Homework 2

Fondations of Machine Learning

Xiang Pan (xp2030)

March 22, 2022

VC Dimension

1

(a)

Show that there exists a set of $n+1$ points in $\mathbb{R}^n$ that can be shattered by $\mathcal{B}_n$. Conclude that VCdim $(\mathcal{B}_n) \geq n+1$

Set condition: We have n+1 points can determine a unique ball in $\mathbb{R}^n$, and we have the ball center c, radius r.

For those negative points $p_n \in P_N$, we have the radius vector from p_n to c, thus we can constuct the new $p'_n = p_n + \delta(c - p_n)$, $\delta > 0$. $p'_n \in P'_N$, P'_N is the set of p'_n.

Thus we can constuct a new ball with new n+1 points set $P'_N \bigcup P_P$. For the new ball B (c', r'), the negative points distance $r > r'$, the positive points $r' \leq r'$. Thus B (c', r') can shatter n+1 points in $\mathbb{R}^n$.

To prove $r > r'$, the B (c,r) and B (c',r') are joint, since P_P are in the B (c,r) and B (c',r'). Note that, P'_N are strictly inside B (c,r), it is easy to check that P_N are strictly outside B (c',r').

There exists a set of (n+1) points in R^n that can be shattered by B(n). We can conclude that VCdim $(\mathcal{B}_n) \geq n+1$.

(b)

Let $B(c,r)$ be the ball of radius r centered at $c \in \mathbb{R}^n$. Then $x \in B(c,r)$

$$\|x - c\|^2 \leq r \tag{1}$$

$$\sum_{i=1}^{n} \|x_i\|^2 - 2\sum_{i=1}^{n} c_i x_i + \sum_{i=1}^{n} c_i^2 - r \leq 0 \tag{2}$$

We can find a hyperplane h that is orthogonal to $B(c,r)$ and x is in $B(c,r)$ if $h \cdot x' + b \leq 0$.

$$h = \begin{bmatrix} 1, \\ -2c_1, \\ -2c_2, \\ . \\ -2c_n \end{bmatrix} \tag{3}$$

$$x' = \begin{bmatrix} \sum_{i=1}^{n} \|x_i\|^2 \\ x_1 \\ x_2 \\ . \\ x_n \end{bmatrix} \tag{4}$$

$$b = \sum c_i^2 - r \tag{5}$$

The VC dimension of $B(c, r)$ is at most as the same as the VC dimension of hyperplane R^{n+1}, which is n+2.

(c)

Show that VCdim$(\mathcal{B}_2) = 3$.

We have already know that

$$3 \leq (\mathcal{B}_2) \leq 4 \tag{6}$$

But for four points, we can use three points $y_1 = 1, y_2 = 1, y_3 = 1$ to uniquely determine a ball b_0 in $\mathbb{R}^2$, we can make the fourth point within the ball, and the fourth point is the center of the ball, and the $y_4 = -1$. Any ball can cover $p1, p2, p3$ must cover the minimal ball b_0, thus $p4$ must within any ball can cover $p1, p2, p3$. The VC dimension is at most 3.

This case can not find a ball to shatter it.

Thus we have VCdim $(\mathcal{B}_2) = 3$

Maximum Margin Multiple Kernel

1

(a)

$$\widehat{\mathcal{M}}_q = \left\{ \boldsymbol{\mu} : \boldsymbol{\mu} \in \Delta_q, \widehat{\gamma}_{K_\mu} \geq \gamma_0 \right\} \tag{7}$$

Δ_q is the set of μ,

$$\Delta_q = \{\boldsymbol{\mu} : \boldsymbol{\mu} \geq 0, \|\boldsymbol{\mu}\|_q = 1\} \text{ with } q \geq 1 \tag{8}$$

