Designer:陳麒宇(R07943034)

Date:2019/12/28

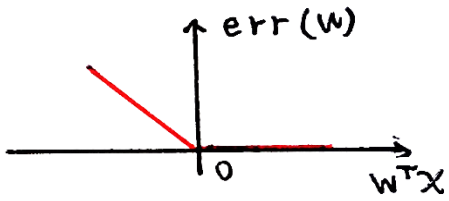
機器學習基石 HW#3

1. Go register for the Coursera version of the first part of the class and solve its homework 3.

2. When using SGD on the following error function and `ignoring' some singular points that are not differentiable, prove or disprove that err(w) = max(0, $-yw^T x$) results in PLA.

For $err(w) = \max(0, -yw^T x)$

1. when $y = +1$



When $w^T x < 0$, $\frac{\partial err}{\partial w} = -yx$

$w_{t+1} \leftarrow w_t - \frac{\partial err}{\partial w} = w_t + yx$

in terms of PLA: 有錯就更正 $(y \neq sign(w^T x))$

$w_{t+1} \leftarrow w_t + [\![y_n \neq sign(w_t^T x_n)]\!] (yx)$

$= w_t + yx$

$\Rightarrow$ SGD on err(w) same as PLA when $y = +1$ and $w^T x < 0$

Designer:陳麒宇(R07943034)

Date:2019/12/28

When $w^Tx > 0$, $\frac{\partial err}{\partial w} = 0$

$$w_{t+1} \leftarrow w_t - \frac{\partial err}{\partial w} = w_t - 0$$

in terms of PLA：沒錯不更正 ($y = sign(w^Tx)$)

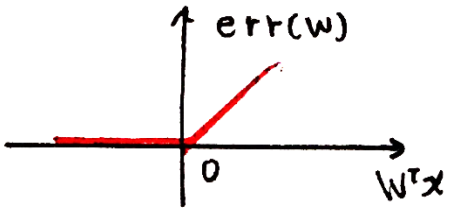
$$w_{t+1} \leftarrow w_t + [\![y_n \neq sign(w_t^T x_n)]\!] (yx)$$

$$= w_t + 0$$

⇒ SGD on err(w) same as PLA

when $y = +1$ and $w^Tx > 0$

2. When $y = -1$.



When $w^Tx < 0$, $\frac{\partial err}{\partial w} = 0$ ($w_{t+1} \leftarrow w_t + 0$)

PLA：沒錯不更正 ($w_{t+1} \leftarrow w_t + 0$)

When $w^Tx > 0$, $\frac{\partial err}{\partial w} = -yx$ ($w_{t+1} \leftarrow w_t + yx$)

PLA：有錯就更正 ($w_{t+1} \leftarrow w_t + yx$)

3. When $y = 0$, the point is not differentiable.

With 1. . 2. and 3., we prove that using SGD on err(w) results in PLA.

Designer:陳麒宇(R07943034)

Date:2019/12/28

3. Write down the derivation steps of Question 9 of Homework 3 on Coursera.

9. Continue from Question 8 and denote the Hessian matrix to be $\nabla^2 E(u,v)$, and assume that the Hessian matrix is positive definite. What is the optimal $(\Delta u, \Delta v)$ to minimize $\hat{E}_2(\Delta u, \Delta v)$? (The direction is called the Newton Direction.)

- ○ $-\nabla^2 E(u,v)\nabla E(u,v)$
- ○ none of the other choices
- ○ $+\left(\nabla^2 E(u,v)\right)^{-1}\nabla E(u,v)$
- ○ $+\nabla^2 E(u,v)\nabla E(u,v)$
- ○ $-\left(\nabla^2 E(u,v)\right)^{-1}\nabla E(u,v)$

Problem: To find the optimal $(\Delta u, \Delta v)$ in

$(u_{t+1}, v_{t+1}) = (u_t, v_t) + (\Delta u, \Delta v)$ to

minimize $\hat{E}_2(\Delta u, \Delta v)$ with $H(u,v) = \nabla^2 E(u,v)$

Ans: $E(x)$ 在 x_k 附近的二階 taylor expansion:

$$E(x) = E(x_k) + (\nabla E(x_k))^T (x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 E(x_k)(x - x_k)$$

Remind: 最好的解使得 $\nabla E(x) = 0$

$$\Rightarrow \nabla E(x) = \nabla E(x_k) + \nabla^2 E(x_k)(x - x_k) = 0$$

$$\Rightarrow x = x_k + \left[-(\nabla^2 E(x_k))^{-1} \nabla E(x_k)\right]$$

替換符號: $(u_{t+1}, v_{t+1}) = (u_t, v_t) + \left[-(\nabla^2 E(u,v))^{-1} \nabla E(u,v)\right]$

