# **Machine Learning HW6**

B03902089 資工三 林良翰

#### **Descent Methods for Probabilistic SVM**

ullet Probabilistic SVM:  $\min_{A,B} rac{1}{N} \sum_{n=1}^{N} \ln \left( 1 + \exp \left( -y_n \left( A \cdot \left( w_{SVM}^T \phi \left( x_n 
ight) + b_{SVM} 
ight) + B 
ight) 
ight) 
ight)$ 

Let

$$z_n=w_{SVM}^T\phi\left(x_n
ight)+b_{SVM}$$
  $p_n= heta\left(-y_n\left(Az_n+B
ight)
ight)$  , where  $heta\left(s
ight)=rac{\exp(s)}{1+\exp(s)}$ 

1.

$$ullet$$
 Let  $s_n = -y_nig(Az_n + Big)$   $\Rightarrow Fig(A, Big) = rac{1}{N}\sum_{n=1}^N \lnig(1 + \exp(s_n)ig)$ 

$$egin{aligned} ullet rac{\partial F(A,B)}{\partial A} &= rac{1}{N} \sum_{n=1}^N rac{1}{1+\exp(s_n)} \exp(s_n) rac{\partial s_n}{\partial A} &= rac{1}{N} \sum_{n=1}^N -p_n y_n z_n \ rac{\partial F(A,B)}{\partial B} &= rac{1}{N} \sum_{n=1}^N rac{1}{1+\exp(s_n)} \exp(s_n) rac{\partial s_n}{\partial B} &= rac{1}{N} \sum_{n=1}^N -p_n y_n \end{aligned}$$

$$ullet \ 
abla F\left(A,B
ight) = rac{1}{N} \sum\limits_{n=1}^{N} \left[ -p_n y_n z_n, -p_n y_n 
ight]^T$$

2.

ullet Definition of Hessian Matrix Hig(fig) of  $fig(x_1,x_2,\ldots,x_nig)$ 

$$H(f) = egin{bmatrix} rac{\partial^2 f}{\partial x_1^2} & rac{\partial^2 f}{\partial x_1 x_2} & \cdots & rac{\partial^2 f}{\partial x_1 x_n} \ rac{\partial^2 f}{\partial x_2 x_1} & rac{\partial^2 f}{\partial x_2^2} & \cdots & rac{\partial^2 f}{\partial x_2 x_n} \ dots & dots & \ddots & dots \ rac{\partial^2 f}{\partial x_n x_1} & rac{\partial^2 f}{\partial x_n x_2} & \cdots & rac{\partial^2 f}{\partial x_n^2} \ \end{bmatrix}$$

ullet Let  $s_n = -y_n ig(A z_n + Big)$ 

$$egin{aligned} rac{\partial -p_n y_n z_n}{\partial A} &= rac{\partial - heta \left(\partial s_n
ight) y_n z_n}{\partial s_n} rac{\partial s_n}{\partial A} \ &= rac{e^{s_n}}{1+e^{s_n}} igg(1-rac{e^{s_n}}{1+e^{s_n}}igg) \left(y_n z_n
ight)^2 \ &= \left(y_n z_n
ight)^2 p_n \left(1-p_n
ight) \end{aligned}$$

$$\begin{split} \frac{\partial -p_n y_n z_n}{\partial B} &= \frac{\partial -\theta \left(\partial s_n\right) y_n z_n}{\partial s_n} \frac{\partial s_n}{\partial B} \\ &= \frac{e^{s_n}}{1+e^{s_n}} \left(1-\frac{e^{s_n}}{1+e^{s_n}}\right) \left(y_n^2 z_n\right) \\ &= \left(y_n^2 z_n\right) p_n \left(1-p_n\right) \\ \frac{\partial -p_n y_n}{\partial B} &= \frac{\partial -\theta \left(\partial s_n\right) y_n}{\partial s_n} \frac{\partial s_n}{\partial B} \\ &= \frac{e^{s_n}}{1+e^{s_n}} \left(1-\frac{e^{s_n}}{1+e^{s_n}}\right) \left(y_n\right)^2 \\ &= \left(y_n\right)^2 p_n \left(1-p_n\right) \\ \bullet & H(F(A,B)) &= \frac{1}{N} \sum_{n=1}^N \left[\frac{\left(y_n z_n\right)^2 p_n \left(1-p_n\right) - \left(y_n^2 z_n\right) p_n \left(1-p_n\right)}{\left(y_n^2 z_n\right) p_n \left(1-p_n\right)}\right] \end{split}$$

