Machine Learning HW6

B03902089 資工三 林良翰

Descent Methods for Probabilistic SVM

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

  $z_n = w_{SVM}^T \phi(x_n) + b_{SVM}$

  $p_n = \theta(-y_n(Az_n + B))$, where $\theta(s) = \frac{\exp(s)}{1+\exp(s)}$

1.

- Let $s_n = -y_n\left(Az_n + B\right)$

  $\Rightarrow F\left(A,B\right) = \frac{1}{N} \sum_{n=1}^{N} \ln\left(1 + \exp(s_n)\right)$
- $\frac{\partial F(A,B)}{\partial A} = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{1+\exp(s_n)} \exp(s_n) \frac{\partial s_n}{\partial A} = \frac{1}{N} \sum_{n=1}^{N} -p_n y_n z_n$

  $\frac{\partial F(A,B)}{\partial B} = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{1+\exp(s_n)} \exp(s_n) \frac{\partial s_n}{\partial B} = \frac{1}{N} \sum_{n=1}^{N} -p_n y_n$
- $\nabla F(A,B) = \frac{1}{N} \sum_{n=1}^{N} [-p_n y_n z_n, -p_n y_n]^T$

2.

- Definition of Hessian Matrix $H\left(f\right)$ of $f\left(x_1, x_2, \dots, x_n\right)$

$$H\left(f\right) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \frac{\partial^2 f}{\partial x_2 x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n x_1} & \frac{\partial^2 f}{\partial x_n x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

- Let $s_n = -y_n\left(Az_n + B\right)$

$$\begin{aligned} \frac{\partial - p_n y_n z_n}{\partial A} &= \frac{\partial - \theta(\partial s_n) y_n z_n}{\partial s_n} \frac{\partial s_n}{\partial A} \\ &= \frac{e^{s_n}}{1 + e^{s_n}} \left(1 - \frac{e^{s_n}}{1 + e^{s_n}}\right) (y_n z_n)^2 \\ &= (y_n z_n)^2 p_n (1 - p_n) \end{aligned}$$

$$
\begin{aligned}
\frac{\partial - p_n y_n z_n}{\partial B} &= \frac{\partial - \theta\left(\partial s_n\right) y_n z_n}{\partial s_n} \frac{\partial s_n}{\partial B} \\
&= \frac{e^{s_n}}{1+e^{s_n}}\left(1 - \frac{e^{s_n}}{1+e^{s_n}}\right)\left(y_n^2 z_n\right) \\
&= \left(y_n^2 z_n\right) p_n \left(1 - p_n\right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial - p_n y_n}{\partial B} &= \frac{\partial - \theta\left(\partial s_n\right) y_n}{\partial s_n} \frac{\partial s_n}{\partial B} \\
&= \frac{e^{s_n}}{1+e^{s_n}}\left(1 - \frac{e^{s_n}}{1+e^{s_n}}\right)\left(y_n\right)^2 \\
&= \left(y_n\right)^2 p_n \left(1 - p_n\right)
\end{aligned}
$$

- $H(F(A,B)) = \frac{1}{N}\sum_{n=1}^{N}\begin{bmatrix} \left(y_n z_n\right)^2 p_n\left(1-p_n\right) & \left(y_n^2 z_n\right) p_n\left(1-p_n\right) \\ \left(y_n^2 z_n\right) p_n\left(1-p_n\right) & \left(y_n\right)^2 p_n\left(1-p_n\right)\end{bmatrix}$

Kernel Ridge Regression

- Guassian Kernel $K\left(x, x'\right) = \exp\left(-\gamma\|x - x'\|^2\right)$
- Kernel Ridge Regression:
  - Want to Minimize: $\min_{\beta} E_{aug}\left(\beta\right) = \min_{\beta} \frac{\lambda}{N}\beta^T K\beta + \frac{1}{N}\left(\beta^T K^T K\beta - 2\beta^T K^T y + y^T y\right)$
  - Solving: $\nabla E_{aug}\left(\beta\right) = \frac{2}{N}K^T\left(\left(\lambda I + K\right)\beta - y\right) = 0$
  - Obtain: $\beta = \left(\lambda I + K\right)^{-1} y$

3.

- $\gamma \to \infty$
- $\lim_{\gamma\to\infty} K\left(x, x'\right) = \lim_{\gamma\to\infty} e^{-\gamma\|x-x'\|^2} = I$
- $\beta = \left(\lambda I + I\right)^{-1} y$

4.

- $\gamma \to 0$
- $\lim_{\gamma\to 0} K\left(x, x'\right) = \lim_{\gamma\to 0} e^{-\gamma\|x-x'\|^2} = J$
  J is the matrix of all ones.
- $\beta = \left(\lambda I + J\right)^{-1} y$

Support Vector Regression

- $\left(P_2\right) \min_{b,w,\xi^{\vee},\xi^{\wedge}} \frac{1}{2}w^T w + C\sum_{n=1}^{N}\left(\left(\xi_n^{\vee}\right)^2 + \left(\xi_n^{\wedge}\right)^2\right)$
  s.t. $-\epsilon - \xi_n^{\vee} \le y_n - w^T\phi\left(x_n\right) - b \le \epsilon + \xi_n^{\wedge}$

5.

