Machine Learning HW3.5

b03902089 林良翰

Linear Regression

1.

- $\sigma^2(1-\frac{d+1}{N}) \geq 0.008$
  $0.01(1-\frac{9}{N}) \geq 0.008$
  $N \geq 45$

Error and SGD

2.

- Upper bound of $[sign(\mathbf{w}^T\mathbf{x}) \neq y]$

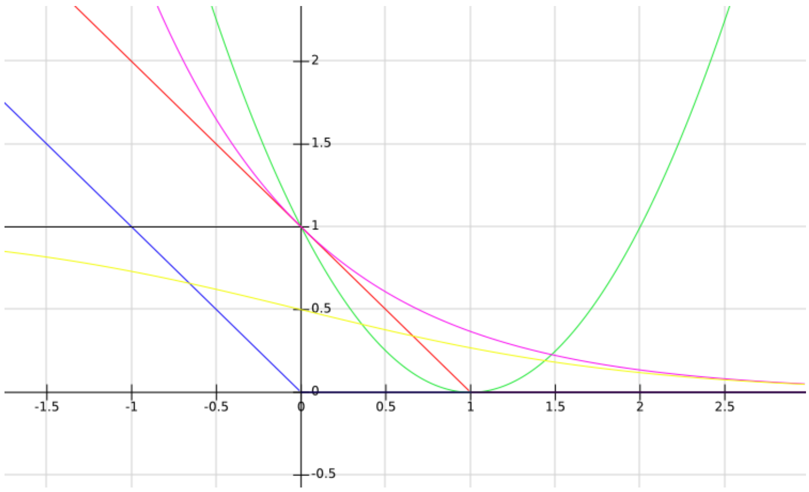
- $[a]$
  If $y = sign(\mathbf{w}^T\mathbf{x}), -y\mathbf{w}^T\mathbf{x} \leq 0 \Rightarrow max(0, 1-y\mathbf{w}^T\mathbf{x}) \geq 0 = [sign(\mathbf{w}^T\mathbf{x}) \neq y]$
  If $y \neq sign(\mathbf{w}^T\mathbf{x}), -y\mathbf{w}^T\mathbf{x} \geq 0 \Rightarrow max(0, 1-y\mathbf{w}^T\mathbf{x}) \geq 1 = [sign(\mathbf{w}^T\mathbf{x}) \neq y]$
  $\Rightarrow max(0, 1-y\mathbf{w}^T\mathbf{x}) \geq [sign(\mathbf{w}^T\mathbf{x}) \neq y]$

- $[b]$
  If $y = sign(\mathbf{w}^T\mathbf{x}), -y\mathbf{w}^T\mathbf{x} \leq 0 \Rightarrow \left(max(0, 1-y\mathbf{w}^T\mathbf{x})\right)^2 \geq 0 = [sign(\mathbf{w}^T\mathbf{x}) \neq y]$
  If $y \neq sign(\mathbf{w}^T\mathbf{x}), -y\mathbf{w}^T\mathbf{x} \geq 0 \Rightarrow \left(max(0, 1-y\mathbf{w}^T\mathbf{x})\right)^2 \geq 1 = [sign(\mathbf{w}^T\mathbf{x}) \neq y]$
  $\Rightarrow \left(max(0, 1-y\mathbf{w}^T\mathbf{x})\right)^2 \geq [sign(\mathbf{w}^T\mathbf{x}) \neq y]$

- $[c]$
  If $y = sign(\mathbf{w}^T\mathbf{x}), -y\mathbf{w}^T\mathbf{x} \leq 0 \Rightarrow max(0, -y\mathbf{w}^T\mathbf{x}) = 0 = [sign(\mathbf{w}^T\mathbf{x}) \neq y]$
  If $y \neq sign(\mathbf{w}^T\mathbf{x}), max(0, -y\mathbf{w}^T\mathbf{x}) \leq 0$, but $[sign(\mathbf{w}^T\mathbf{x}) \neq y] = 1$
  $\Rightarrow max(0, -y\mathbf{w}^T\mathbf{x})$ is not the upper bound of $[sign(\mathbf{w}^T\mathbf{x}) \neq y]$

