

Derive the formula that describes the update rule of parameters in logistic regression. (e.g., $w_i \leftarrow w_i - \ldots$) (Hint: Gradient descent)

2. Closed-Form Linear Regression Solution

In the lecture, we've learnt how to solve linear regression problem via gradient descent. Here you will derive the closed-form solution for such kind of problems.

In the following questions, unless otherwise specified, we denote $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ as a dataset of N input-output pairs, where $\mathbf{x}_i \in \mathbb{R}^k$ denotes the vectorial input and $y_i \in \mathbb{R}$ denotes the corresponding scalar output.

2-(a)

Let's begin with a specific dataset

$$S = \{(x_i, y_i)\}_{i=1}^5 = \{(1, 1.5), (2, 2.4), (3, 3.5), (4, 4.1), (5, 5.3)\}$$

Please find the linear regression model $(w,b) \in \mathbb{R} \times \mathbb{R}$ that minimizes the sum of squares loss

$$L_{ssq}(\mathbf{w}, b) = \frac{1}{2 \times 5} \sum_{i=1}^{5} (y_i - (\mathbf{w}^T \mathbf{x}_i + b))^2$$

2-(b)

Please find the linear regression model $(\mathbf{w}, b) \in \mathbb{R}^k \times \mathbb{R}$ that minimizes the sum of squares loss

$$L_{ssq}(\mathbf{w}, b) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - (\mathbf{w}^T \mathbf{x}_i + b))^2$$

2-(c)

A key motivation for regularization is to avoid overfitting. A common choice is to add a L^2-regularization term into the original loss function

$$L_{reg}(\mathbf{w}, b) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - (\mathbf{w}^T \mathbf{x}_i + b))^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

where $\lambda \geq 0$ and for $\mathbf{w} = [w_1\ w_2\ \dots\ w_k]^T$, one denotes $\|\mathbf{w}\|^2 = w_1^2 + \dots + w_k^2$.

Please find the linear regression model $(\mathbf{w}, b)$ that minimizes the aforementioned regularized sum of squares loss.

3. Noise and regulation

Consider the linear model $f_{\mathbf{w},b} : \mathbb{R}^k \to \mathbb{R}$, where $\mathbf{w} \in \mathbb{R}^k$ and $b \in \mathbb{R}$, defined as

$$f_{\mathbf{w},b}(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$$

Given dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, if the inputs $\mathbf{x}_i \in \mathbb{R}^k$ are contaminated with input noise $\eta_i \in \mathbb{R}^k$, we may consider the expected sum-of-squares loss in the presence of input noise as

$$\tilde{L}_{ssq}(\mathbf{w}, b) = \mathbb{E}\left[\frac{1}{2N}\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_i + \eta_i) - y_i)^2\right]$$

where the expectation is taken over the randomness of input noises $\eta_1, \dots, \eta_N$.

Now assume the input noises $\eta_i = [\eta_{i,1}\ \eta_{i,2}\ \dots \eta_{i,k}]$ are random vectors with zero mean $\mathbb{E}[\eta_{i,j}] = 0$, and the covariance between components is given by

$$\mathbb{E}[\eta_{i,j}\eta_{i',j'}] = \delta_{i,i'}\delta_{j,j'}\sigma^2$$

where $\delta_{i,i'} = \begin{cases} 1 & \text{, if } i = i' \\ 0 & \text{, otherwise.} \end{cases}$ denotes the Kronecker delta.

Please show that

$$\tilde{L}_{ssq}(\mathbf{w}, b) = \frac{1}{2N}\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2 + \frac{\sigma^2}{2}\|\mathbf{w}\|^2$$

That is, minimizing the expected sum-of-squares loss in the presence of input noise is equivalent to minimizing noise-free sum-of-squares loss with the addition of a L^2-regularization term on the weights.

- Hint: $\|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x} = Trace(\mathbf{x}\mathbf{x}^T)$.

