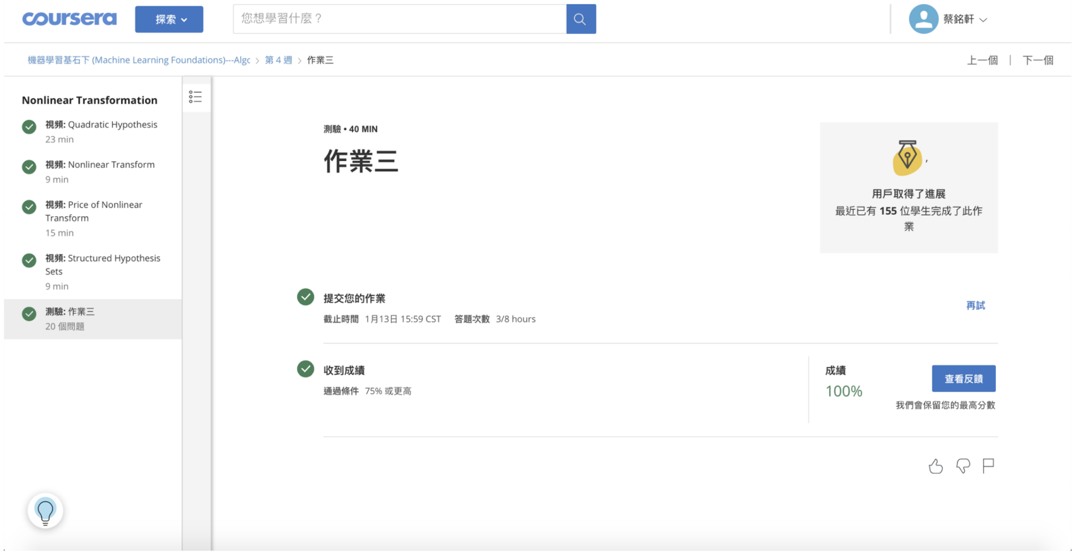
2019 ML foundation HW 3

B07902064 資工二 蔡銘軒

June 29, 2020

Problem 1



機器學習基石下 (Machine Learning Foundations)---Algc > 第 4 週 > 作業三

Nonlinear Transformation

- 視頻: Quadratic Hypothesis
  23 min
- 視頻: Nonlinear Transform
  9 min
- 視頻: Price of Nonlinear Transform
  15 min
- 視頻: Structured Hypothesis Sets
  9 min
- 測驗: 作業三
  20 個問題

測驗 • 40 MIN

作業三

用戶取得了進展
最近已有 155 位學生完成了此作業

提交您的作業
截止時間 1月13日 15:59 CST 答題次數 3/8 hours

收到成績
通過條件 75% 或更高

成績
100%
我們會保留您的最高分數

Problem 2

We consider the case when $y = 1$.

- $\mathbf{w}^T\mathbf{x} \leq 0$
  - **SGD:** The gradient of $err(\mathbf{w}) = -y\mathbf{w}^T\mathbf{x}$ is $-\mathbf{x}$. We update by $\mathbf{w} \leftarrow \mathbf{w} + \mathbf{x} = \mathbf{w} + y \cdot \mathbf{x}$.
  - **PLA:** Assuming $\text{sign}(0) = -1$, then $\mathbf{w}^T\mathbf{x} \leq 0$ indicates that we make a mistake on the point $(\mathbf{x}, y)$, and we update by $\mathbf{w} \leftarrow \mathbf{w} + y \cdot \mathbf{w}$.
- $\mathbf{w}^T\mathbf{x} > 0$
  - **SGD:** The gradient of $err(\mathbf{w}) = 0$ is 0. We update by $\mathbf{w} \leftarrow \mathbf{w} + 0 \cdot \mathbf{x} = \mathbf{w}$.
  - **PLA:** $\mathbf{w}^T\mathbf{x} > 0$ shows that we are correct on the point $(\mathbf{x}, y)$, so we do not update $\mathbf{w}$.

The case when $y = -1$ is similar, and we see that when using SGD with $err(\mathbf{w}) = \max(0, -y\mathbf{w}^T\mathbf{x})$, the result is identical to that of PLA with learning rate 1.