- $[d]$
  If $y = sign(\mathbf{w}^T\mathbf{x}), 0 < \theta(-y\mathbf{w}^T\mathbf{x}) \leq 0.5 \Rightarrow \theta(-y\mathbf{w}^T\mathbf{x}) \geq [sign(\mathbf{w}^T\mathbf{x}) \neq y] = 0$
  If $y \neq sign(\mathbf{w}^T\mathbf{x}), 0.5 \leq \theta(-y\mathbf{w}^T\mathbf{x}) < 1 \Rightarrow \theta(-y\mathbf{w}^T\mathbf{x}) \leq sign(\mathbf{w}^T\mathbf{x}) \neq y] = 1$
  $\Rightarrow \theta(-y\mathbf{w}^T\mathbf{x})$ is not the upper bound of $[sign(\mathbf{w}^T\mathbf{x}) \neq y]$

- $[e]$
  If $y = sign(\mathbf{w}^T\mathbf{x}), 0 < e^{-y\mathbf{w}^T\mathbf{x}} \leq 1 \Rightarrow e^{-y\mathbf{w}^T\mathbf{x}} \geq [sign(\mathbf{w}^T\mathbf{x}) \neq y] = 0$
  If $y \neq sign(\mathbf{w}^T\mathbf{x}), e^{-y\mathbf{w}^T\mathbf{x}} \geq 1 \Rightarrow e^{-y\mathbf{w}^T\mathbf{x}} \geq sign(\mathbf{w}^T\mathbf{x}) \neq y] = 1$
  $\Rightarrow e^{-y\mathbf{w}^T\mathbf{x}} \geq [sign(\mathbf{w}^T\mathbf{x}) \neq y]$

- 

Gradient Descent and Beyond

3.

- $err(\mathbf{w}) = max(0, -y\mathbf{w}^T\mathbf{x})$
- If $y = sign(\mathbf{w}^T\mathbf{x}), -y\mathbf{w}^T\mathbf{x} \leq 0 \Rightarrow max(0, -y\mathbf{w}^T\mathbf{x}) = 0, \frac{\partial err(\mathbf{w})}{\partial \mathbf{w}} = 0$
- If $y \neq sign(\mathbf{w}^T\mathbf{x}), -y\mathbf{w}^T\mathbf{x} \geq 0 \Rightarrow max(0, -y\mathbf{w}^T\mathbf{x}) = 0, \frac{\partial err(\mathbf{w})}{\partial \mathbf{w}} = -y\mathbf{x}$
  (Ignore the point that isn't differentiable)
- According to the perceptron algorithm
  $\mathbf{w}_{t+1} = \mathbf{w}_t + y_n\mathbf{x}_n = \mathbf{w}_t - \frac{\partial err(\mathbf{w})}{\partial \mathbf{w}}$
  We can observe that $\mathbf{w}_{t+1}$ changes only if the label is wrong, otherwise it remains the same.

4.

- $E(u_5, v_5) \approx 2.825$

```
import math as m

ETA = 0.01
E = lambda u, v : m.exp(u) + m.exp(2*v) + m.exp(u*v) + u**2 - 2*u*v + 2
GE = lambda u, v : (m.exp(u) + v*m.exp(u*v) + 2*u - 2*v - 3, 2*m.exp(2*

u = v = 0
print ('E(u0, v0) =', E(u, v))
for i in range(5):
    g = GE(u, v)
    u -= ETA * g[0]
    v -= ETA * g[1]
    print ('E(u', i+1, ', v', i+1, ') = ', E(u, v), sep='')
```

5.

- $b = \frac{1}{0!}E(0, 0) = 3$

- $b_v = \frac{1}{1!}\frac{\partial E(0,0)}{\partial v} = 0$
- $b_u = \frac{1}{1!}\frac{\partial E(0,0)}{\partial u} = -2$
- $b_{uv} = \frac{1}{2!}\frac{\partial E(0,0)}{\partial u \partial u} = -1$
- $b_{vv} = \frac{1}{2!}\frac{\partial E(0,0)}{\partial^2 v} = 4$
- $b_{uu} = \frac{1}{2!}\frac{\partial E(0,0)}{\partial^2 u} = 1.5$
- $(b_{uu}, b_{vv}, b_{uv}, b_u, b_v, b) = (1.5, 4, -1, -2, 0, 3)$

6.

