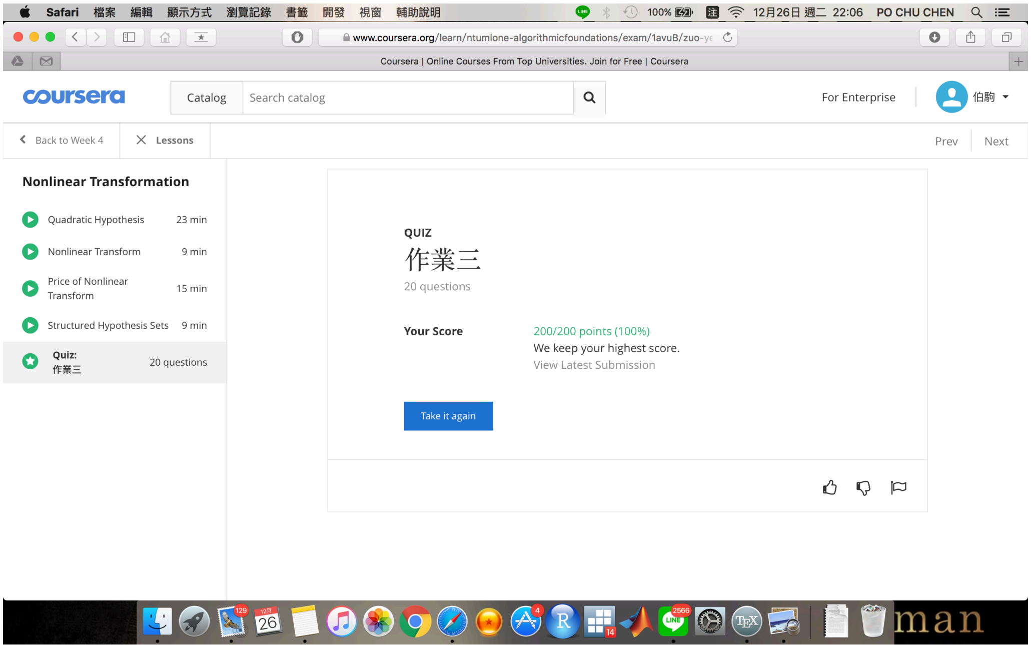
Home Work 3

Machine Learning Foundations

R04323050

經濟碩三　　陳伯駒

1.


Safari 檔案 編輯 顯示方式 瀏覽記錄 書籤 開發 視窗 輔助說明
100%
12月26日 週二 22:06 PO CHU CHEN
www.coursera.org/learn/ntumlone-algorithmicfoundations/exam/1avuB/zuo-ye
Coursera | Online Courses From Top Universities. Join for Free | Coursera
coursera
Catalog
Search catalog
For Enterprise
伯駒
Back to Week 4
Lessons
Prev
Next
Take it again
man

Nonlinear Transformation

Quadratic Hypothesis	23 min
Nonlinear Transform	9 min
Price of Nonlinear Transform	15 min
Structured Hypothesis Sets	9 min
Quiz: 作業三	20 questions

QUIZ

作業三

20 questions

Your Score　200/200 points (100%)

We keep your highest score.

View Latest Submission

2.

<u>Claim:</u> $(I - H)$ is idempotent. i.e $(I - H)^2 = (I - H)$, where $H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

$$
\begin{aligned}
H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T, \quad H \cdot H &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\
&= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\
&= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}I\mathbf{X}^T \\
&= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = H \\
\text{Hence,} \quad (I-H)^2 &= (I-H)\cdot(I-H) \\
&= I(I-H) - H(I-H) = I - H - H + H \cdot H \\
&= I - 2H + H = I - H
\end{aligned}
$$

3.

$err(\mathbf{w}) = \max(0, -y\mathbf{w}^T\mathbf{x})$, W.L.O.G. let $y = 1$

① $\mathbf{w}^T\mathbf{x} < 0$, $err(\mathbf{w}) = -y\mathbf{w}^T\mathbf{x}, \nabla_{\mathbf{w}} err(\mathbf{w}) = -y\mathbf{x}$, $y \neq sgn(\mathbf{w}^T\mathbf{x})$

② $\mathbf{w}^T\mathbf{x} > 0$, $err(\mathbf{w}) = 0, \nabla_{\mathbf{w}} err(\mathbf{w}) = 0$, $y = sgn(\mathbf{w}^T\mathbf{x})$

Plugging the formula of SGD： $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \eta \cdot (-\nabla_{\mathbf{w}} err(\mathbf{w}))$

Let $\eta = 1$, then $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \nabla_{\mathbf{w}} err(\mathbf{w}) = \mathbf{w}_t + [\![y \neq sgn(\mathbf{w}^T\mathbf{x})]\!] \cdot (y\mathbf{x})$

It's obvious that the above result is a PLA iteration in P.11 of slides 11.