Problem 3

By the second-order Taylor's expansion, we have:

$$\hat{E_2}(\Delta u, \Delta v) = E(u,v) + \frac{1}{2}E_{uu}(u,v)(\Delta u)^2 + \frac{1}{2}E_{vv}(u,v)(\Delta v)^2 + E_{uv}(u,v)\Delta u\Delta v + E_u(u,v)\Delta u + E_v(u,v)\Delta v$$

To find the minimum, we take the partial derivatives with respect to Δu and Δv, and solve the equation which set both of them to zero.

$$\begin{cases} \frac{\partial \hat{E_2}(\Delta u,\Delta v)}{\partial \Delta u} & = E_{uu}(u,v)\Delta u + E_{uv}(u,v)\Delta v + E_u(u,v) = 0 \\ \frac{\partial \hat{E_2}(\Delta u,\Delta v)}{\partial \Delta v} & = E_{vv}(u,v)\Delta v + E_{uv}(u,v)\Delta u + E_v(u,v) = 0 \end{cases}$$

$$\begin{aligned} &\Longrightarrow \begin{bmatrix} E_{uu}(u,v) & E_{uv}(u,v) \\ E_{vu}(u,v) & E_{vv}(u,v) \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = -\begin{bmatrix} E_u(u,v) \\ E_v(u,v) \end{bmatrix} \\ &\Longrightarrow \nabla^2 E(u,v) \cdot \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = -\nabla E(u,v) \\ &\Longrightarrow \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = -(\nabla^2 E(u,v))^{-1} \nabla E(u,v) \end{aligned}$$

In the last step, we use the property of a positive definite matrix — it is invertible.

Problem 4

Let $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K)$. Following the notation and concept in lecture slides, we have

$$\begin{aligned} &\max_{\mathbf{w}} \prod_{n=1}^{N} P(\mathbf{x}_n) \frac{e^{\mathbf{w}_{y_n}^T \mathbf{x}_n}}{\sum_{k=1}^{K} e^{\mathbf{w}_k^T \mathbf{x}_n}} \\ \Longrightarrow &\max_{\mathbf{w}} \prod_{n=1}^{N} \frac{e^{\mathbf{w}_{y_n}^T \mathbf{x}_n}}{\sum_{k=1}^{K} e^{\mathbf{w}_k^T \mathbf{x}_n}} \quad (P(\mathbf{x}_n) \text{ remains the same across different } \mathbf{w}\text{, so we leave it out)} \\ \Longrightarrow &\max_{\mathbf{w}} \sum_{n=1}^{N} ln(\frac{e^{\mathbf{w}_{y_n}^T \mathbf{x}_n}}{\sum_{k=1}^{K} e^{\mathbf{w}_k^T \mathbf{x}_n}}) \quad \text{(take } ln() \text{ so that the product becomes the sum)} \\ \Longrightarrow &\min_{\mathbf{w}} \sum_{n=1}^{N} -ln(\frac{e^{\mathbf{w}_{y_n}^T \mathbf{x}_n}}{\sum_{k=1}^{K} e^{\mathbf{w}_k^T \mathbf{x}_n}}) \quad \text{(reverse the sign and find the minimum instead of the maximum)} \\ \Longrightarrow &\min_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^{N} (ln(\sum_{k=1}^{K} e^{\mathbf{w}_k^T \mathbf{x}_n}) - \mathbf{w}_{y_n}^T \mathbf{x}_n) \quad \text{(add a constant } \frac{1}{N} \text{ and express } ln() \text{ in a different form)} \end{aligned}$$

Problem 5

$$\begin{aligned} &\frac{1}{N+K}\left(\sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x}_n)^2 + \sum_{k=1}^{K}(\tilde{y_k} - \mathbf{w}^T\tilde{\mathbf{x}}_k)\right) \\ &= \frac{1}{N+K}\left(\|\mathbf{X}\mathbf{w} - y\|^2 + \|\tilde{\mathbf{X}}\mathbf{w} - \tilde{y}\|^2\right) \end{aligned}$$

To find the minimum, we take its gradient and solve the equation that set it to zero. As the constant $\frac{1}{N+K}$ does not affect the gradient when it is set to zero, we leave it out hereafter.