- Let $A_n = y_n - w^T\phi\left(x_n\right) - b$
- Lagrange Multiplier Method:

$$\mathcal{L}\left(P_2\right) \min_{b,w,\xi^\vee,\xi^\wedge} \max_{\alpha^\vee,\alpha^\wedge} \tfrac{1}{2} w^T w + C \sum_{n=1}^{N} \left(\left(\xi_n^\vee\right)^2 + \left(\xi_n^\wedge\right)^2\right)$$
$$+ \alpha^\vee \sum_{n=1}^{N} \left(A_n + \left(\epsilon + \xi_n^\vee\right)\right) + \alpha^\wedge \sum_{n=1}^{N} \left(A_n - \left(\epsilon + \xi_n^\wedge\right)\right)$$

- Partial derivative on ξ
  $\frac{\partial \mathcal{L}}{\partial \xi_n^\vee} = 2C\xi_n^\vee + \alpha^\vee = 0 \Rightarrow \alpha^\vee = -2C\xi_n^\vee$
  $\frac{\partial \mathcal{L}}{\partial \xi_n^\wedge} = 2C\xi_n^\wedge - \alpha^\wedge = 0 \Rightarrow \alpha^\wedge = 2C\xi_n^\wedge$
  $$\Rightarrow L\left(P_2\right) \min_{b,w,\xi^\vee,\xi^\wedge} \tfrac{1}{2} w^T w - C \sum_{n=1}^{N} \left(\left(\xi_n^\vee\right)^2 + \left(\xi_n^\wedge\right)^2\right)$$
  $$- 2C \sum_{n=1}^{N} \xi_n \left(A_n + \epsilon\right) + 2C \sum_{n=1}^{N} \xi_n^\wedge \left(A_n - \epsilon\right)$$
- Partial derivative on ξ again
  $\frac{\partial \mathcal{L}}{\partial \xi_n^\vee} = -2C\xi_n^\vee - 2C\left(A_n + \epsilon\right) = 0 \Rightarrow \xi_n^\vee = -\left(A_n + \epsilon\right)$ and $A_n \leq -\epsilon$
  $\frac{\partial \mathcal{L}}{\partial \xi_n^\wedge} = -2C\xi_n^\wedge + 2C\left(A_n - \epsilon\right) = 0 \Rightarrow \xi_n^\wedge = \left(A_n - \epsilon\right)$ and $A_n \geq +\epsilon$
  $$\Rightarrow L\left(P_2\right) \min_{b,w} \tfrac{1}{2} w^T w + C \sum_{n=1}^{N} \left(\left[A_n \leq -\epsilon\right]\left(A_n + \epsilon\right)^2 + \left[A_n \geq +\epsilon\right]\left(A_n - \epsilon\right)^2\right)$$
- Transform the above equation into non-linear
  $$\mathcal{L}\left(P_2\right) \min_{b,w} \tfrac{1}{2} w^T w + C \sum_{n=1}^{N} \left(\max\left(0, \left|A_n\right| - \epsilon\right)\right)^2$$
  $$\Rightarrow \mathcal{L}\left(P_2\right) \min_{b,w} \tfrac{1}{2} w^T w + C \sum_{n=1}^{N} \left(\max\left(0, \left|y_n - w^T\phi\left(x_n\right) - b\right| - \epsilon\right)\right)^2$$

6.

- Let $s_n = \sum_{m=1}^{N} \left(\beta_m K\left(x_n, x_m\right) + b\right)$
- From 5.
  $$F\left(b,\beta\right) = \tfrac{1}{2} \sum_{m=1}^{N} \left(\sum_{n=1}^{N} \beta_n \beta_m K\left(x_n, x_m\right)\right) + C \sum_{n=1}^{N} \left(\max\left(0, \left|y_n - w^T\phi\left(x_n\right) - b\right| - \epsilon\right)\right)^2$$
- $\frac{\partial s_n}{\partial \beta_m} = K\left(x_n, x_m\right)$ denote as K
  $$\begin{aligned}
  \frac{\partial F\left(b,\beta\right)}{\partial \beta_m} &= \frac{1}{2}\sum_{n=1}^{N} \beta_n K + C \sum_{n=1}^{N} \left[\left|y_n - s_n\right| \geq \epsilon\right] \frac{\partial\left(\left|y_n - s_n\right| - \epsilon\right)^2}{\partial \beta_m} \\
  &= \frac{1}{2}\sum_{n=1}^{N} \beta_n K + \begin{cases} y_n - s_n \geq 0, C \sum_{n=1}^{N} \left[\left|y_n - s_n\right| \geq \epsilon\right] \frac{\partial\left(y_n - s_n - \epsilon\right)^2}{\partial \beta_m} \\ y_n - s_n \leq 0, C \sum_{n=1}^{N} \left[\left|y_n - s_n\right| \geq \epsilon\right] \frac{\partial\left(s_n - y_n - \epsilon\right)^2}{\partial \beta_m} \end{cases} \\
  &= \frac{1}{2}\sum_{n=1}^{N} \beta_n K + \begin{cases} y_n - s_n \geq 0, 2C \sum_{n=1}^{N} \left[\left|y_n - s_n\right| \geq \epsilon\right]\left(y_n - s_n - \epsilon\right)\left(-K\right) \\ y_n - s_n \leq 0, 2C \sum_{n=1}^{N} \left[\left|y_n - s_n\right| \geq \epsilon\right]\left(s_n - y_n - \epsilon\right) K \end{cases} \\
  &= \frac{1}{2}\sum_{n=1}^{N} \beta_n K + 2C \sum_{n=1}^{N} \left[\left|y_n - s_n\right| \geq \epsilon\right]\left(\left|y_n - s_n\right| - \epsilon\right) sign\left(y_n - s_n\right) K
  \end{aligned}$$

