cs446 hw2

Junsheng Huang

February 2024

1 Soft-margin SVM

consider Lagrangian function $L(\boldsymbol{w}, \boldsymbol{\varepsilon}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ and KKT constrain:

$$\begin{cases} 1 - \varepsilon_i - y_i(\boldsymbol{w_T x_i}) \leq 0 \\ -\varepsilon_i \leq 0 \\ i = 1, 2, 3...n \end{cases}$$

$$L(\boldsymbol{w}, \boldsymbol{\varepsilon}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum \varepsilon_i + \sum \alpha_i[1 - \varepsilon_i - y_i(\boldsymbol{w^T x_i})] - \sum \beta_i \varepsilon_i \quad (1)$$

We need to $\min_{\boldsymbol{w},\boldsymbol{\varepsilon}} max_{\boldsymbol{\alpha},\boldsymbol{\beta}} L(\boldsymbol{w}, \boldsymbol{\varepsilon}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, which is equal to dual form $max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \min_{\boldsymbol{w},\boldsymbol{\varepsilon}} L(\boldsymbol{w}, \boldsymbol{\varepsilon}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. First consider $\min_{\boldsymbol{w},\boldsymbol{\varepsilon}} L(\boldsymbol{w}, \boldsymbol{\varepsilon}, \boldsymbol{\alpha}, \boldsymbol{\beta})$:

$$\begin{cases} \dfrac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum \alpha_i y_i \boldsymbol{x_i} \\ \dfrac{\partial L}{\partial \boldsymbol{\varepsilon_i}} = C - \alpha_i - \beta_i \end{cases}$$

so:

$$L(\boldsymbol{w}, \boldsymbol{\varepsilon}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}(\sum \alpha^i y_i \boldsymbol{x_i})^2 + C\sum \varepsilon_i + \sum \alpha_i \varepsilon_i - \sum \beta_i \varepsilon_i - \sum \alpha_i y_i(\boldsymbol{w^T x_i}) \quad (2)$$

$$= \sum \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i^T x_j} \quad (3)$$

In conclusion, the dual form is: $max_\alpha \sum \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \boldsymbol{x_i^T x_j}, 0 \leq \alpha_i \leq C, \sum_{i=1}^N \alpha_i y_i = 0$

2 SVM,RBF Kernel and Nearest Neighbor

2.1

the prediction is: $f(x) = \boldsymbol{\hat{w}^T x} = (\sum \hat{\alpha}_i y_i \boldsymbol{x_i})^T \boldsymbol{x}$

2.2

the prediction is:

$$\begin{aligned} f_\sigma(x) &= \boldsymbol{\hat{w}^T x} && (4)\\ &= (\sum \hat{\alpha}_i y_i \phi(\boldsymbol{x_i}))^T \phi(\boldsymbol{x}) && (5)\\ &= \sum \hat{\alpha}_i y_i \kappa(\boldsymbol{x_i}, \boldsymbol{x}) && (6)\\ &= \sum \hat{\alpha}_i y_i exp(-\frac{\|\boldsymbol{x_i} - \boldsymbol{x_j}\|_2^2}{2\sigma^2}) && (7) \end{aligned}$$

2.3

$$\frac{f_\sigma(x)}{exp(\frac{-\rho^2}{2\sigma^2})} = \frac{\sum_{i \in S} \hat{\alpha}_i y_i exp(\frac{-\|\boldsymbol{x_i}-\boldsymbol{x_j}\|_2^2}{2\sigma^2})}{exp(\frac{-\rho^2}{2\sigma^2})} \tag{8}$$

consider the sum into two parts: T and $S \setminus T$
for $i \in T$: $\|\boldsymbol{x_i} - \boldsymbol{x_j}\|_2^2 = \rho^2$, so we have:

$$\frac{\sum_{i \in T} \hat{\alpha}_i y_i exp(\frac{-\|\boldsymbol{x_i}-\boldsymbol{x_j}\|_2^2}{2\sigma^2})}{exp(\frac{-\rho^2}{2\sigma^2})} = \sum_{i \in T} \hat{\alpha}_i y_i \tag{9}$$