## **Kernel Ridge Regression**

- ullet Guassian Kernel  $K\left(x,x'
  ight)=\exp{\left(-\gamma\|x-x'\|^2
  ight)}$
- Kernel Ridge Regression:

$$\circ$$
 Want to Minimize:  $\min_{eta} E_{aug}\left(eta
ight) = \min_{eta} rac{\lambda}{N} eta^T K eta + rac{1}{N} \left(eta^T K^T K eta - 2eta^T K^T y + y^T y
ight)$ 

$$\circ$$
 Solving:  $abla E_{aug}\left(eta
ight)=rac{2}{N}K^{T}\left(\left(\lambda I+K
ight)eta-y
ight)=0$ 

• Obtain: 
$$\beta = (\lambda I + K)^{-1} y$$

3.

• 
$$\gamma \to \infty$$

$$egin{aligned} & \lim_{\gamma o \infty} K\left(x, x'
ight) = \lim_{\gamma o \infty} e^{-\gamma \|x - x'\|^2} = I \ & egin{aligned} & eta = (\lambda I + I)^{-1} y \end{aligned}$$

$$\bullet \quad \beta = (\lambda I + I)^{-1} y$$

4.

$$\bullet$$
  $\gamma \rightarrow 0$ 

$$egin{aligned} ullet & \gamma o 0 \ ullet & \lim_{\gamma o 0} K\left(x, x'
ight) = \lim_{\gamma o 0} e^{-\gamma \|x - x'\|^2} = J \end{aligned}$$

 $\boldsymbol{J}$  is the matrix of all ones.

$$\bullet \quad \beta = (\lambda I + J)^{-1} y$$

#### **Support Vector Regression**

$$\begin{split} \bullet \quad & (P_2) \min_{b,w,\xi^\vee,\xi^\wedge} \tfrac{1}{2} w^T w + C \sum_{n=1}^N \left( \left( \xi_n^\vee \right)^2 + \left( \xi_n^\wedge \right)^2 \right) \\ \text{s.t.} \quad & -\epsilon - \xi_n^\vee \leq y_n - w^T \phi \left( x_n \right) - b \leq \epsilon + \xi_n^\wedge \end{split}$$

5.

$$ullet$$
 Let  $A_n = y_n - w^T \phi\left(x_n
ight) - b$ 

• Lagrange Multiplier Method:

$$egin{aligned} \mathcal{L}\left(P_{2}
ight) \min_{b,w,\xi^{ee},\xi^{\wedge}lpha^{ee},lpha^{\wedge}} \max_{2}rac{1}{2}w^{T}w + C\sum_{n=1}^{N}\left(\left(\xi_{n}^{ee}
ight)^{2} + \left(\xi_{n}^{\wedge}
ight)^{2}
ight) \ &+ lpha^{ee}\sum_{n=1}^{N}\left(A_{n} + \left(\epsilon + \xi_{n}^{ee}
ight)
ight) + lpha^{\wedge}\sum_{n=1}^{N}\left(A_{n} - \left(\epsilon + \xi_{n}^{\wedge}
ight)
ight) \end{aligned}$$