$$\begin{aligned} &\left(\|\mathbf{X}\mathbf{w} - y\|^2 + \|\tilde{\mathbf{X}}\mathbf{w} - \tilde{y}\|^2\right) \\ &= \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{w}^T\mathbf{X}^Ty + y^Ty + \mathbf{w}^T\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\mathbf{w} - 2\mathbf{w}^T\tilde{\mathbf{X}}^T\tilde{y} + \tilde{y}^T\tilde{y} \end{aligned}$$

Taking the gradient and setting it to zero, we have

$$\begin{aligned} 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T y + 2\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\mathbf{w} - 2\tilde{\mathbf{X}}^T\tilde{y} = 0 \\ \implies \left(\mathbf{X}^T\mathbf{X} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)\mathbf{w} = \mathbf{X}^T y + \tilde{\mathbf{X}}^T\tilde{y} \\ \implies \mathbf{w} = \left(\mathbf{X}^T\mathbf{X} + \tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-1}\left(\mathbf{X}^T y + \tilde{\mathbf{X}}^T\tilde{y}\right) \end{aligned}$$

Problem 6

$$\begin{aligned} &\frac{\lambda}{N}\|\mathbf{w}\|^2 + \frac{1}{N}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 \\ =&\frac{\lambda}{N}\|\mathbf{w}\|^2 + \frac{1}{N}(\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{w}^T\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}) \end{aligned}$$

To find the $\mathbf{w}$ that gives the minimum, we take its gradient and solve the equation that set it to zero. We have

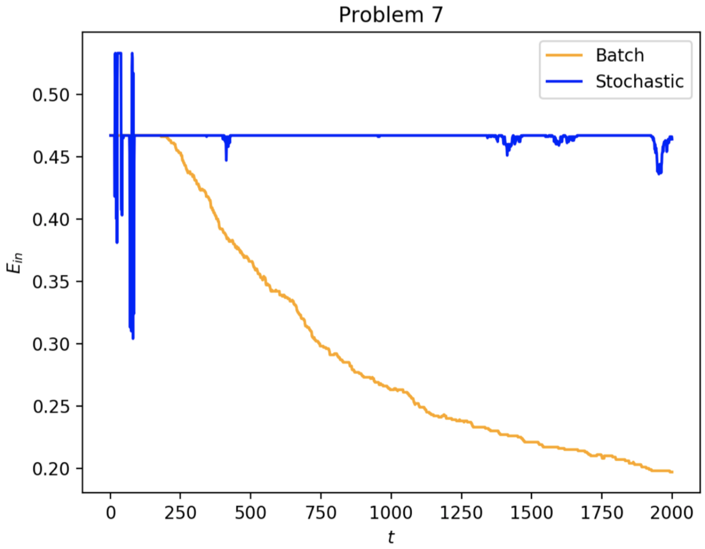
$$\begin{aligned} \frac{2\lambda}{N}\mathbf{w} + \frac{2}{N}(\mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{X}^T\mathbf{y}) = 0 \\ (\lambda\mathbf{I} + \mathbf{X}^T\mathbf{X})\mathbf{w} = \mathbf{X}^T\mathbf{y} \\ \mathbf{w} = (\lambda\mathbf{I} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \end{aligned}$$

Compared with the result in Problem 5, we see that when $\tilde{\mathbf{X}} = \sqrt{\lambda}\mathbf{I}$ and $\tilde{\mathbf{y}} = 0$, we have

$$\begin{aligned} (\mathbf{X}^T\mathbf{X} + \sqrt{\lambda}\mathbf{I}^T\sqrt{\lambda}\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{y} + \tilde{\mathbf{X}}^T 0) \\ = (\lambda\mathbf{I} + \mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y}) \end{aligned}$$

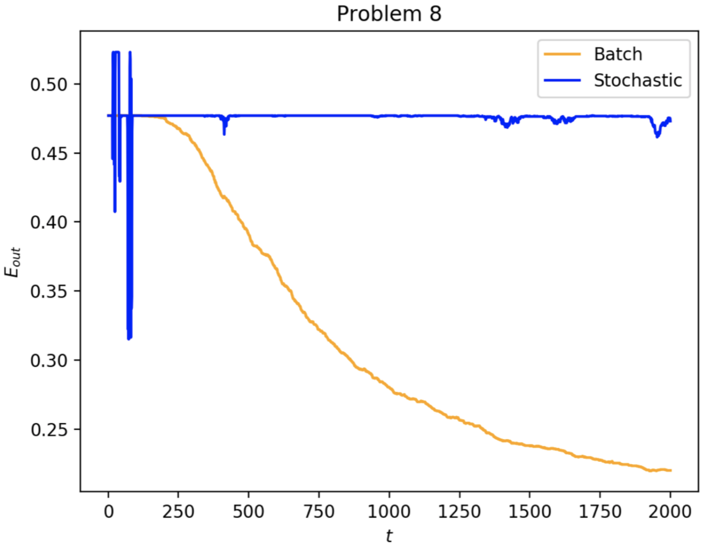
So $\tilde{\mathbf{X}} = \sqrt{\lambda}\mathbf{I}$ and $\tilde{\mathbf{y}} = 0$.

