Machine Learning Foundation (NTU, Fall 2018) Homework #4

陳熙 R07922151

1.

2. 因為 $E_{\text{aug}}(\mathbf{w}) = E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w}^T\mathbf{w}$，所以有

$$\nabla E_{\text{aug}}(\mathbf{w}) = \nabla E_{\text{in}}(\mathbf{w}) + \nabla\left(\frac{\lambda}{N}\mathbf{w}^T\mathbf{w}\right)$$

$$= \nabla E_{\text{in}}(\mathbf{w}) + \frac{2\lambda}{N}\mathbf{w}$$

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta\nabla E_{\text{aug}}(\mathbf{w}(t))$$

$$= \mathbf{w}(t) - \eta\left(\nabla E_{\text{in}}(\mathbf{w}(t)) + \frac{2\lambda}{N}\mathbf{w}(t)\right)$$

$$= \left(1 - \frac{2\eta\lambda}{N}\right)\mathbf{w}(t) - \eta\nabla E_{\text{in}}(\mathbf{w}(t))$$

3. 對於 linear model，當用(-1, 0)和(ρ, 1)作為訓練樣本，(1, 0)作為 validation 樣本時，$h_1(x) = a_1(x+1) = a_1x + a_1$，因為 $h_1(\rho) = 1$，所以，$a_1 = \frac{1}{\rho+1}$，

$h_1(x) = \frac{1}{\rho+1}(x+1)$，此時 $e_1 = (h_1(1) - 0)^2 = \left(\frac{2}{\rho+1}\right)^2 = \frac{4}{(\rho+1)^2}$。

當用(-1, 0)和(1, 0)作為訓練樣本，(ρ, 1)作為 validation 樣本時，$h_1(x) = 0$，

此時，$e_2=\left(h_1(\rho)-1\right)^2=(0-1)^2=1$。

當用$(\rho, 1)$和$(1, 0)$作為訓練樣本，$(-1, 0)$作為 validation 樣本時，$h_1(x)=a_1(x-1)$，

因為 $h_1(\rho)=1$，所以，$a_1=\dfrac{1}{p-1}$，$h_1(x)=\dfrac{1}{p-1}(x-1)$，此時，

$$e_3=\left(h_1(-1)-0\right)^2=\left(\frac{-2}{\rho-1}\right)^2=\frac{4}{(\rho-1)^2}\text{。}$$

綜上，$E_{\text{loo}}=\dfrac{1}{3}\displaystyle\sum_{n=1}^{3}e_n=\dfrac{1}{3}\left(\dfrac{4}{(\rho+1)^2}+\dfrac{4}{(\rho-1)^2}+1\right)$。

4. Coursera 上問題 12 中 $\mathbf{w}_{\text{reg}}=\arg\min\limits_{\mathbf{w}}\dfrac{\lambda}{N}\|\mathbf{w}\|^2+\dfrac{1}{N}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2$，可以寫為

$$\begin{aligned}\mathbf{w}_{\text{reg}}&=\arg\min_{\mathbf{w}}\frac{\lambda}{N}\mathbf{w}^T\mathbf{w}+\frac{1}{N}(\mathbf{X}\mathbf{w}-\mathbf{y})^T(\mathbf{X}\mathbf{w}-\mathbf{y})\\&=\arg\min_{\mathbf{w}}\frac{\lambda}{N}\mathbf{w}^T\mathbf{w}+\frac{1}{N}\left(\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}-\mathbf{y}^T\mathbf{X}\mathbf{w}-\mathbf{w}^T\mathbf{X}^T\mathbf{y}+\mathbf{y}^T\mathbf{y}\right)\end{aligned}$$

令 $f(\mathbf{w})=\dfrac{\lambda}{N}\mathbf{w}^T\mathbf{w}+\dfrac{1}{N}(\mathbf{X}\mathbf{w}-\mathbf{y})^T(\mathbf{X}\mathbf{w}-\mathbf{y})$，則

$$\begin{aligned}\nabla f(\mathbf{w})&=\nabla\frac{\lambda}{N}\mathbf{w}^T\mathbf{w}+\nabla\frac{1}{N}\left(\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}-\mathbf{y}^T\mathbf{X}\mathbf{w}-\mathbf{w}^T\mathbf{X}^T\mathbf{y}+\mathbf{y}^T\mathbf{y}\right)\\&=\frac{2\lambda}{N}\mathbf{w}+\frac{1}{N}\left(2\mathbf{X}^T\mathbf{X}\mathbf{w}-2\mathbf{X}^T\mathbf{y}\right)\\&=\frac{2}{N}\left(\left(\lambda\mathbf{I}+\mathbf{X}^T\mathbf{X}\right)\mathbf{w}-\mathbf{X}^T\mathbf{y}\right)\\&=\frac{2}{N}\left(\left(\lambda+\sum_{t=1}^{N}\|\mathbf{x}(t)\|^2\right)\mathbf{w}-\sum_{t=1}^{N}y(t)\mathbf{x}(t)\right)\\&=\frac{2}{N}\left(\lambda\mathbf{w}+\sum_{t=1}^{N}\left(\|\mathbf{x}(t)\|^2\mathbf{w}-y(t)\mathbf{x}(t)\right)\right)\\&=\frac{1}{N}\sum_{t=1}^{N}\left(\frac{2\lambda}{N}\mathbf{w}+2\left(\|\mathbf{x}(t)\|^2\mathbf{w}-y(t)\mathbf{x}(t)\right)\right)\end{aligned}$$