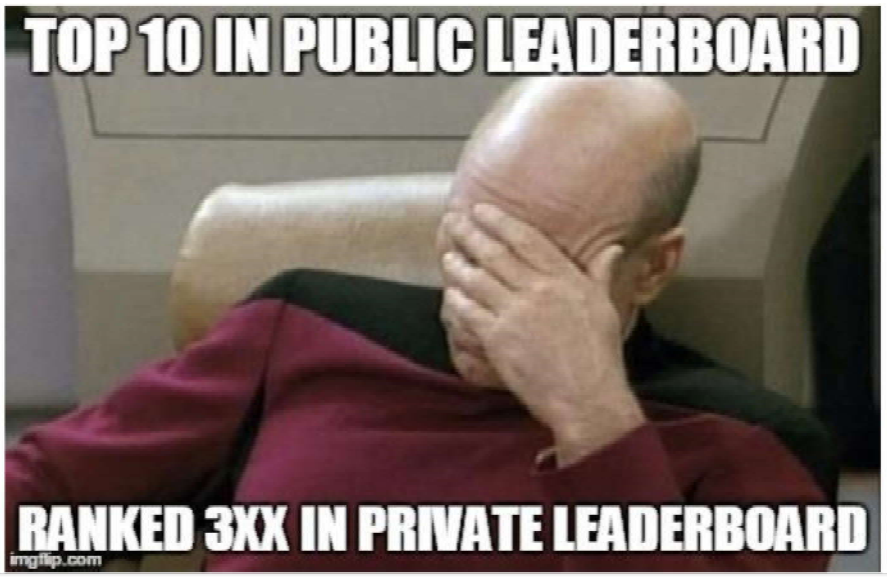
4. Kaggle Hacker

In the lecture, we've learnt the importance of validation. It is said that fine tuning your model based on Kaggle public leaderboard always causes "disaster" on private test dataset.

Let's not talk about whether it'll lead to disastrous results or not. The fact is that most students even don't know how to "overfit" public leaderboard except for submitting many and many times.

In this problem, you'll see how to take advantages of public leaderboard in hw1 kaggle competition. (In theory XD)

Warning



Suppose you have trained $K+1$ models $g_0, g_1, \cdots g_K$, and in particular $g_0(\mathbf{x}) = 0$ is the zero function.

Assume the testing dataset is $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where you only know x_i while y_i is hidden. Nevertheless, you are allowed to observe the sum of squares testing error

$$e_k = \frac{1}{N}\sum_{i=1}^{N}(g_k(\mathbf{x}_i) - y_i)^2, \;\; k = 0, 1, \cdots K$$

Of course, you know $s_k = \frac{1}{N}\sum_{i=1}^{N}(g_k(\mathbf{x}_i))^2$.

4-(a)

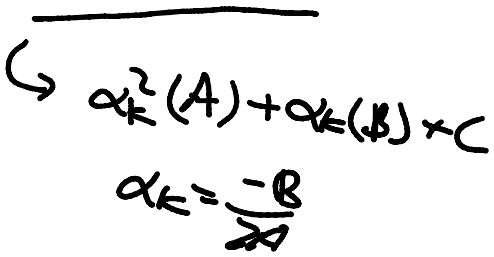
Please express $\sum_{i=1}^{N} g_k(\mathbf{x}_i)y_i$ in terms of $N, e_0, e_1, \cdots, e_K, s_1, \cdots, s_K$. Prove your answer.

- Hint: $e_0 = \frac{1}{N}\sum_{i=1}^{N} y_i^2$

4-(b)

For the given $K+1$ models in the previous problem, explain how to solve

$\min_{\alpha_1,\cdots\alpha_K} L_{test}(\sum_{k=1}^{K}\alpha_k g_k) = \min[\frac{1}{N}\sum_{i=1}^{N}(\sum_{k=1}^{K}\alpha_k g_k(\mathbf{x}_i) - y_i)^2]$, and obtain the optimal weights $\alpha_1, \cdots \alpha_K$.



1).