K is the combined kernel function, μ is the weight for each kernel component,

$$K = \sum_{k=1}^{p} \mu_k K_k \tag{9}$$

To explain γ, we need to show where the γ_0 from.
From the paper,

$$\max_{\boldsymbol{\mu} \in \Delta_q} \sum_{i=1}^{m} \min_{y \neq y_i} \boldsymbol{\mu} \cdot \boldsymbol{\eta}\left(x_i, y_i, y\right) \tag{10}$$

which equals to

$$\max_{\boldsymbol{\mu} \in \Delta_q} \frac{1}{m} \sum_{i=1}^{m} \min_{y \neq y_i} \boldsymbol{\mu} \cdot \boldsymbol{\eta}\left(x_i, y_i, y\right) \tag{11}$$

.

Converting the optimization to convex optimization, we have

$$\max_{\substack{\boldsymbol{\mu} \in \Delta_q \\ \gamma}} \sum_{i=1}^{m} \gamma_i \text{ s.t. } \forall i \in [1, m], \forall y \neq y_i, \boldsymbol{\mu} \cdot \boldsymbol{\eta}\left(x_i, y_i, y\right) \geq \gamma_i \tag{12}$$

We can directly solve the optimization problem, however we can convert it to a minimization problem with constraint $\widehat{\gamma}_{K_{\boldsymbol{\mu}}} \geq \gamma_0$.

$$\widehat{\gamma}_{K_\mu} = \frac{1}{m} \sum_{i=1}^{m} \min_{y \neq y_i} \boldsymbol{\mu} \cdot \boldsymbol{\eta}\left(x_i, y_i, y\right) \tag{13}$$

For γ_0, setting it equal to the maximum feasible value will guarantee that the selected $\boldsymbol{\mu}$ is also give us the solution as(12).

(b)

$$\begin{aligned} &\min_{\boldsymbol{\mu} \in \widehat{\mathcal{M}}_q} \min_{\mathbf{w}, \boldsymbol{\xi}} \frac{1}{2} \sum_{y=1}^{c} \sum_{k=1}^{p} \frac{\left\|\mathbf{w}_{y,k}\right\|^2}{\mu_k} + C \sum_{i=1}^{m} \xi_i, \\ &\text{subject to: } \forall i \in [1, m], \xi_i \geq 0, \forall y \neq y_i \\ &\qquad \xi_i \geq 1 - \left(\mathbf{w}_{y_i} \cdot \Phi\left(x_i\right) - \mathbf{w}_y \cdot \Phi\left(x_i\right)\right) \end{aligned} \tag{14}$$

We can trasnform the optimization problem to,

$$\begin{aligned} &\min_{\boldsymbol{\mu} \in \widehat{\mathcal{M}}_q} \min_{\mathbf{w}, \boldsymbol{\xi}} \beta \frac{1}{2} \sum_{y=1}^{c} \left\|\mathbf{w}_y \cdot \frac{1}{\sqrt{\boldsymbol{\mu}}}\right\|^2 + \sum_{i=1}^{m} \xi_i, \\ &\text{subject to: } \forall i \in [1, m], \xi_i \geq 0, \\ &\qquad \left(\mathbf{w}_{y_i} \cdot \Phi\left(x_i\right) - \mathbf{w}_y \cdot \Phi\left(x_i\right)\right) + \delta_{y_i, y} - 1 + \xi_i \geq 0 \end{aligned} \tag{15}$$

$$\beta = \frac{1}{C} \tag{16}$$

$w \in R^{m \times c}$

$\delta_{y_i,y}$ is the indicator function, if $\delta_{y_i,y} = 1$ if $y_i = y$, otherwise it is 0.