• Partial derivative on £

$$egin{aligned} rac{\partial \mathcal{L}}{\partial \xi_n^{ee}} &= 2C \xi_n^{ee} + lpha^{ee} = 0 \Rightarrow lpha^{ee} = -2C \xi_n^{ee} \ rac{\partial \mathcal{L}}{\partial \xi_n^{ee}} &= 2C \xi_n^{\wedge} - lpha^{\wedge} = 0 \Rightarrow lpha^{\wedge} = 2C \xi_n^{\wedge} \ \Rightarrow L\left(P_2
ight) \min_{b,w,\xi^{ee},\xi^{\wedge}} rac{1}{2} w^T w - C \sum_{n=1}^N \left( \left( \xi_n^{ee} 
ight)^2 + \left( \xi_n^{\wedge} 
ight)^2 
ight) \ - 2C \sum_{n=1}^N \xi_n \left( A_n + \epsilon 
ight) + 2C \sum_{n=1}^N \xi_n^{\wedge} \left( A_n - \epsilon 
ight) \end{aligned}$$

• Partial derivative on **\xi** again

$$egin{array}{l} rac{\partial \mathcal{L}}{\partial \xi_n^{ee}} &= -2C\xi_n^{ee} - 2C\left(A_n + \epsilon
ight) = 0 \Rightarrow \xi_n^{ee} = -\left(A_n + \epsilon
ight) ext{ and } A_n \leq -\epsilon \ rac{\partial \mathcal{L}}{\partial \xi_n^{ee}} &= -2C\xi_n^{\wedge} + 2C\left(A_n - \epsilon
ight) = 0 \Rightarrow \xi_n^{\wedge} = \left(A_n - \epsilon
ight) ext{ and } A_n \geq +\epsilon \ \Rightarrow L\left(P_2
ight) \min_{b,w} rac{1}{2} w^T w + C \sum_{n=1}^N \left(\left[A_n \leq -\epsilon
ight] \left(A_n + \epsilon
ight)^2 + \left[A_n \geq +\epsilon
ight] \left(A_n - \epsilon
ight)^2
ight) \end{array}$$

• Transform the above equation into non-linear

$$egin{aligned} \mathcal{L}\left(P_{2}
ight) \min_{b,w} rac{1}{2} w^{T} w + C \sum_{n=1}^{N} \left( \max\left(0,\left|A_{n}
ight| - \epsilon 
ight) 
ight)^{2} \ \Rightarrow \mathcal{L}\left(P_{2}
ight) \min_{b,w} rac{1}{2} w^{T} w + C \sum_{n=1}^{N} \left( \max\left(0,\left|y_{n} - w^{T} \phi\left(x_{n}
ight) - b
ight| - \epsilon 
ight) 
ight)^{2} \end{aligned}$$

6.

$$ullet$$
 Let  $s_n = \sum\limits_{m=1}^{N} \left( eta_m K\left( x_n, x_m 
ight) + b 
ight)$ 

• From 5.

$$F\left(b,eta
ight) = rac{1}{2}\sum_{m=1}^{N}\left(\sum_{n=1}^{N}eta_{n}eta_{m}K\left(x_{n},x_{m}
ight)
ight) + C\sum_{n=1}^{N}\left(\max\left(0,\left|y_{n}-w^{T}\phi\left(x_{n}
ight)-b
ight|-\epsilon
ight)
ight)^{2}$$

$$ullet rac{\partial s_n}{\partial eta_m} = K\left(x_n, x_m
ight)$$
 denote as  $K$ 