- *Def*. of Newton Direction (https://en.wikipedia.org/wiki/Newton's_method_in_optimization): find an $\mathbf{x}$ $s.t.$
  $f'(\mathbf{x}) + \Delta\mathbf{x} f''(\mathbf{x}) = 0$
  $\Rightarrow \Delta\mathbf{x} = -\frac{f'(\mathbf{x})}{f''(\mathbf{x})}$
- $\nabla E(u, v) + \Delta(u, v)\nabla^2 E(u, v) = 0$
  $\Rightarrow (\Delta u, \Delta v) = -\frac{\nabla E(u,v)}{\nabla^2 E(u,v)}$
- Details of $\nabla E(u, v) + \Delta(u, v)\nabla^2 E(u, v) = 0$
  Denote $f_{xy}(x, y) = \frac{\partial E(x,y)}{\partial x \partial y}$
  $f_u(u, v) + f_v(u, v) + \Delta(u, v)\Big(\nabla f_u(u, v) + \nabla f_v(u, v)\Big) = 0$
  $f_u(u, v) + f_v(u, v) + \Delta u\Big(\nabla f_u(u, v) + \nabla f_v(u, v)\Big) + \Delta v\Big(\nabla f_u(u, v) + \nabla f_v(u, v)\Big) = 0$
  $f_u(u, v) + f_v(u, v) + \Delta u\Big(f_{uu}(u, v) + f_{vu}(u, v)\Big) + \Delta v\Big(f_{uv}(u, v) + f_{vv}(u, v)\Big) = 0$
  $\Rightarrow \Delta u = -\frac{f_u(u,v)}{f_{uu}(u,v) + f_{vu}(u,v)}$
  $\Rightarrow \Delta v = -\frac{f_v(u,v)}{f_{uv}(u,v) + f_{vv}(u,v)}$

7.

- $E(u_5, v_5) \approx 2.361$

```
import math as m

e = m.e

E = lambda u, v : e**u + e**(2*v) + e**(u*v) + u**2 - 2*u*v + 2*v**2 -
Gu = lambda u, v : e**u + v*e**(u*v) + 2*u - 2*v - 3
Gv = lambda u, v : 2*e**(2*v) + u*e**(u*v) - 2*u + 4*v - 2
Guu = lambda u, v : e**u + (v**2)*e**(u*v) + 2
Gvv = lambda u, v : 4*e**(2*v) + (u**2)*e**(u*v) + 4
Guv = lambda u, v : e**(u*v) - 2

u = v = 0
print('E(u0, v0) =', E(u, v))
for i in range(5):
    nu = -1 * Gu(u, v) / (Guu(u, v) + Guv(u, v))
    nv = -1 * Gv(u, v) / (Gvv(u, v) + Guv(u, v))
    u += nu
    v += nv
    print ('E(u', i+1, ', v', i+1, ') = ', E(u, v), sep='')
```

Regularization and Weight Decay

8.

- $\mathbf{w}(t+1) \leftarrow \mathbf{w}(t) - \eta \nabla E_{aug}\left(\mathbf{w}(t)\right)$
- $\nabla E_{aug}(\mathbf{w})$
  $= \nabla\left(E_{in}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w}^T\mathbf{w}\right)$
  $= \nabla E_{in}(\mathbf{w}) + \frac{2\lambda}{N}\mathbf{w}$
- $\Rightarrow \mathbf{w}(t+1) \leftarrow \left(1 - \frac{2\eta\lambda}{N}\right)\mathbf{w}(t) - \eta \nabla E_{in}\left(\mathbf{w}(t)\right)$
- $\alpha = 1 - \frac{2\eta\lambda}{N}$
- $\beta = -\eta$

Virtual Examples and Regularization

9.

- $X = \underbrace{\begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}}_{N\times(d+1)}, y = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{N\times 1}, \tilde{X} = \underbrace{\begin{bmatrix} \tilde{x}_1^T \\ \tilde{x}_2^T \\ \vdots \\ \tilde{x}_K^T \end{bmatrix}}_{K\times(d+1)}, \tilde{y} = \underbrace{\begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \vdots \\ \tilde{y}_K \end{bmatrix}}_{K\times 1}, w = \underbrace{\begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}}_{(d+1)\times 1}$