1. Logistic Regression

1-(a)

Suppose we have a logistic regression model with four features that learns the following bias and weights:
$b = 1, w_1 = -1, w_2 = 2, w_3 = -1, w_4 = 5$

Suppose the following feature values for a given example:

x_1	x_2	x_3	x_4
7	0	3	10

Function set:

$$f_{w,b}(x) = P_{w,b}(C_1|x) = \sigma(\sum_i w_i x_i + b)$$

Please calculate the logistic regression prediction for the above particular example. (The answer should be a scalar indicating the posterior probability of class C_1)

logistic regression prediction = sigmoid(sum of wixi + b)
sum of wixi + b = -1(7) + 2(0) + (-1)(3) + 5(10) = 40 + 1 = 41
sigmoid(41) = 1/(1+e^(-41)) ~= 1
1b).

x^1	x^2	x^3	...	x^N
$\hat{y}^1 = 1$ (Class 1)	$\hat{y}^2 = 1$ (Class 1)	$\hat{y}^3 = 0$ (Class 2)	...	$\hat{y}^N = 1$ (Class 1)

Assume the data is generated so that the probability of sample x belonging to C_1 is

$$f_{w,b}(x) = P_{w,b}(C_1|x)$$

Given a set of w and b, the probability of generating the data is as follows(assuming the data is generated independently):

$$L(w,b) = f_{w,b}(x^1)f_{w,b}(x^2)(1 - f_{w,b}(x^3)) \cdots f_{w,b}(x^N)$$

Please write down the loss function $L(w,b)$ defined as the negative of the log likelihood (Hint: Cross entropy)

$$L(w,b) = \begin{cases} -\log f_{w,b}(x) & \text{if } y=1 \\ -\log(1 - f_{w,b}(x)) & \text{if } y=0 \end{cases}$$

$$L(w,b) = -y \log(f_{w,b}(x)) - (1-y)\log(f_{w,b}(x))$$

1-(c)

Derive the formula that describes the update rule of parameters in logistic regression. (e.g., $w_i \leftarrow w_i - \ldots$) (Hint: Gradient descent)

We want to minimize the loss function.

$$\min_{w,b} L(w,b)$$

$$w_i \leftarrow w_i - \alpha \frac{\partial}{\partial w_i} J(w,b)$$

where α is the learning rate and $J(w,b)$ is the cost function with

$$J(w,b) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)} \times \log(f_{w,b}(x^{(i)})) + (1-y^{(i)}) \times \log(f_{w,b}(x^{(i)}))\right]$$

$$\frac{\partial}{\partial w_i} L(w,b) = \frac{1}{m}\sum_{i=1}^{m}(f_{w,b}(x^{(i)}) - y^{(i)})\, f_{w,b}(x^{(i)})$$

$$w_i \leftarrow w_i - \alpha \frac{1}{m}\sum_{i=1}^{m}(f_{w,b}(x^{(i)}) - y^{(i)})\, f_{w,b}(x^{(i)})$$

2-(a)

Let's begin with a specific dataset

$$S = \{(x_i, y_i)\}_{i=1}^{5} = \{(1, 1.5), (2, 2.4), (3, 3.5), (4, 4.1), (5, 5.3)\}$$

$x_i y_i = 1.5, 4.8, 10.5, 16.4, 26.5$

$\sum x_i y_i = 59.7$

Please find the linear regression model $(w,b) \in \mathbb{R} \times \mathbb{R}$ that minimizes the sum of squares loss

$$L_{sq}(\mathbf{w}, b) = \frac{1}{2 \times 5}\sum_{i=1}^{5}(y_i - (\mathbf{w}^T\mathbf{x}_i + b))^2$$

So, the fitted equation is: $Y = 0.57 + 0.93X$

$\Rightarrow w = 0.93,\ b = 0.57$

2a). $L_{ssq}(w,b) = \frac{1}{2\times 5}\sum_{i=1}^{5}(y_i - (wx_i+b))^2$

Minimize $L_{ssq}(w,b)$, taking derivatives of $L_{ssq}(w,b)$ w.r.t. to w,b respectively.