$$egin{align*} rac{\partial eta_m}{\partial F(b,eta)} &= rac{1}{2} \sum_{n=1}^N eta_n K + C \sum_{n=1}^N [|y_n - s_n| \geq \epsilon] \, rac{\partial (|y_n - s_n| - \epsilon)^2}{\partial eta_m} \ &= rac{1}{2} \sum_{n=1}^N eta_n K + \left\{egin{align*} y_n - s_n \geq 0, C \sum_{n=1}^N [|y_n - s_n| \geq \epsilon] \, rac{\partial (y_n - s_n - \epsilon)^2}{\partial eta_m} \ y_n - s_n \leq 0, C \sum_{n=1}^N [|y_n - s_n| \geq \epsilon] \, rac{\partial (s_n - y_n - \epsilon)^2}{\partial eta_m} \ &= rac{1}{2} \sum_{n=1}^N eta_n K + \left\{egin{align*} y_n - s_n \geq 0, 2C \sum_{n=1}^N [|y_n - s_n| \geq \epsilon] \, (y_n - s_n - \epsilon) \, (-K) \ y_n - s_n \leq 0, 2C \sum_{n=1}^N [|y_n - s_n| \geq \epsilon] \, (s_n - y_n - \epsilon) \, K \ &= rac{1}{2} \sum_{n=1}^N eta_n K + 2C \sum_{n=1}^N [|y_n - s_n| \geq \epsilon] \, (|y_n - s_n| - \epsilon) \, sign \, (y_n - s_n) \, K \end{array}
ight.$$

## Blending

- Since there are only 2 points, the best hypothesis is simply the line passing through these 2 points. Suppose the 2 points are  $(x_1, x_1^2)$  and  $(x_2, x_2^2)$
- The best hypothesis can be represented as

$$h\left(x
ight) = rac{x_{1}^{2}-x_{2}^{2}}{x_{1}-x_{2}}(x-x_{1}) + x_{1}^{2} = \left(x_{1}+x_{2}
ight)x - x_{1}x_{2}$$

- $\bar{g}(x) = E[h(x)] = E[x_1 + x_2]x + E[x_1x_2]$
- Since the  $\boldsymbol{x}$  value of points are sampled from unifrom distribution over [0,1] $\Rightarrow ar{g}\left(x
  ight) = E\left[x_1 + x_2
  ight]x + E\left[x_1x_2
  ight] = x - rac{1}{4}$

### **Test Set Linear Regression**

8.

- Define a cheating hypothesis.
  - $g_i\left(x_i\right)=[i=j]$ , where  $1\leq i,j\leq N$ . The function will output 1 if i=j, else output 0. Define a special hypothesis that will always output 0.

$$g_0\left(x_j
ight)=0$$
, where  $1\leq j\leq N$ .

• Construct a series of cheating hypothesis

$$[g_0, g_1, g_2, \dots, g_{n-2}, g_{n-1}]$$

- ullet Query RMSE for N times to obtain  $RMSE\left(g_{i}
  ight)$ , where  $0\leq i\leq N-1$
- Now we can compute every  $ilde{y}_i$

$$ilde{y}_i = rac{1}{2} \Big( N \left( \left[ RMSE\left(g_0
ight) 
ight]^2 - \left[ RMSE\left(g_i
ight) 
ight]^2 \Big) + 1 \Big)$$

ullet  $ilde{y}_n$  can be computed from all the other  $ilde{y}_i$  with  $RMSE\left(g_0
ight)$ Thus we need N queries.

9.

- Continue from 8., we use  $g_0$  again  $g_0\left(x_j\right)=0$ , where  $1\leq j\leq N$ .
- List out two equations below

$$egin{aligned} \left[RMSE\left(g_{0}
ight)
ight]^{2} &= rac{1}{N}\sum_{i=1}^{N}\left( ilde{y}_{i}
ight)^{2} = rac{1}{N} ilde{y}^{T} ilde{y} \ &\left[RMSE\left(g
ight)
ight]^{2} = rac{1}{N}\sum_{i=1}^{N}\left( ilde{y}_{i} - g\left(x_{i}
ight)
ight)^{2} = rac{1}{N}\left( ilde{y}^{T} ilde{y} - 2g^{T} ilde{y} + g^{T}g
ight) \end{aligned}$$

$$ullet$$
 We can obtain  $g^T ilde{y}$  by the equations above  $g^T ilde{y}=rac{1}{2}\Big(N\left([RMSE\left(g_0
ight)]^2-[RMSE\left(g
ight)]^2\Big)+g^Tg\Big)$ 

• Thus we only need 2 gueries.