- We can obtain

$$
\begin{aligned}
E_{in}(\mathbf{w}) &= \frac{1}{N+K}\left(\|X\mathbf{w}-y\|^2+\left\|\tilde{X}\mathbf{w}-\tilde{y}\right\|^2\right)\\
&= \frac{1}{N+K}\left((w^TX^TXw-2w^TX^Ty+y^Ty)+(w^T\tilde{X}^T\tilde{X}w-2w^T\tilde{X}^T\tilde{y}+\tilde{y}^T\tilde{y})\right)\\
\frac{\partial E_{in}(\mathbf{w})}{\partial \mathbf{w}} &= \frac{2}{N+K}\left((X^TXw-X^Ty)+(\tilde{X}^T\tilde{X}w-\tilde{X}^T\tilde{y})\right)
\end{aligned}
$$

- To obtain optimal $\mathbf{w}$, we need to solve

$$
\frac{\partial E_{in}(\mathbf{w})}{\partial \mathbf{w}} = \underbrace{\begin{bmatrix}0\\0\\\vdots\\0\end{bmatrix}}_{(d+1)\times 1}
$$

- $\Rightarrow \mathbf{w} = \left(X^TX+\tilde{X}^T\tilde{X}\right)^{-1}\left(X^Ty+\tilde{X}^T\tilde{y}\right)$

10.

- we have known
  $\mathbf{w}_{REG} = \left(X^TX+\lambda I\right)^{-1}\left(X^Ty\right)$
  $\mathbf{w}_{VIR} = \left(X^TX+\tilde{X}^T\tilde{X}\right)^{-1}\left(X^Ty+\tilde{X}^T\tilde{y}\right)$
- Let $\tilde{X} = \sqrt{\lambda}I, \tilde{y} = 0\ s.t.\ \mathbf{w}_{VIR} = \mathbf{w}_{REG}$

Experiments with Logistic Regression

11.

- $E_{out} = 0.475$

12.

- $E_{out} = 0.473$

Experiment with Regularized Linear Regression and Validation

13.

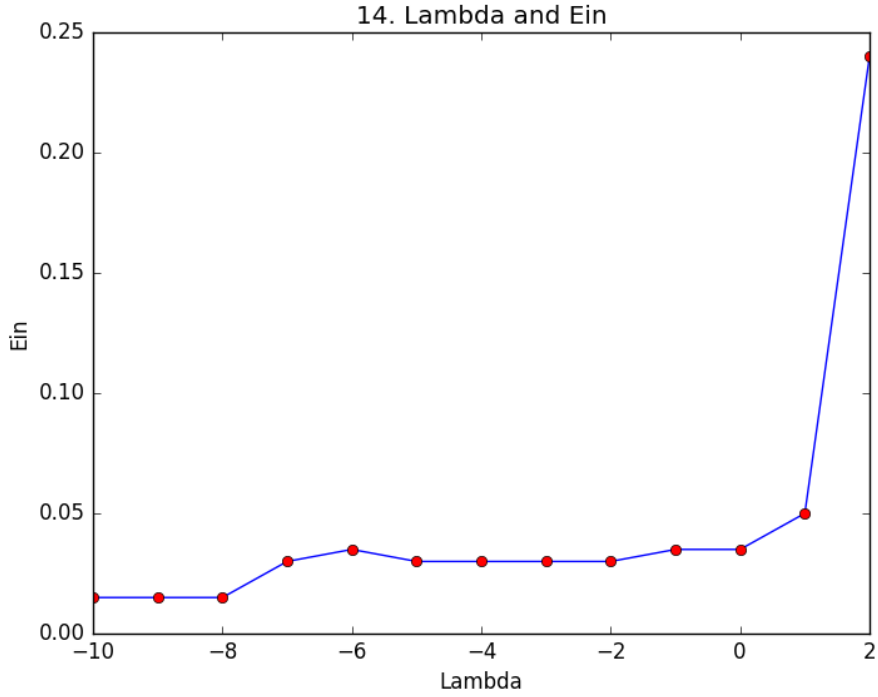
- Using whole data D
- $\lambda = 1.126$
- $E_{in} = 0.035, E_{out} = 0.02$

14.