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \beta \frac{1}{2} \sum_{y=1}^{c} \left\| \mathbf{w}_y \cdot \frac{1}{\sqrt{\boldsymbol{\mu}}} \right\|^2 + \sum_{i=1}^{m} \xi_i - \sum_{i,y} \alpha_{i,y} \left(\mathbf{w}_{y_i} \cdot \Phi\left(x_i\right) - \mathbf{w}_y \cdot \Phi\left(x_i\right) \right) + \delta_{y_i,y} - 1 + \xi_i \tag{17}$$

Then we can get the KKT,

$$\frac{\partial}{\partial \xi_i} \mathcal{L} = 1 - \sum_{y} \alpha_{i,y} = 0 \quad \Rightarrow \quad \sum_{y} \alpha_{i,y} = 1 \tag{18}$$

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{w}_y} \mathcal{L} &= \beta \boldsymbol{w}_y \cdot \frac{1}{\boldsymbol{\mu}} + \sum_{i} \alpha_{i,y} \Phi(x_i) - \sum_{i, y_i = y} \underbrace{\left(\sum_{q} \alpha_{i,q} \right)}_{=1} \Phi(x_i) \\ &= \beta \boldsymbol{w}_y \cdot \frac{1}{\boldsymbol{\mu}} + \sum_{i} \alpha_{i,y} \Phi(x_i) - \sum_{i} \delta_{y_i y} \Phi(x_i) = 0 \end{aligned} \tag{19}$$

$$\boldsymbol{w}_y = \beta^{-1} \left[\sum_{i} \left(\delta_{y_i,y} - \alpha_{i,y} \right) \Phi(x_i) \right] \cdot \boldsymbol{\mu} \tag{20}$$

By using the KKT(18),

$$\begin{aligned} \mathcal{L}(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}) &= \beta \frac{1}{2} \sum_{y=1}^{c} \left\| \mathbf{w}_y \cdot \frac{1}{\sqrt{\boldsymbol{\mu}}} \right\|^2 + \sum_{i=1}^{m} \xi_i - \sum_{i,y} \alpha_{i,y} \left(\mathbf{w}_{y_i} \cdot \Phi\left(x_i\right) - \mathbf{w}_y \cdot \Phi\left(x_i\right) \right) + \delta_{y_i,y} - 1 + \xi_i \\ &= \beta \frac{1}{2} \sum_{y=1}^{c} \left\| \mathbf{w}_y \cdot \frac{1}{\sqrt{\boldsymbol{\mu}}} \right\|^2 + \sum_{i=1}^{m} \xi_i - \sum_{i,y} \alpha_{i,y} \xi_i + \sum_{i,y} \alpha_{i,y} \delta_{y_i,y} - \sum_{i,y} \alpha_{i,y} \left(\mathbf{w}_{y_i} \cdot \Phi\left(x_i\right) - \mathbf{w}_y \cdot \Phi\left(x_i\right) \right) \\ &= \beta \frac{1}{2} \sum_{y=1}^{c} \left\| \mathbf{w}_y \cdot \frac{1}{\sqrt{\boldsymbol{\mu}}} \right\|^2 + \sum_{i=1}^{m} \xi_i - \sum_{i} \xi_i \underbrace{\sum_{y} \alpha_{i,y}}_{=1} + \sum_{i,y} \alpha_{i,y} \delta_{y_i,y} \\ &- \sum_{i,y} \alpha_{i,y} \left(\mathbf{w}_{y_i} \cdot \Phi\left(x_i\right) - \mathbf{w}_y \cdot \Phi\left(x_i\right) \right) \\ &= \underbrace{\frac{1}{2} \beta \sum_{y=1}^{c} \left\| \mathbf{w}_y \cdot \frac{1}{\sqrt{\boldsymbol{\mu}}} \right\|^2}_{=P1} + \underbrace{\sum_{i,y} \alpha_{i,y} \delta_{y_i,y}}_{=P2} + \underbrace{\sum_{i,y} \alpha_{i,y} \left(\mathbf{w}_y \cdot \Phi\left(x_i\right) \right)}_{=P3} - \underbrace{\sum_{i,y} \alpha_{i,y} \left(\mathbf{w}_{y_i} \cdot \Phi\left(x_i\right) \right)}_{=P4} \end{aligned} \tag{21}$$