10.

• Continue from 8. 9., we use  $g_0$ ,  $g^T \tilde{y}$  again

$$g^T ilde{y} = rac{1}{2} \Big( N \left( \left[ RMSE\left(g_0
ight) 
ight]^2 - \left[ RMSE\left(g
ight) 
ight]^2 \Big) + g^T g \Big)$$

ullet The problem is to obtain optimal  $[lpha_1,lpha_2,\ldots,lpha_K]$  for

$$egin{aligned} \min_{lpha_1,lpha_2,\ldots,lpha_K} RMSE\left(\sum\limits_{k=1}^K lpha_k g_k
ight) \ RMSE\left(\sum\limits_{k=1}^K lpha_k g_k
ight) &= rac{1}{N}igg( ilde{y}^T ilde{y} - 2igg(\sum\limits_{k=1}^K lpha_k g_k
ight)^T ilde{y} + igg(\sum\limits_{k=1}^K lpha_k g_k
ight)^Tigg(\sum\limits_{k=1}^K lpha_k g_k
ight) \end{aligned}$$

• Partial derivative by  $\alpha_s$ , where  $1 \leq s \leq K$ 

$$rac{\partial RMSE\left(\sum\limits_{k=1}^{K}lpha_kg_k
ight)}{\partiallpha_s}=-2(g_s)^T ilde{y}+2(g_s)^T\sum\limits_{k=1}^{K}lpha_kg_k=0$$

From 9., we can calculate  $(g_s)^T ilde{y}$ 

• We can obtain K equations to solve all  $\alpha$   $\Rightarrow$  Require K+1 queries.

#### **Experiment with Kernel Ridge Regression**

11.

• 400 training data, 100 testing data.

 $\bullet$   $E_{in}$ 

$\lambda \setminus \gamma$	32	2	0.125
0.001	0.0	0.0	0.0
1	0.0	0.0	0.03
1000	0.0	0.0	0.2425

ullet  $\gamma=32,2$  or  $(\lambda,\gamma)=(0.001,0.125)$  have the minimum  $E_{in}=0.0$ 

12.

 $\bullet$   $E_{out}$ 

$\lambda \setminus \gamma$	32	2	0.125
0.001	0.45	0.44	0.46
1	0.45	0.44	0.45
1000	0.45	0.44	0.39

ullet  $(\lambda,\gamma)=(1000,0.125)$  has the minimum  $E_{in}=0.39$ 

#### **Experiment with Support Vector Regression**

13.

- 400 training data, 100 testing data.
- $\bullet$   $E_{in}$

$\lambda \setminus \gamma$	32	2	0.125
0.001	0.4	0.4	0.4
1	0.0	0.0	0.035
1000	0.0	0.0	0.0

ullet  $\gamma=1000$  or  $(\lambda,\gamma)=(1,32)$  or (1,2) have minimum  $E_{in}=0.0$ 

#### 14.

 $\bullet$   $E_{out}$ 

$\lambda \setminus \gamma$	32	2	0.125
0.001	0.48	0.48	0.48
1	0.48	0.48	0.42
1000	0.48	0.48	0.47

ullet  $(\lambda,\gamma)=(1,0.125)$  has the minimum  $E_{in}=0.42$ 

## **Experiment with Bagging Ridge Regression**

#### 15. 16.

- Bootstrap aggregation on 400 training data, 200 iterations.
- 100 testing data.
- ullet  $E_{in}$  and  $E_{out}$

λ	$E_{in}$	$E_{out}$
0.01	0.3115	0.3710
0.1	0.3105	0.3677
1	0.3126	0.3693
10	0.3109	0.3690
100	0.3128	0.3692

ullet  $\lambda=0.1$  has the smallest  $E_{in}=0.3105$  and  $E_{out}=0.3677$