Designer:陳麒宇(R07943034)

Date:2019/12/28

4. Write down the derivation steps of Question 16 of Homework 3 on Coursera.

16. For Questions 16-17, you will derive an algorithm for the multinomial (multiclass) logistic regression model. For a K-class classification problem, we will denote the output space $\mathcal{Y} = \{1, 2, \ldots, K\}$. The hypotheses considered by the model are indexed by a list of weight vectors $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K)$, each weight vector of length $d + 1$. Each list represents a hypothesis

$$h_y(\mathbf{x}) = \Big(\exp(\mathbf{w}_y^T \mathbf{x})\Big) / \Big(\sum_{k=1}^{K} \exp(\mathbf{w}_k^T \mathbf{x})\Big)$$

that can be used to approximate the target distribution $P(y|\mathbf{x})$. The model then seeks for the maximum likelihood solution over all such hypotheses.

For general K, derive an $E_{\text{in}}(\mathbf{w}_1, \cdots, \mathbf{w}_K)$ like page 11 of Lecture 10 slides by minimizing the negative log likelihood. What is the resulting E_{in}?

- ○ $\frac{1}{N} \sum_{n=1}^{N} \Big(\sum_{k=1}^{K} \mathbf{w}_k^T \mathbf{x}_n - \mathbf{w}_{y_n}^T \mathbf{x}_n \Big)$
- ○ $\frac{1}{N} \sum_{n=1}^{N} \Big(\sum_{k=1}^{K} \big(\mathbf{w}_k^T \mathbf{x}_n - \mathbf{w}_{y_n}^T \mathbf{x}_n \big) \Big)$
- ○ $\frac{1}{N} \sum_{n=1}^{N} \Big(\ln\Big(\sum_{k=1}^{K} \exp(\mathbf{w}_k^T \mathbf{x}_n) \Big) - \mathbf{w}_{y_n}^T \mathbf{x}_n \Big)$
- ○ none of the other choices

$$\max_{w} \text{likelihood}(w) \propto \prod_{n=1}^{N} h_n(x)$$

$$\Rightarrow \min_{w} \frac{1}{N} \sum_{n=1}^{N} -\ln(h_n(x))$$

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

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

Designer:陳麒宇(R07943034)

Date:2019/12/28

5. Write down the derivation steps of Question 11 of Homework 4 on Coursera.

11. For Questions 11-12, consider linear regression with virtual examples. That is, we add K virtual examples $(\tilde{\mathbf{x}}_1, \tilde{y}_1), (\tilde{\mathbf{x}}_2, \tilde{y}_2), \ldots, (\tilde{\mathbf{x}}_K, \tilde{y}_K)$ to the training data set, and solve

$$\min_{\mathbf{w}} \frac{1}{N+K} \left(\sum_{n=1}^{N} (y_n - \mathbf{w}^T \mathbf{x}_n)^2 + \sum_{k=1}^{K} (\tilde{y}_k - \mathbf{w}^T \tilde{\mathbf{x}}_k)^2 \right).$$

We will show that using some "special" virtual examples, which were claimed to be a possible way to combat overfitting in Lecture 9, is related to regularization, another possible way to combat overfitting discussed in Lecture 10. Let $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1 \tilde{\mathbf{x}}_2 \ldots \tilde{\mathbf{x}}_K]^T$, and $\tilde{\mathbf{y}} = [\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_K]^T$.

What is the optimal $\mathbf{w}$ to the optimization problem above, assuming that all the inversions exist?

- ○ $(\mathbf{X}^T\mathbf{X})^{-1}(\tilde{\mathbf{X}}^T\tilde{\mathbf{y}})$
- ○ $(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y} + \tilde{\mathbf{X}}^T\tilde{\mathbf{y}})$
- ○ $(\mathbf{X}^T\mathbf{X} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}(\tilde{\mathbf{X}}^T\tilde{\mathbf{y}})$
- ○ $(\mathbf{X}^T\mathbf{X} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}(\mathbf{X}^T\mathbf{y} + \tilde{\mathbf{X}}^T\tilde{\mathbf{y}})$
- ○ none of the other choices

$$\min_{w} E_{in}(w) = \frac{1}{N+K} \left(\sum_{n=1}^{N} (y_n - w^T x_n)^2 + \sum_{k=1}^{K} (\tilde{y}_k - w^T \tilde{x}_k)^2 \right)$$

$$\nabla E_{in}(w) = \frac{1}{N+K} \left[(2X^TX w - 2X^Ty) + (2\tilde{X}^T\tilde{X} w - 2\tilde{X}^T\tilde{y}) \right] = 0$$