- Using whole data D
- $\log_{10}\lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$
-

λ	E_{in}	E_{out}
10^{2}	0.240	0.261
10^{1}	0.050	0.045
10^{0}	0.035	0.020
10^{-1}	0.035	0.016
10^{-2}	0.030	0.016
10^{-3}	0.030	0.016
10^{-4}	0.030	0.016
10^{-5}	0.030	0.016
10^{-6}	0.035	0.016
10^{-7}	0.030	0.015
10^{-8}	0.015	0.020
10^{-9}	0.015	0.020
10^{-10}	0.015	0.020

- $\arg\min_{\lambda} E_{in}(g_\lambda) = 10^{-8}$
- $E_{in}(g_{10^{-8}}) = 0.015, E_{out}(g_{10^{-8}}) = 0.02$

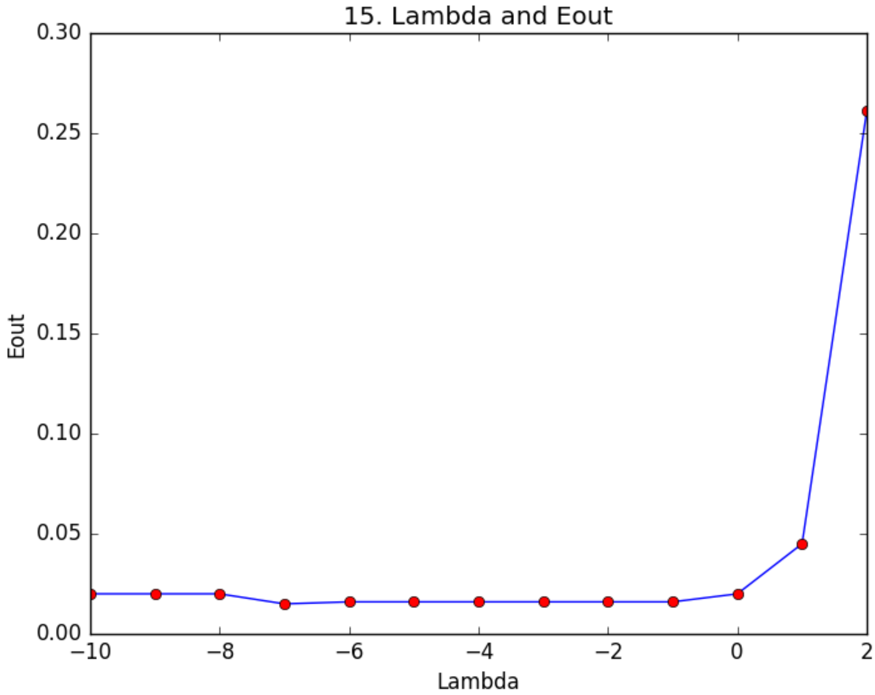


15.

- Continue from problem 14.

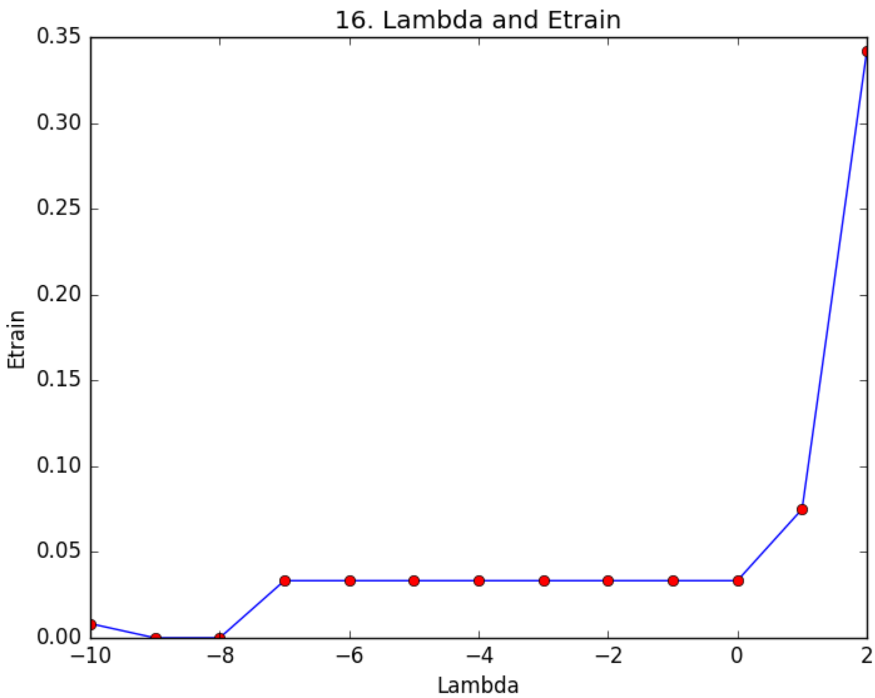
- $\arg\min_{\lambda} E_{out}(g_\lambda) = 10^{-7}$