4.

Given $\widehat{E_2}(\Delta u, \Delta v)$ is the second-order Taylor's expansion of $E(u, v)$.

$\widehat{E_2}(\Delta u, \Delta v) = E(u,v) + \begin{bmatrix} \frac{\partial E}{\partial u} & \frac{\partial E}{\partial v} \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \Delta u & \Delta v \end{bmatrix} \begin{bmatrix} \frac{\partial^2 E}{\partial u^2} & \frac{\partial^2 E}{\partial v \partial u} \\ \frac{\partial^2 E}{\partial u \partial v} & \frac{\partial^2 E}{\partial v^2} \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix}$, where we let $\mathbf{z} = \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix}$, $\nabla E(u,v) = \begin{bmatrix} \frac{\partial E}{\partial u} & \frac{\partial E}{\partial v} \end{bmatrix}$, $\nabla^2 E(u,v) = \begin{bmatrix} \frac{\partial^2 E}{\partial u^2} & \frac{\partial^2 E}{\partial v \partial u} \\ \frac{\partial^2 E}{\partial u \partial v} & \frac{\partial^2 E}{\partial v^2} \end{bmatrix}$.

Hence, $\widehat{E_2}(\nabla u, \nabla v) = E(u,v) + \nabla E(u,v)\mathbf{z} + \frac{1}{2}\mathbf{z}^T \nabla^2 E(u,v)\, \mathbf{z}$

F.O.C. $\frac{\partial \widehat{E_2}}{\partial \mathbf{z}} = \nabla E(u,v) + \nabla^2 E(u,v)\, \mathbf{z} = 0 \Rightarrow \mathbf{z}^* = \frac{-\nabla E(u,v)}{\nabla^2 E(u,v)} = -(\nabla^2 E(u,v))^{-1} \nabla E(u,v)$

S.O.C. $\frac{\partial^2 \widehat{E_2}}{\partial \mathbf{z}^2} = \nabla^2 E(u,v)$, which is a positive definite matrix given by question. Thus, we know $\mathbf{z}^*$ can achieve a minimum solution.

5.

$$\max_{\mathbf{w}}\ likelihood(\mathbf{w}) \propto \prod_{n=1}^{N} h_y(x_n)$$

$$\Leftrightarrow \max_{\mathbf{w}}\ \ln \prod_{n=1}^{N} \frac{exp(\mathbf{w}_{y_n}^T x_n)}{\sum_{i=1}^{K} exp(\mathbf{w}_i^T x_n)} = \max_{\mathbf{w}} \sum_{n=1}^{N} \ln \frac{exp(\mathbf{w}_{y_n}^T x_n)}{\sum_{i=1}^{K} exp(\mathbf{w}_i^T x_n)}$$