Problem 7

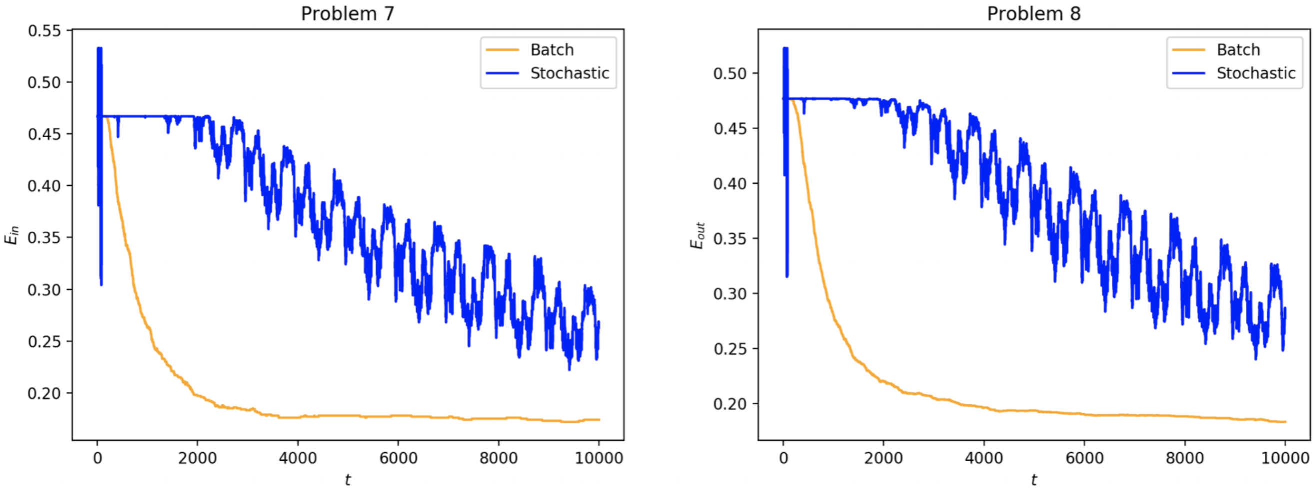


The E_{in} of gradient descent decreases as the number of iterations increases. On the other hand, stochastic gradient descent does not show much improvement in terms of E_{in} compared to gradient descent. However, as the number of iterations increases, the fluctuation of E_{in} is becoming less dramatic. Except for the huge fluctuation in the beginning of schochastic gradient desecnt, the performance of gradient descent is generally better.

Problem 8



The pattern of E_{out} is similar to that of E_{in}. Gradient descent shows steady improvement while schochastic gradient descent fluctuates around 0.47 as the number of iteration increases. We can also see that E_{out} is a little more than E_{in}, which matches the theory taught in class. I've conducted another experiment by increasing the number of iterations from 2000 to 10000. The result is shown in the following figures.



We can see both E_{in} and E_{out} of gradient descent are stable and below 0.2 after enough iterations, while both E_{in} and E_{out} of schochastic gradient descent show a tendency to decrease. We can expect after more iterations, they might converge with the error rates of gradient descent.