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

$$\Rightarrow \text{optimal } w = (X^TX + \tilde{X}^T\tilde{X})^{-1}(X^Ty + \tilde{X}^T\tilde{y})$$

6. Write down the derivation steps of Question 12 of Homework 4 on Coursera.

12. For what $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ will the solution of the linear regression problem above equal to

$$\mathbf{w}_{\text{reg}} = \text{argmin}_{\mathbf{w}} \frac{\lambda}{N} \|\mathbf{w}\|^2 + \frac{1}{N} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 ?$$

- ○ $\tilde{\mathbf{X}} = \mathbf{I}, \tilde{\mathbf{y}} = \mathbf{0}$
- ○ $\tilde{\mathbf{X}} = \sqrt{\lambda}\mathbf{I}, \tilde{\mathbf{y}} = \mathbf{0}$
- ○ $\tilde{\mathbf{X}} = \lambda\mathbf{I}, \tilde{\mathbf{y}} = \mathbf{1}$
- ○ $\tilde{\mathbf{X}} = \sqrt{\lambda}\mathbf{X}, \tilde{\mathbf{y}} = \mathbf{y}$
- ○ none of the other choices

根據上題與課程中推導的結果，

我們只需要求 $(X^TX + \tilde{X}^T\tilde{X})^{-1}(X^Ty + \tilde{X}^T\tilde{y})$

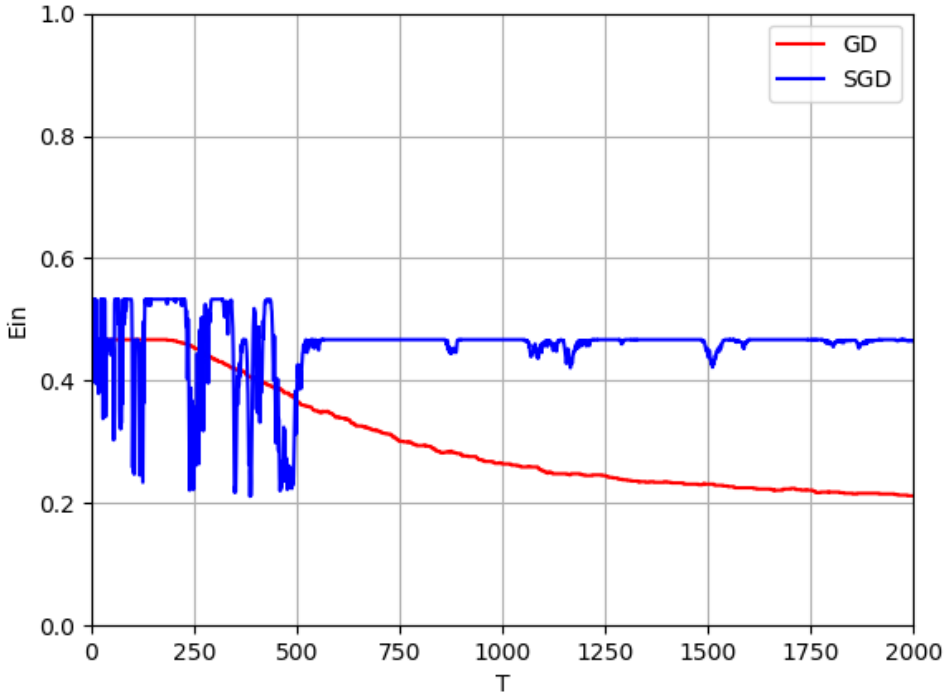
$= (X^TX + \lambda I)^{-1}(X^Ty)$

$\Rightarrow \begin{cases} \tilde{X}^T\tilde{X} = \lambda I \\ \tilde{X}^T\tilde{y} = 0 \end{cases}$