  $E_{out}(g_{10^{-7}}) = 0.016$



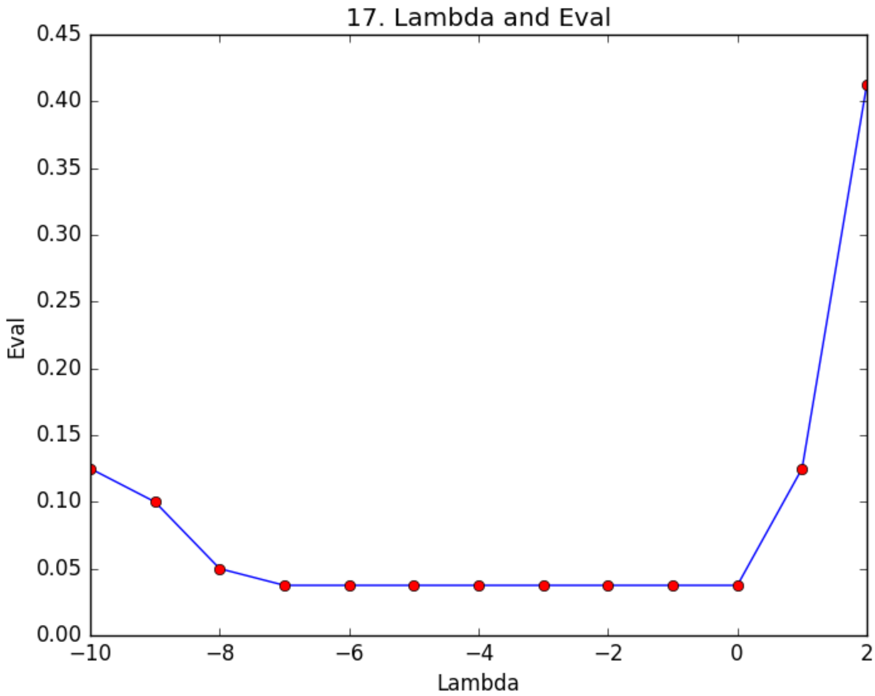
16.

- Using D_{train} and D_{val}
- $\log_{10}\lambda = \{2, 1, 0, -1, \ldots, -8, -9, -10\}$
- $\arg\min_{\lambda} E_{train}(g_\lambda^-) = 10^{-8}$
- $E_{train}(g_{10^{-8}}^-) = 0.0, E_{out}(g_{10^{-8}}^-) = 0.025$



17.

- Continue from problem 16.
- $\arg\min_{\lambda} E_{val}(g_{\lambda}^{-}) = 10^0$
- $E_{val}(g_{10^0}^{-}) = 0.037, E_{out}(g_{10^0}^{-}) = 0.028$



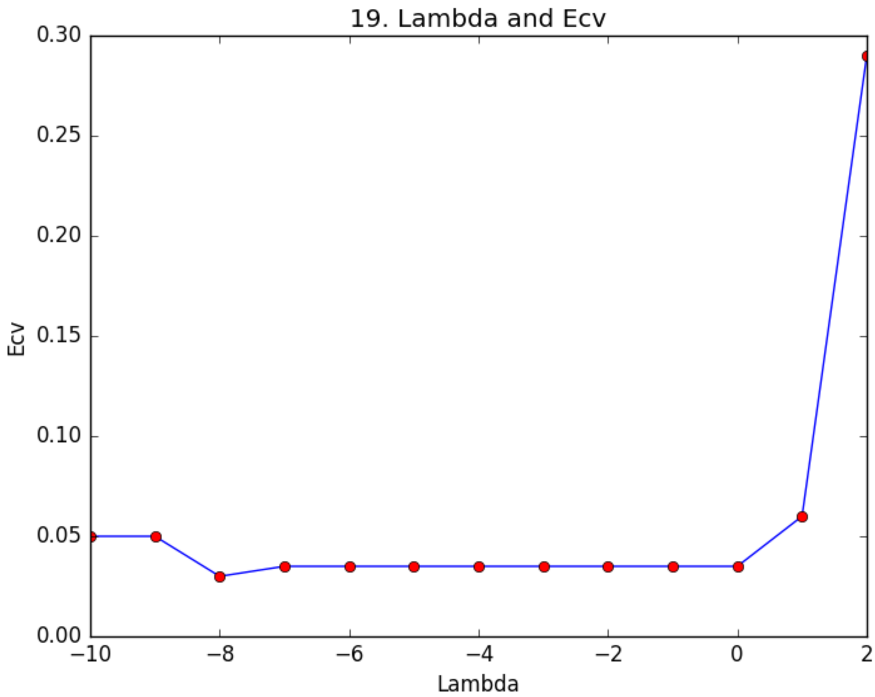
18.