for $i \in S \setminus T$, we have:

$$\frac{\sum_{i \in S \setminus T} \hat{\alpha}_i y_i exp(\frac{-\|\boldsymbol{x_i}-\boldsymbol{x_j}\|_2^2}{2\sigma^2})}{exp(\frac{-\rho^2}{2\sigma^2})} = \sum_{i \in S \setminus T} \hat{\alpha}_i y_i exp(\frac{\rho^2 - \|\boldsymbol{x_i}-\boldsymbol{x_j}\|_2^2}{2\sigma^2}) \tag{10}$$

since $\rho^2 - \|\boldsymbol{x_i} - \boldsymbol{x_j}\|_2^2 \le 0$,

$$\lim_{\sigma \to 0} exp(\frac{\rho^2 - \|\boldsymbol{x_i}-\boldsymbol{x_j}\|_2^2}{2\sigma^2}) = 0 \tag{11}$$

So we have:

$$\lim_{\sigma \to 0} \frac{f_\sigma(\boldsymbol{x})}{exp(\frac{-\rho^2}{2\sigma^2})} = \sum_{i \in T} \hat{\alpha}_i y_i \tag{12}$$

3 Decision Tree and Adaboost

3.1

the sample entropy of D is $I(D) = -\sum p(c|D)log(p(x|D)) = -(\frac{1}{2}log(\frac{1}{2})*2) = 1$

3.2

the rule for the split is if $x_1 \geq 5$, label is 1, else -1 ,the maximum information gain is:
$IG(D, f) = 1 - I(D|f) = 1 - \frac{2}{3}(-\frac{3}{4}log(\frac{3}{4}) - \frac{1}{4}log(\frac{1}{4})) - \frac{1}{3}(-log(1)) = 0.46$

3.3

We further divide the child node $x_1 < 5$, the rule is if $x_2 \geq 2$, label is -1, else 1, the maximum information gain is:
$IG(D, f) = I(D) - I(D|f) = -\frac{3}{4}log(\frac{3}{4}) - \frac{1}{4}log(\frac{1}{4}) - \frac{3}{4}(-log(1)) - \frac{1}{4}(-log(1)) = 0.81$

3.4

when t = 1:
$\gamma_1^{(i)} = \frac{1}{6}$ for $i = 1, 2...6$
$f_1(\boldsymbol{x}^{(i)}) = 1$ if $x_1^{(i)} \geq 5$, else $f_1(\boldsymbol{x}^{(i)}) = -1$ That means, $f_1(\boldsymbol{x}) = sign(x_1 - 5)$
$\epsilon_1 = \sum_{i=1}^{6} \gamma_1^i y^{(i)} f_1(\boldsymbol{x}^{(i)}) = \frac{1}{6}(5-1) = \frac{2}{3}$
$\alpha_1 = \frac{1}{2}ln(\frac{1+\epsilon_1}{1-\epsilon_1}) = \frac{1}{2}ln(5)$
when t =2: $\gamma_2^{(i)} = \frac{1}{6}exp(-\frac{1}{2}ln(5))$ for $i = 1, 3, 4, 5, 6$ and $\gamma_2^{(i)} = \frac{1}{6}exp(\frac{1}{2}ln(5))$ for $i = 2$, after normalization, it would be:
$\gamma_2^{(i)} = \frac{1}{10}$ for $i = 1, 3, 4, 5, 6$ and $\gamma_2^{(i)} = \frac{1}{2}$ for $i = 2$
$f_2(\boldsymbol{x}^{(i)}) = 1$ if $x_1^{(i)} \geq 2$, else $f_2(\boldsymbol{x}^{(i)}) = -1$ That means, $f_2(\boldsymbol{x}) = sign(x_1 - 2)$
$\epsilon_2 = \sum_{i=1}^{6} \gamma_1^i y^{(i)} f_1(\boldsymbol{x}^{(i)}) = \frac{1}{2} + \frac{3}{10} - \frac{2}{10} = \frac{3}{5}$
$\alpha_2 = \frac{1}{2}ln(\frac{1+\epsilon_2}{1-\epsilon_2}) = ln(2)$