把問題簡單化，只需要求一組解

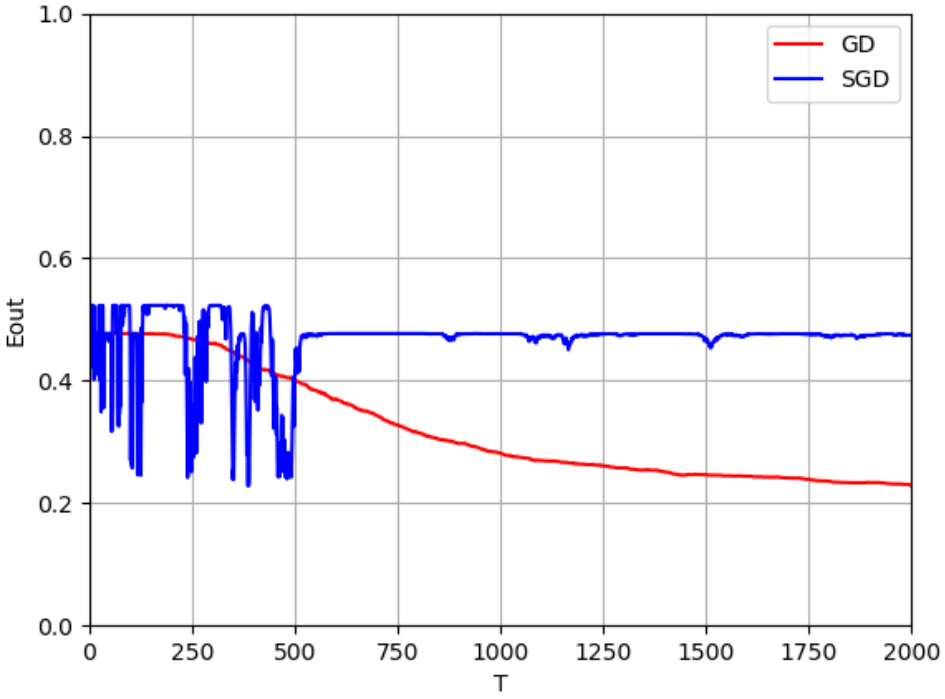
令 $X^T = 0$，上式不成立，因此令 $\begin{cases} \tilde{y} = 0 \\ \tilde{X} = \sqrt{\lambda} I \end{cases}$

7. For Questions 19 and 20 of Homework 3 on Coursera, plot a figure that shows Ein(wt) as a function of t for both the gradient descent version and the stochastic gradient descent version on the same figure. Describe your findings. Please print out the figure for grading.



由上圖結果可以得知，使用 SGD 的方法來隨機選擇梯度更新有點類似 PLA 的概念，隨機選一個點來更新權重，但是不確定更新後的結果會不會比原本好，看圖可以發現 gradient descent 的方法 error 會穩定下降，但需要花費比較多時間，使用 SGD 速度會比較快，但是迭代數不夠多時類似於瞎猜

8. For Questions 19 and 20 of Homework 3 on Coursera, plot a figure that shows Eout(wt) as a function of t for both the gradient descent version and the stochastic gradient descent version on the same figure. Describe your findings. Please print out the figure for grading.



將 E_{out} 跟 E_{in} 比較後會發現兩種梯度更新方法的 E_{out} 都趨近於 Ein，由此可知我們可以先分別嘗試尋找更好的 hypothesis 來降低 Ein，例如增加迭代次數或是提高 learning rate 等方式，並設定訓練停止條件，像是直到平均梯度趨近於 0 或是到達迭代次數上限便停止，藉此讓 E_{out} 有效的收斂

9. Assume that the singular value decomposition (SVD) of X is X = U□V^T , where U 2 RN__ satisfies U^{T}U = I_, V 2 R(d+1)__ satisfies V^{T}V = I_, and □ 2 R___ is a positive diagonal matrix.

   (a) Prove that w_{lin} = V□□1UT y satisfies X^{T}Xw_{lin} = X^{T}y, and hence is a solution. (Note: V□□1UT is actually one way to define the pseudo inverse X$^+$)

(a) 將 SVD of X 與 w_{lin} 代入 $X^TXw = X^Ty$

$X^T(U\Gamma V^T)(V\Gamma^{-1}U^T)y = X^Ty$ 看是否成立

已知 $V^TV = I_\rho$, $\Gamma\Gamma^{-1} = I_\rho$, $U^TU = I_\rho$

由於 $I_\rho^{-1} = I_\rho$, 因此 $UU^T = I_\rho$

$V^TV = I$
$\Gamma\Gamma^{-1} = I$ $\Rightarrow$ $X^T III y = X^Ty$ 等式成立
$UU^T = I$

   (b) Prove that for any solution that satisfies X^{T}Xw = X^{T}y, ||w_{lin}|| <= ||w||. That is, the solution we have constructed is the \shortest" weight vector that minimizes Ein.

   用以下兩張圖來證明這個觀點，對於 0/1 error 來說不論 w^{T}x 多大或多小 Ein 都只會是 0 或是 1，但是對於 squared error 來說距離極值越遠則 error 越大，因此我們會嘗試找到一個絕對值較小的 w，使得 Ein 能夠有效的收斂