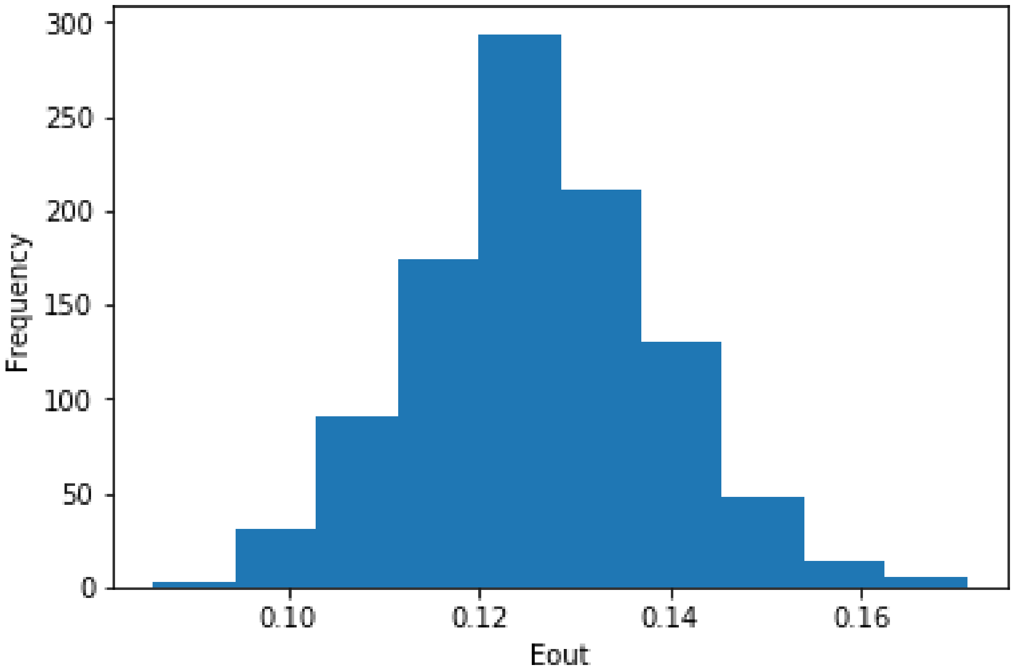
$$\Leftrightarrow \min_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^{N} -\ln \frac{exp(\mathbf{w}_{y_n}^T x_n)}{\sum_{i=1}^{K} exp(\mathbf{w}_i^T x_n)}.$$

Thus, $E_{in} = \dfrac{1}{N} \displaystyle\sum_{n=1}^{N} \left[\ln(\sum_{k=1}^{K} exp(\mathbf{w}_k^T x_n)) - \mathbf{w}_{y_n}^T x_n \right]$

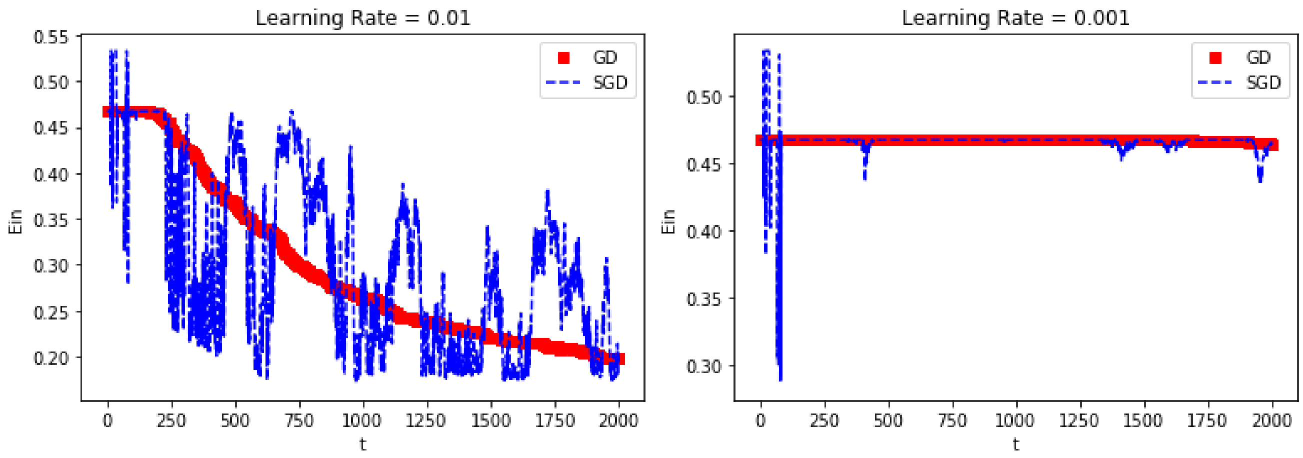
6.

$$\frac{\partial Ein}{\partial \mathbf{w}_i} = \frac{1}{N} \sum_{n=1}^{N} \left\{ \frac{exp(\mathbf{w}_i^T x_n)}{\sum_{i=1}^{K} exp(\mathbf{w}_i^T x_n)} x_n - [\![y_n = i]\!] x_n \right\} = \frac{1}{N} \sum_{n=1}^{N} ((h_i(x_n) - [\![y_n = i]\!]) x_n)$$

7.