3.5

the rule of classifier is:
$F_T(\boldsymbol{x}) = sign(\alpha_1 f_1(\boldsymbol{x}) + \alpha_2 f_2(\boldsymbol{x})) = sign(\frac{1}{2}ln(5)sign(x_1-5) + ln(2)sign(x_1-2))$
Verify for each case:
$F_T(\boldsymbol{x}^{(1)}) = sign(\frac{1}{2}ln(5)(-1) + ln2(-1)) = -1$ (correct)
$F_T(\boldsymbol{x}^{(2)}) = sign(\frac{1}{2}ln(5)(-1) + ln2(1)) = -1$ (wrong)

$F_T(\boldsymbol{x}^{(3)}) = sign(\frac{1}{2}ln(5)(-1) + ln2(-1)) = -1$ (correct)
$F_T(\boldsymbol{x}^{(4)}) = sign(\frac{1}{2}ln(5)(-1) + ln2(-1)) = -1$ (correct)
$F_T(\boldsymbol{x}^{(5)}) = sign(\frac{1}{2}ln(5)(1) + ln2(-1)) = 1$ (correct)
$F_T(\boldsymbol{x}^{(6)}) = sign(\frac{1}{2}ln(5)(1) + ln2(-1)) = 1$ (correct)

4 Learning Theory

4.1

we have probability of no less than $1 - \delta$ to have $|p - \hat{p}| \leq \sqrt{\frac{1og(\frac{2}{\delta})}{2n}}$, and we have $\delta = 0.05$, so $\sqrt{\frac{ln(40)}{2n}} \leq 0.05$, equals to $n > 737.7$, so at least 738 samples are needed.

4.2

4.2.1

$$VC(\mathcal{F}_{affine}) = 2 \tag{13}$$

This is because when $VCdim = 2$, consider $(1,1)(1,0)(0,1)(0,0)$, they can be scattered by finding the line to intersect the x-axis with the point. However, for $VCdim = 3$, consider $(1,0,1)$: it can't be scatted by a line because a line can only have one intersection with the x-axis, so can't divide three parts out.

4.2.2

$$VC(\mathcal{F}^k_{affine}) = k + 1 \tag{14}$$

consider $\boldsymbol{w^T x} + w_0 = \begin{bmatrix} \boldsymbol{x}^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{w} \\ w_0 \end{bmatrix}$ and all data point as $X = \begin{bmatrix} \boldsymbol{x}_1^T & 1 \\ \boldsymbol{x}_2^T & 1 \\ ... & \\ \boldsymbol{x}_n^T & 1 \end{bmatrix}$

Now, we can consider equation $X \begin{bmatrix} \boldsymbol{w} \\ w_0 \end{bmatrix} = \boldsymbol{y}$c

for $VCdim = k + 1$, consider X is full-ranked, this equation is always solvable and thus exist $\boldsymbol{w}^T$,w_0 to scatter the data point.c

for $VCdim = k + 2$,there always exist $\boldsymbol{y}$ such that rank$[X, \boldsymbol{y}]$ ¿ rank[X], so the equation for this y is unsolvable, thus not exist$\boldsymbol{w}^T$,w_0 to scatter the data point.

Here is an example:
$\boldsymbol{x}_j = \sum_{i\neq j} a_i \boldsymbol{x}_i$, $y_i = sign(a_i)$, $y_j = -1$, and we have $y_i = sign(\boldsymbol{w}^T\boldsymbol{x}_i) = sign(a_i)$, $y_j = sign(\boldsymbol{w}^T\boldsymbol{x}_j) = \sum_{i\neq j} a_i \boldsymbol{w}^T\boldsymbol{x}_i = 1$, which is contradict.