Problem 9

(a) Substitute $\mathrm{U}\Gamma\mathrm{V}^T$ for X and $\mathrm{V}\Gamma^{-1}\mathrm{U}^T\mathbf{y}$ for $\mathbf{w}_{\mathrm{lin}}$, we have

$$\begin{aligned}
&\mathbf{X}^T\mathbf{X}\mathbf{w}_{\mathrm{lin}}\\
=&\mathbf{X}^T(\mathbf{U}\Gamma\mathbf{V}^T)(\mathbf{V}\Gamma^{-1}\mathbf{U}^T\mathbf{y})\\
=&\mathbf{X}^T(\mathbf{U}\Gamma\Gamma^{-1}\mathbf{U}^T\mathbf{y}) \quad (\text{using } \mathbf{V}^T\mathbf{V}=\mathbf{I}_\rho)\\
=&\mathbf{X}^T(\mathbf{U}\mathbf{U}^T\mathbf{y}) \quad (\text{using } \Gamma\Gamma^{-1}=\mathbf{I}_\rho)\\
=&(\mathbf{V}\Gamma^T\mathbf{U}^T)(\mathbf{U}\mathbf{U}^T)\mathbf{y} \quad (\text{expanding } \mathbf{X}^T)\\
=&(\mathbf{V}\Gamma^T\mathbf{U}^T)\mathbf{y} \quad (\text{using } \mathbf{U}^T\mathbf{U}=\mathbf{I}_\rho)\\
=&\mathbf{X}^T\mathbf{y}
\end{aligned}$$

(b) We first introduce some notations and some mathematical facts:

- Let $A \in \mathbb{R}^{m\times n}$, we define $R(A)$ to be the column space of A.
- For $B \in \mathbb{R}^{m\times n}$, let $A = B^\dagger B$, then A is an orthogonal projection matrix of $\mathbb{R}^n$ onto $R(B^\dagger)$.

Lemma 1. *Let U be a subspace of an inner product space W. Let $z = x + y$ be a vector in W, such that $x \in U$ and $y \in U^\perp$. Then for any $z \neq x$, we have $\|z\| > \|x\|$.*

Proof. Immediate from $\|z\|^2 = \|x\|^2 + \|y\|^2$ with $\|y\| = \|z - x\| > 0$ □

Let $\mathbf{w}$ be a vector that satisfies $\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{X}\mathbf{w}_{\mathrm{lin}} = \mathbf{X}^T\mathbf{y}$. We show that the orthogonal projection of $\mathbf{w}$ onto $R((\mathbf{X}^T\mathbf{X})^\dagger)$ is $\mathbf{w}_{\mathrm{lin}}$.

$$\begin{aligned}
(\mathbf{X}^T\mathbf{X})^\dagger(\mathbf{X}^T\mathbf{X})\mathbf{w} =&(\mathbf{X}^T\mathbf{X})^\dagger(\mathbf{X}^T\mathbf{X})\mathbf{w}_{\mathrm{lin}}\\
=&(\mathbf{X}^T\mathbf{X})^\dagger\mathbf{X}^T\mathbf{y}\\
=&(\mathbf{V}\Gamma^T\mathbf{U}^T\mathbf{U}\Gamma\mathbf{V}^F)^\dagger(\mathbf{V}\Gamma^T\mathbf{U}^T)\mathbf{y}\\
=&(\mathbf{V}\Gamma^T\Gamma\mathbf{V}^T)^\dagger(\mathbf{V}\Gamma^T\mathbf{U}^T)\mathbf{y} \ \ (\text{using } \mathbf{U}^T\mathbf{U}=\mathbf{I}_\rho)\\
=&(\mathbf{V}(\Gamma^T\Gamma)^{-1}\mathbf{V}^T)(\mathbf{V}\Gamma^T\mathbf{U}^T)\mathbf{y} \ \ (\text{using the propery mentioned in } (a))\\
=&(\mathbf{V}\Gamma^{-1}(\Gamma^T)^{-1}\Gamma^T\mathbf{U}^T)\mathbf{y} \ \ (\text{using } \mathbf{V}^T\mathbf{V}=\mathbf{I}_\rho)\\
=&(\mathbf{V}\Gamma^{-1}\mathbf{U}^T)\mathbf{y}\\
=&\mathbf{w}_{\mathrm{lin}}
\end{aligned}$$

By lemma 1, for $\mathbf{w} \neq \mathbf{w}_{\mathrm{lin}}$, we have $\|\mathbf{w}_{\mathrm{lin}}\| < \|\mathbf{w}\|$. We conclude that for every $\mathbf{w}$ that satisfies $\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y}$, we have $\|\mathbf{w}_{\mathrm{lin}}\| \leq \|\mathbf{w}\|$. The equality holds when $\mathbf{w} = \mathbf{w}_{\mathrm{lin}}$.