$$\frac{\partial L(w,b)}{\partial w} = \frac{1}{2\times 5}\sum_{i=1}^{5}\frac{\partial}{\partial w}\left\{(y_i - (wx_i+b))^2\right\}$$

$$= 2(y_i - (wx_i+b))(-x_i)$$

$$= \frac{1}{2\times 5}\sum_{i=1}^{5} 2((wx_i+b) - y_i)(x_i)$$

$$= \frac{1}{5}\sum_{i=1}^{5}\left[x_i(wx_i+b) - x_i y_i\right]$$

Setting $\frac{\partial L(w,b)}{w}$ to be 0,

$$0 = \frac{1}{5}\sum_{i=1}^{5}\left[x_i(wx_i+b) - x_i y_i\right]$$

$$\sum_{i=1}^{5}(wx_i^2 + bx_i) = \sum_{i=1}^{5} x_i y_i \quad -①$$

Similarly, $\frac{\partial L(w,b)}{\partial b} = \frac{1}{5}\sum_{i=1}^{5}\left[wx_i + b - y_i\right]$

Setting $\frac{\partial L(w,b)}{\partial b} = 0$. we have

$$\sum_{i=1}^{5}(wx_i+b) = \sum_{i=1}^{5} y_i \quad -②$$

$\sum_{i=1}^{5}$ $\sum_{i=1}^{5}$

$$\sum_{i=1}^{5} wx_i + 5b = \sum_{i=1}^{5} y_i$$

Grouping ①, ②, we have:

$$\sum_{i=1}^{5} (wx_i^2 + bx_i) = \sum_{i=1}^{5} x_i y_i$$

$$\sum_{i=1}^{5} (wx_i + b) = \sum_{i=1}^{5} y_i$$

$$\begin{cases} w\sum_{i=1}^{5} x_i^2 + b\sum_{i=1}^{5} x_i = \sum_{i=1}^{5} x_i y_i \\ w\sum_{i=1}^{5} x_i + 5b = \sum_{i=1}^{5} y_i \end{cases}$$

$$55w + 15b = 59.7$$

$$15w + 5b = 16.8$$

$$5b = 16.8 - 15w$$

$$55w + 3(16.8 - 15w) = 59.7$$

$$55w + 50.4 - 45w = 59.7$$

$$10w = 9.3$$

$$w = 0.93$$

$$b = \frac{16.8 - 15(0.93)}{5} = 0.57$$

3. Noise and regulation

Consider the linear model $f_{\mathbf{w},b} : \mathbb{R}^k \to \mathbb{R}$, where $\mathbf{w} \in \mathbb{R}^k$ and $b \in \mathbb{R}$, defined as

$$f_{\mathbf{w},b}(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$$

Given dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, if the inputs $\mathbf{x}_i \in \mathbb{R}^k$ are contaminated with input noise $\eta_i \in \mathbb{R}^k$, we may consider the expected sum-of-squares loss in the presence of input noise as

$$\tilde{L}_{ssq}(\mathbf{w}, b) = \mathbb{E}\left[\frac{1}{2N}\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_i + \eta_i) - y_i)^2\right]$$

where the expectation is taken over the randomness of input noises $\eta_1, \ldots, \eta_N$.

Now assume the input noises $\eta_i = [\eta_{i,1}\ \eta_{i,2} \ldots \eta_{i,k}]$ are random vectors with zero mean $\mathbb{E}[\eta_{i,j}] = 0$, and the covariance between components is given by

$$\mathbb{E}[\eta_{i,j}\eta_{i',j'}] = \delta_{i,i'}\delta_{j,j'}\sigma^2$$

where $\delta_{i,i'} = \begin{cases} 1 & \text{, if } i = i' \\ 0 & \text{, otherwise.} \end{cases}$ denotes the Kronecker delta.

Please show that

$$\tilde{L}_{ssq}(\mathbf{w}, b) = \frac{1}{2N}\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2 + \frac{\sigma^2}{2}\|\mathbf{w}\|^2$$

That is, minimizing the expected sum-of-squares loss in the presence of input noise is equivalent to minimizing noise-free sum-of-squares loss with the addition of a L^2-regularization term on the weights.