且$\nabla^2 f(\mathbf{w})=\dfrac{2}{N}\left(\lambda\mathbf{I}+\mathbf{X}^T\mathbf{X}\right)\succ 0$。

SGD 的 update rule 為：

$$\begin{aligned}\mathbf{w}(t+1)&=\mathbf{w}(t)-\eta\nabla_t f(\mathbf{w}(t))\\&=\mathbf{w}(t)-\eta\left(\frac{2\lambda}{N}\mathbf{w}+2\left(\|\mathbf{x}(t)\|^2\mathbf{w}-y(t)\mathbf{x}(t)\right)\right)\\&=\left(1-\frac{2\eta\lambda}{N}\right)\mathbf{w}(t)-2\eta\left(\|\mathbf{x}(t)\|^2\mathbf{w}(t)-y(t)\mathbf{x}(t)\right)\end{aligned}$$

Coursera 上問題 3 的 update rule 為 $\mathbf{w}(t+1)=\left(1-\frac{2\eta\lambda}{N}\right)\mathbf{w}(t)-\eta\nabla E_{\text{in}}(\mathbf{w}(t))$，當 $E_{\text{in}}(\mathbf{w}(t))=\left(\mathbf{w}(t)^T\mathbf{x}(t)-y(t)\right)^2$ 時，$\nabla E_{\text{in}}(\mathbf{w}(t))=2\left(\|\mathbf{x}(t)\|^2\mathbf{w}(t)-y(t)\mathbf{x}(t)\right)$，$\mathbf{w}(t+1)=\left(1-\frac{2\eta\lambda}{N}\right)\mathbf{w}(t)-2\eta\left(\|\mathbf{x}(t)\|^2\mathbf{w}(t)-y(t)\mathbf{x}(t)\right)$，此時兩個問題的 update rule 相同。說明用 SGD 加 virtual examples 的方法和 regularization 的方法實現的效果相同。

5. 根據題意，因為 x 在 $[0,2\pi]$ 均勻分佈，square error 為

$$\begin{aligned}E(w)&=\int_0^{2\pi}(\sin ax-wx)^2\,dx\\&=\int_0^{2\pi}(\sin ax)^2\,dx+\int_0^{2\pi}(wx)^2\,dx-2\int_0^{2\pi}wx\sin ax\,dx\end{aligned}$$

令 $I_1=\int_0^{2\pi}(\sin ax)^2\,dx$，$I_2=\int_0^{2\pi}(wx)^2\,dx$，$I_3=2\int_0^{2\pi}wx\sin ax\,dx$，則

$$I_1=\int_0^{2\pi}\left(\frac{1}{2}-\frac{1}{2}\cos 2ax\right)dx=\pi-\frac{1}{2}\times\frac{1}{2a}\left[\sin 2ax\right]_{x=0}^{2\pi}=\pi-\frac{\sin 4a\pi}{4a}$$

$$I_2=w^2\int_0^{2\pi}x^2dx=\frac{w^2}{3}\left[x^3\right]_{x=0}^{2\pi}=\frac{8\pi^3w^2}{3}$$

$$\begin{aligned}I_3&=2w\int_0^{2\pi}x\sin ax\,dx\\&=-\frac{2w}{a}\int_0^{2\pi}x\,d(\cos ax)\\&=-\frac{2w}{a}\left(\left[x\cos ax\right]_{x=0}^{2\pi}-\int_0^{2\pi}\cos ax\,dx\right)\\&=-\frac{2w}{a}\left(2\pi\cos 2a\pi-\frac{1}{a}\left[\sin ax\right]_{x=0}^{2\pi}\right)\\&=-\frac{2w}{a}\left(2\pi\cos 2a\pi-\frac{\sin 2a\pi}{a}\right)\end{aligned}$$

$$E(w)=\pi-\frac{\sin 4a\pi}{4a}+\frac{8\pi^3w^2}{3}+\frac{2w}{a}\left(2\pi\cos 2a\pi-\frac{\sin 2a\pi}{a}\right)$$

當 $w=-\dfrac{\frac{2}{a}\left(2\pi\cos 2a\pi-\frac{\sin 2a\pi}{a}\right)}{2\times\left(\frac{8\pi^3}{3}\right)}=-\dfrac{6\pi a\cos(2\pi a)-3\sin(2\pi a)}{8\pi^3a^2}$ 時， $E(w)$ 有最小值，所以對每個 x 的 deterministic noise 為 $\left|\sin(ax)+\dfrac{6\pi a\cos(2\pi a)-3\sin(2\pi a)}{8\pi^3a^2}x\right|$ 。