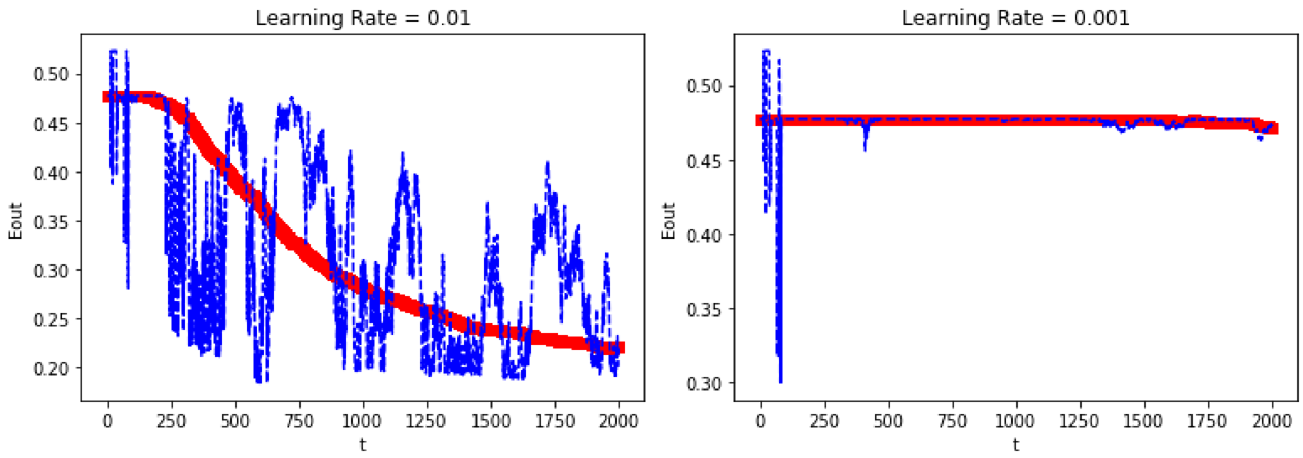


8.



Comparing the GD with SGD version, we can observe that the descent of error rate for GD version is more stable than SGD one's, though after 2000 iterations, their error rate could be almost the same. Moreover, it seems the fixed learning rate = 0.01 is more suitable than 0.001 since it decrease the error rate to a lower level after iterations, implying that $\eta = 0.001$ may be too low steps for learning.

9.



We can find that the trend in figure Ein and Eout are almost same, which highly corresponds to what we expect: $E_{in} \approx E_{out}$. Combining with the result in previous question, we find that this algorithm achieves the learning goal.

10.

(a). We can consider the problem as solving: $\mathbf{X} \cdot \mathbf{w}_{LIN} \approx \mathbf{y}$

We want to find a vector $\mathbf{w}_{LIN}$ such that $\|\mathbf{X} \cdot \mathbf{w}_{LIN} - \mathbf{y}\| \leq \|\mathbf{X} \cdot \mathbf{u} - \mathbf{y}\|, \forall \mathbf{u} \in \mathbb{R}^N$

In P.15 of slides 9, we know $\|\mathbf{X} \cdot \mathbf{w}_{LIN} - \mathbf{y}\| \Leftrightarrow \mathbf{X} \cdot \mathbf{w}_{LIN} = \mathbf{H} \cdot \mathbf{y}$, where $\mathbf{H}$ is a orthogonal projection matrix for $Col(\mathbf{X})$.

Claim: $\mathbf{X}$ is a $N \times (d+1)$ matrix with $rank(\mathbf{X}) = \rho$, and the SVD of $\mathbf{X}$ is $\mathbf{X} = U\Gamma V^T$ given in question setting, then $\mathbf{H} = UU^T$.

Let $UU^T = P$, we can observe that $P^2 = P^T = P$. By theorem[1], we know P is a orthogonal projection matrix of $Col(P)$. We want to show $Col(P) = Col(\mathbf{X})$, then $P = \mathbf{H} = UU^T$.