4.2.3

$$VC(\mathcal{F}_{cos}) = \infty \tag{15}$$

consider data set $\mathcal{D} = \left\{x_i = \frac{3\pi}{4}8^i\right\}_{i=1}^n$ for $\forall S \in \mathbf{D}$, we can always find predictor $\mathbf{F}_{cos} = \{\mathbf{1}\{cos(cx) > 0\}\}$ with $c = \sum_{i:y_i=-1} 8^{-i}$
for any point $x_j = \frac{3\pi}{4}8^i$ with $y_i = -1$, we have:

$$\begin{aligned} cx_j &= \frac{3\pi}{4}8^i \sum_{i:y_i=-1} 8^{-i} && (16)\\ &= \frac{3\pi}{4} + \frac{3\pi}{4}(\sum_{i<j} 8^{j-i} + \sum_{i>j} 8^{j-i}) && (17)\end{aligned}$$

for $i < j$ part, the value would be $2n\pi$; for $i > j$ part, the value would be $[0, \frac{3\pi}{4})$(consider the sum of geometric sequence).
Thus the value would be $[\frac{3\pi}{4} + 2n\pi, \frac{3\pi}{2} + 2n\pi)$, and $cos(cx_j) < 0$, $\mathcal{F}_{cos}(x_j) = -1$
for any point $x_j = \frac{3\pi}{4}8^i$ with $y_i = 1$, we have:

$$\begin{aligned} cx_j &= \frac{3\pi}{4}8^i \sum_{i:y_i=-1} 8^{-i} && (18)\\ &= \frac{3\pi}{4}(\sum_{i<j} 8^{j-i} + \sum_{i>j} 8^{j-i}) && (19)\end{aligned}$$

for $i < j$ part, the value would be $2n\pi$; for $i > j$ part, the value would be $[0, \frac{3\pi}{16})$.(consider the sum of geometric sequence)
Thus the value would be $[2n\pi, \frac{3\pi}{16} + 2n\pi)$, and $cos(cx_j) \geq 0$, $\mathcal{F}_{cos}(x_j) = 1$
Since that, we prove there exists predictor that can satisfy data set with $VCdim = n$, so $VCdim = \infty$.

5 Coding: SVM

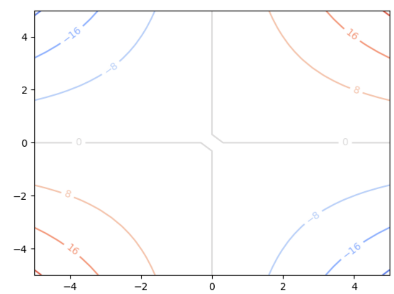

Figure 1: polynomial kernel with degree 2

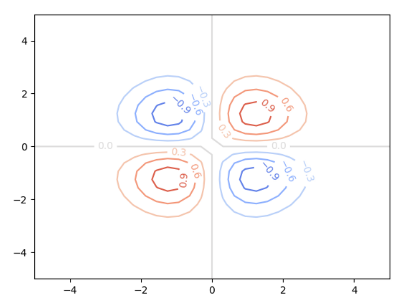

Figure 2: RBF kernel with $\sigma = 1$

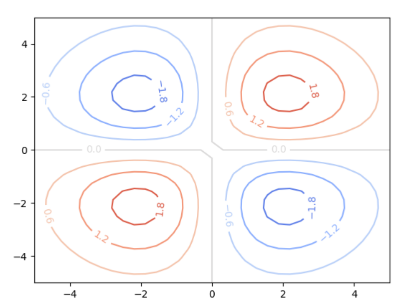

Figure 3: RBF kernel with $\sigma = 2$

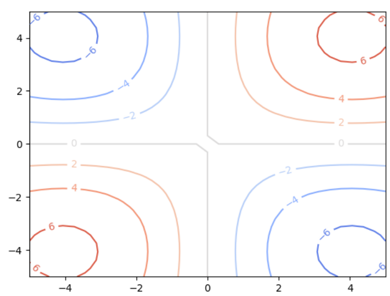

Figure 4: RBF kernel with $\sigma = 4$