Blending

7.

- 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
  $h(x) = \frac{x_1^2 - x_2^2}{x_1 - x_2}(x - x_1) + x_1^2 = (x_1 + x_2)x - x_1 x_2$
- $\bar{g}(x) = E[h(x)] = E[x_1 + x_2]x + E[x_1 x_2]$
- Since the x value of points are sampled from unifrom distribution over $[0, 1]$
  $\Rightarrow \bar{g}(x) = E[x_1 + x_2]x + E[x_1 x_2] = x - \frac{1}{4}$

Test Set Linear Regression

8.

- Define a cheating hypothesis.
  $g_i(x_j) = [i = j]$, 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.
  $g_0(x_j) = 0$, where $1 \leq j \leq N$.
- Construct a series of cheating hypothesis
  $[g_0, g_1, g_2, \ldots, g_{n-2}, g_{n-1}]$
- Query RMSE for N times to obtain $RMSE(g_i)$, where $0 \leq i \leq N - 1$
- Now we can compute every $\tilde{y}_i$
  $\tilde{y}_i = \frac{1}{2}\left(N\left([RMSE(g_0)]^2 - [RMSE(g_i)]^2\right) + 1\right)$
- $\tilde{y}_n$ can be computed from all the other $\tilde{y}_i$ with $RMSE(g_0)$
  Thus we need N queries.

9.

- Continue from 8., we use g_0 again
  $g_0(x_j) = 0$, where $1 \leq j \leq N$.
- List out two equations below
  $[RMSE(g_0)]^2 = \frac{1}{N}\sum_{i=1}^{N}(\tilde{y}_i)^2 = \frac{1}{N}\tilde{y}^T\tilde{y}$
  $[RMSE(g)]^2 = \frac{1}{N}\sum_{i=1}^{N}(\tilde{y}_i - g(x_i))^2 = \frac{1}{N}\left(\tilde{y}^T\tilde{y} - 2g^T\tilde{y} + g^T g\right)$
- We can obtain $g^T\tilde{y}$ by the equations above
  $g^T\tilde{y} = \frac{1}{2}\left(N\left([RMSE(g_0)]^2 - [RMSE(g)]^2\right) + g^T g\right)$
- Thus we only need 2 queries.

10.

- Continue from 8. 9., we use g_0, $g^T\tilde{y}$ again
  $g^T\tilde{y} = \frac{1}{2}\left(N\left([RMSE(g_0)]^2 - [RMSE(g)]^2\right) + g^T g\right)$

- The problem is to obtain optimal $[\alpha_1, \alpha_2, \ldots, \alpha_K]$ for
$$\min_{\alpha_1,\alpha_2,\ldots,\alpha_K} RMSE\left(\sum_{k=1}^{K} \alpha_k g_k\right)$$
$$RMSE\left(\sum_{k=1}^{K} \alpha_k g_k\right) = \frac{1}{N}\left(\tilde{y}^T\tilde{y} - 2\left(\sum_{k=1}^{K} \alpha_k g_k\right)^T \tilde{y} + \left(\sum_{k=1}^{K} \alpha_k g_k\right)^T \left(\sum_{k=1}^{K} \alpha_k g_k\right)\right)$$
- Partial derivative by α_s, where $1 \le s \le K$
$$\frac{\partial RMSE\left(\sum_{k=1}^{K} \alpha_k g_k\right)}{\partial \alpha_s} = -2(g_s)^T\tilde{y} + 2(g_s)^T \sum_{k=1}^{K} \alpha_k g_k = 0$$
From 9., we can calculate $(g_s)^T\tilde{y}$
- We can obtain K equations to solve all α
$\Rightarrow$ Require $K+1$ queries.

Experiment with Kernel Ridge Regression

11.

- 400 training data, 100 testing data.
- 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

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

12.

- 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

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

Experiment with Support Vector Regression

13.

- 400 training data, 100 testing data.
- 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

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

14.

- 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

- $(\lambda, \gamma) = (1, 0.125)$ 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.
- 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

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