- Continue from problem 17., using whole data D with optimal $\lambda = 10^0$ in 17.
- According to the table of problem 14.
  $E_{in}(g_{10^0}) = 0.035, E_{out}(g_{10^0}) = 0.02$

19.

- Split first 200 data of D into 5 parts, and use different part to be D_{val}, while other parts are D_{train}
- $\arg\min_{\lambda} E_{cv}(g_{\lambda}^{cv}) = 10^{-8}$

- $E_{cv}(g^{cv}_{10^{-8}}) = 0.03$



20.

- Continue from problem 19., using whole data D with optimal $\lambda = 10^{-8}$ in 19.
- According to the table of problem 14.
  $E_{in}(g_{10^{-8}}) = 0.015, E_{out}(g_{10^{-8}}) = 0.02$

More on Virtual Examples

21.

$$set \quad E_{aug} = \frac{1}{N}\|Xw - y\|^2 + \frac{\lambda}{N}\|\Gamma w\|^2$$

$$w_{reg} = \underset{w}{\operatorname{argmin}} \frac{1}{N}\|Xw - y\|^2 + \frac{\lambda}{N}\|\Gamma w\|^2$$

$$= \underset{w}{\operatorname{argmin}} \frac{1}{N}\left(w^T X^T X w - 2w^T X^T y + y^T y\right) + \frac{\lambda}{N}\left(w^T \Gamma^T \Gamma w\right)$$

$$\frac{\partial E_{aug}}{\partial w} = \frac{1}{N}\left(2X^T X w - 2X^T y\right) + \frac{\lambda}{N}\left(2\Gamma^T \Gamma w\right) = 0$$

$$\rightarrow \left(X^T X + \lambda \Gamma^T \Gamma\right) w_{reg} = X^T y$$

$$\rightarrow w_{reg} = \left(X^T X + \lambda \Gamma^T \Gamma\right)^{-1} X^T y$$

From problem 11, we can thus have

$w_{reg} = \left(X^T X + \lambda \Gamma^T \Gamma\right)^{-1} X^T y = \left(X^T X + \tilde{X}^T \tilde{X}\right)^{-1}\left(X^T y + \tilde{X}^T \tilde{y}\right)$, so we can then

choose $\tilde{X} = \sqrt{\lambda}\Gamma, \ \tilde{y} = 0$ as virtual examples.

22.

$$set\ \ E_{aug} = \frac{1}{N}\|Xw - y\|^2 + \frac{\lambda}{N}\|w - w_{hint}\|^2$$

$$w_{reg} = \underset{w}{\operatorname{argmin}} \frac{1}{N}\|Xw - y\|^2 + \frac{\lambda}{N}\|w - w_{hint}\|^2$$

$$= \underset{w}{\operatorname{argmin}} \frac{1}{N}\left(w^T X^T X w - 2w^T X^T y + y^T y\right) + \frac{\lambda}{N}\left(w^T w - 2w^T w_{hint} + w_{hint}^T w_{hint}\right)$$

$$\frac{\partial E_{aug}}{\partial w} = \frac{1}{N}\left(2X^T X w - 2X^T y\right) + \frac{\lambda}{N}\left(2w - 2w_{hint}\right) = 0$$

$$\rightarrow \left(X^T X + \lambda I\right) w_{reg} = X^T y + w_{hint}$$

$$\rightarrow w_{reg} = \left(X^T X + \lambda I\right)^{-1}\left(X^T y + w_{hint}\right)$$

From problem 11, we can thus have

$w_{reg} = \left(X^T X + \lambda I\right)^{-1}\left(X^T y + w_{hint}\right) = \left(X^T X + \tilde{X}^T \tilde{X}\right)^{-1}\left(X^T y + \tilde{X}^T \tilde{y}\right)$, so we can then

choose $\tilde{X} = \sqrt{\lambda} I,\ \ \tilde{y} = \frac{w_{hint}}{\sqrt{\lambda}}$ as virtual examples.