$$
\begin{aligned}
P3 &= \sum_{i,y} \alpha_{i,y} \left(\mathbf{w}_y \cdot \Phi\left(x_i\right) \right) && (22)\\
&= \beta^{-1} \sum_{i,y} \alpha_{i,y} \left(\left[\sum_j \left(\delta_{y_j,y} - \alpha_{j,y} \right) \Phi(x_j) \right] \cdot \boldsymbol{\mu} \cdot \Phi\left(x_i\right) \right) && (23)\\
&= \beta^{-1} \sum_{i,y} \alpha_{i,y} \boldsymbol{\mu} \cdot \left(\sum_j \left(\delta_{y_j,y} - \alpha_{j,y} \right) \Phi(x_j) \cdot \Phi\left(x_i\right) \right) && (24)\\
&= \beta^{-1} \sum_{i,y} \alpha_{i,y} \boldsymbol{\mu} \cdot \left(\sum_j \left(\delta_{y_j,y} - \alpha_{j,y} \right) K(x_i, x_j) \right) && (25)\\
&= \beta^{-1} \sum_{i,j} K(x_i, x_j) \boldsymbol{\mu} \cdot \left(\sum_y \alpha_{i,y} \left(\delta_{y_j,y} - \alpha_{j,y} \right) \right) && (26)
\end{aligned}
$$

$$
\begin{aligned}
P4 &= \sum_{i,y} \alpha_{i,y} \left(\mathbf{w}_{y_i} \cdot \Phi\left(x_i\right) \right) && (27)\\
&= \beta^{-1} \sum_{i,y} \alpha_{i,y} \left(\left[\sum_j \left(\delta_{y_j,y_i} - \alpha_{j,y_i} \right) \Phi(x_j) \right] \cdot \boldsymbol{\mu} \cdot \Phi\left(x_i\right) \right) && (28)\\
&= \beta^{-1} \sum_{i,y} \alpha_{i,y} \boldsymbol{\mu} \cdot \left(\sum_j \left(\delta_{y_j,y_i} - \alpha_{j,y_i} \right) \Phi(x_j) \cdot \Phi\left(x_i\right) \right) && (29)\\
&= \beta^{-1} \sum_{i,y} \alpha_{i,y} \boldsymbol{\mu} \cdot \left(\sum_j \left(\delta_{y_j,y_i} - \alpha_{j,y_i} \right) K(x_i, x_j) \right) && (30)\\
&= \beta^{-1} \sum_{i,j} K(x_i, x_j) \boldsymbol{\mu} \cdot \left(\delta_{y_j,y_i} - \alpha_{j,y_i} \right) \underbrace{\left(\sum_y \alpha_{i,y} \right)}_{=1} && (31)\\
&= \beta^{-1} \sum_{i,j} K(x_i, x_j) \boldsymbol{\mu} \sum_y \delta_{y_j,y} (\delta_{y_j,y} - \alpha_{j,y}) && (32)
\end{aligned}
$$

$$P1 = \beta \frac{1}{2} \sum_{y=1}^{c} \left\| \mathbf{w}_y \cdot \frac{1}{\sqrt{\boldsymbol{\mu}}} \right\|^2 \tag{33}$$

$$= \beta \frac{1}{2} \sum_{y=1}^{c} (\mathbf{w}_y \cdot \frac{1}{\sqrt{\boldsymbol{\mu}}}) \cdot (\mathbf{w}_y \cdot \frac{1}{\sqrt{\boldsymbol{\mu}}}) \tag{34}$$

$$= \beta^{-1} \frac{1}{2} \sum_{y=1}^{c} \left[\sum_i \left(\delta_{y_i,y} - \alpha_{i,y}\right) \Phi(x_i) \right] \cdot \left[\sum_j \left(\delta_{y_j,y} - \alpha_{j,y}\right) \Phi(x_j) \right] \tag{35}$$