- Hint: $\|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x} = Trace(\mathbf{x}\mathbf{x}^T)$.

$$\tilde{L}_{ssq}(w,b) = E\left[\frac{1}{2N}\sum_{i=1}^{N}(f_{w,b}(x_i+\eta_i)-y_i)^2\right]$$

$$= \frac{1}{2N}\sum_{i=1}^{N}\left[E(f_{w,b}(x_i+\eta_i)-y_i)^2\right]$$

$$= \frac{1}{2N}\sum_{i=1}^{N}\left[E(w^Tx_i + w^T\eta_i + b - y_i)^2\right]$$

$$= \frac{1}{2N}\sum_{i=1}^{N}\left[E\left((w^Tx_i + b - y_i)^2 + 2(w^Tx_i+b-y_i)w^T\eta_i + (w^T\eta_i)^2\right]\right.$$

$$= \frac{1}{2N}\sum_{i=1}^{N}\left[E\|f_{w,b}(x_i)-y_i\|^2 + 2(f_{w,b}(x_i)-y_i)^2 w^T\eta_i + (w^T\eta_i)^2\right]$$

$$= \frac{1}{2N}\sum_{i=1}^{N}\left[(f_{w,b}(x_i)-y_i)^2 + 0 + E[(w^T\eta_i)^2]\right]$$

$$= \frac{1}{2N}\sum_{i=1}^{N}(f_{w,b}(x_i)-y_i)^2 + \frac{\sigma^2}{2}\|w\|^2$$

Suppose you have trained $K+1$ models $g_0, g_1, \cdots g_K$, and in particular $g_0(\mathbf{x}) = 0$ is the zero function.

Assume the testing dataset is $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where you only know x_i while y_i is hidden. Nevertheless, you are allowed to observe the sum of squares testing error

$$e_k = \frac{1}{N}\sum_{i=1}^{N}(g_k(\mathbf{x}_i) - y_i)^2, \quad k = 0, 1, \cdots K$$

Of course, you know $s_k = \frac{1}{N}\sum_{i=1}^{N}(g_k(\mathbf{x}_i))^2$.

4-(a)

Please express $\sum_{i=1}^{N} g_k(\mathbf{x}_i)y_i$ in terms of $N, e_0, e_1, \cdots, e_K, s_1, \cdots, s_K$. Prove your answer.

- Hint: $e_0 = \frac{1}{N}\sum_{i=1}^{N} y_i^2$

4-(b)

For the given $K+1$ models in the previous problem, explain how to solve

$\min_{\alpha_1, \cdots \alpha_K} L_{test}(\sum_{k=1}^{K}\alpha_k g_k) = \min[\frac{1}{N}\sum_{i=1}^{N}(\sum_{k=1}^{K}\alpha_k g_k(\mathbf{x}_i) - y_i)^2]$, and obtain the optimal weights $\alpha_1, \cdots \alpha_K$.

4a). $e_k = \frac{1}{N}\sum_{i=1}^{N}[g_k(x_i)^2 - 2g_k(x_i)(y_i) + y_i^2]$

$e_k = s_k - \frac{2}{N}\sum_{i=1}^{N} g_k(x_i)y_i + e_0$

$e_k - e_0 - s_k = -\frac{2}{N}\sum_{i=1}^{N} g_k(x_i)y_i$

$\frac{N}{2}(s_k + e_0 - e_k) = \sum_{i=1}^{N} g_k(x_i)y_i$

$\sum_{i=1}^{N} g_k(x_i)y_i = \frac{N}{2}(s_k + e_0 - e_k)$

4b). $\frac{1}{N}\sum_{i=1}^{N}(\sum_{k=1}^{K}\alpha_k g_k(x_i) - y_i)^2$

$= \frac{1}{N}\sum_{i=1}^{N}\left[\left(\sum_{k=1}^{K}\alpha_k g_k(x_i)\right)^2 - 2y_i\sum_{k=1}^{K}\alpha_k g_k(x_i) + y_i^2\right]$

to get a minimum for the above equation,

set $\sum_{k=1}^{K}\alpha_k g_k(x_i) = \frac{-(-2y_i)}{2} = y_i$

since the minimum of quadratic function

is $\frac{-B}{2A}$ for $Ax^2 + Bx + C = 0$.