First, we modify matrix Γ by defining :

$$\Gamma^\dagger = \begin{bmatrix} \frac{1}{\gamma_1} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\gamma_2} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{\gamma_3} & \cdots & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \cdots & \frac{1}{\gamma_\rho} \end{bmatrix} \text{, where we know } \Gamma = \begin{bmatrix} \gamma_1 & 0 & 0 & \cdots & 0 \\ 0 & \gamma_2 & 0 & \cdots & 0 \\ 0 & 0 & \gamma_3 & \cdots & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \cdots & \gamma_\rho \end{bmatrix}, \gamma_1, \gamma_2, ..., \gamma_\rho > 0$$

It's actually easy to observe that $\Gamma^\dagger = \Gamma^{-1}$, then

$$\begin{aligned} \mathbf{X}(V\Gamma^{-1}U^T) &= U\Gamma V^T V\Gamma^{-1}U^T \\ &= U\Gamma\Gamma^{-1}U^T \\ &= UU^T = P \end{aligned}$$

Therefore, for any vector $\mathbf{v} \in \mathbb{R}^N$, we have $P \cdot \mathbf{v} = \mathbf{X} \cdot \mathbf{z}$, where $\mathbf{z} = (V\Gamma^{-1}U^T)\mathbf{v}$, meaning $P \cdot \mathbf{v}$ is in $Col(\mathbf{X})$, i.e $Col(P)$ is a subspace of $Col(\mathbf{X})$. Also, $rank(P) = $ [2]ρ, i.e $dim(Col()P) = \rho$, then by theorem[3], $Col(P) = Col(\mathbf{X})$, hence $P = \mathbf{H} = UU^T$.

[1] Suppose P is a $n \times n$ matrix such that $P^2 = P^T = P$, then P is a orthogonal projection matrix for $Col(P)$. The detailed proof can be checked in Lawarence and Stephen, *Elementary Linear Algebra: A Matrix Approach.*

[2] By the property of Gramian Matrix, for any real matrix $A_{m\times n}$, $rank(A) = rank(AA^T) = rank(A^TA)$. Thus, in this question, we have $rank(U) = rank(U^TU) = rank(UU^T) = rank(I_\rho) = rank(P) = \rho$

[3] If V and W are both the subspace of $\mathbb{R}^n$, and $dim(V) = dim(W)$, then $V = W$. The detailed proof can be checked in Lawarence and Stephen , *Elementary Linear Algebra: A Matrix Approach*

For the vector $\mathbf{w}_{LIN}$ such that $\|\mathbf{X}\cdot\mathbf{u}-\mathbf{y}\|$ is minimum, we let $\mathbf{w}_{LIN} = V\Gamma^{-1}U^T\mathbf{y}$, then $\mathbf{X}\cdot\mathbf{w}_{LIN} = \mathbf{X}(V\Gamma^{-1}U^T)\mathbf{y} = \mathbf{H}\cdot\mathbf{y}$, i.e $\mathbf{w}_{LIN} = V\Gamma^{-1}U^T\mathbf{y}$ is a solution.

(b). Suppose there is another vector $\mathbf{s}\in\mathbb{R}^{d+1}$ can also minimize $\|\mathbf{X}\cdot\mathbf{u}-\mathbf{y}\|$ and $\mathbf{s}\neq\mathbf{w}_{LIN}$. Let $\mathbf{t}=\mathbf{s}-\mathbf{w}_{LIN}(\neq\mathbf{0})$, then

$$\mathbf{X}\cdot\mathbf{t} = \mathbf{X}(\mathbf{s}-\mathbf{w}) = \mathbf{H}\cdot\mathbf{y}-\mathbf{H}\cdot\mathbf{y} = \mathbf{0}$$

Plugging SVD for $\mathbf{X}$, besides U and Γ are invertible, we have

$$U\Gamma V^T\mathbf{t} = \mathbf{0}$$
$$\Rightarrow \Gamma V^T\mathbf{t} = \mathbf{0}$$
$$\Rightarrow V^T\mathbf{t} = \mathbf{0}$$

Also, we know V^T is a orthogonal matrix, it will reserve the inner-product.
$\therefore\ \mathbf{t}\cdot\mathbf{w} = (V^T\mathbf{t})\cdot(V^T\mathbf{w}) = 0$
Now by $\mathbf{s}=\mathbf{t}+\mathbf{w_{LIN}}$, $\|\mathbf{s}\|^{\mathbf{2}} = \|\mathbf{t}+\mathbf{w_{LIN}}\|^{\mathbf{2}} = \|\mathbf{t}\|^{\mathbf{2}}+\|\mathbf{w_{LIN}}\|^{\mathbf{2}} > \|\mathbf{w_{LIN}}\|^{\mathbf{2}}$, i.e $\mathbf{w}_{LIN}$ has the least norm(shortest).