$$= \frac{1}{2} \beta^{-1} \sum_{i,j} K(x_i, x_j) \sum_y (\delta_{y_i,y} - \alpha_{i,y})(\delta_{y_j,y} - \alpha_{j,y}) \tag{36}$$

$$P3 - P4 = \beta^{-1} \sum_{i,j} K(x_i, x_j) \boldsymbol{\mu} \sum_y \alpha_{i,y} \left(\delta_{y_j,y} - \alpha_{j,y}\right) \tag{37}$$

$$- \beta^{-1} \sum_{i,j} K(x_i, x_j) \boldsymbol{\mu} \sum_y \delta_{y_j,y} (\delta_{y_j,y} - \alpha_{j,y}) \tag{38}$$

$$= -\beta^{-1} \sum_{i,j} K(x_i, x_j) \sum_y (\delta_{y_i,y} - \alpha_{i,y})(\delta_{y_j,y} - \alpha_{j,y}) \tag{39}$$

$$\begin{aligned}
\mathcal{L}(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}) &= \underbrace{\frac{1}{2} \sum_{y=1}^{c} \left\| \mathbf{w}_y \cdot \frac{1}{\sqrt{\boldsymbol{\mu}}} \right\|^2}_{=P1} + \underbrace{\sum_{i,y} \alpha_{i,y} \delta_{y_i,y}}_{=P2} + \underbrace{\sum_{i,y} \alpha_{i,y} \left(\mathbf{w}_y \cdot \Phi\left(x_i\right)\right)}_{=P3} - \underbrace{\sum_{i,y} \alpha_{i,y} \left(\mathbf{w}_{y_i} \cdot \Phi\left(x_i\right)\right)}_{=P4} \\
&= \frac{1}{2} \beta^{-1} \sum_{i,j} K(x_i, x_j) \sum_y (\delta_{y_i,y} - \alpha_{i,y})(\delta_{y_j,y} - \alpha_{j,y}) \boldsymbol{\mu} \\
&\quad - \beta^{-1} \sum_{i,j} K(x_i, x_j) \sum_y (\delta_{y_i,y} - \alpha_{i,y})(\delta_{y_j,y} - \alpha_{j,y}) \boldsymbol{\mu} \\
&\quad + \sum_{i,y} \alpha_{i,y} \delta_{y_i,y} \\
&= -\frac{1}{2} \beta^{-1} \sum_{i,j} K(x_i, x_j) \sum_y (\delta_{y_i,y} - \alpha_{i,y})(\delta_{y_j,y} - \alpha_{j,y}) \boldsymbol{\mu} \\
&\quad + \sum_{i,y} \alpha_{i,y} \delta_{y_i,y} \\
&= -\frac{1}{2} \beta^{-1} \sum_{i,j} K(x_i, x_j) \sum_y (\delta_{y_i,y} - \alpha_{i,y})(\delta_{y_j,y} - \alpha_{j,y}) \boldsymbol{\mu} + \sum_{i,y} \alpha_{i,y} \delta_{y_i,y} \\
&= -\frac{C}{2} \sum_{i,j} K(x_i, x_j) \sum_y (\delta_{y_i,y} - \alpha_{i,y})(\delta_{y_j,y} - \alpha_{j,y}) \boldsymbol{\mu} + \sum_i \alpha_i \cdot \mathbf{1}_{y_i} \qquad (\text{Using } C = \frac{1}{\beta})
\end{aligned} \tag{40}$$

$$\boldsymbol{w}_y = \beta^{-1}\left[\sum_i \left(\delta_{y_i,y} - \alpha_{i,y}\right)\Phi(x_i)\right]\cdot\boldsymbol{\mu} \tag{41}$$

$$= \beta^{-1}\left[\sum_{i:y_i=y}\left(1-\alpha_{i,y}\right)\Phi(x_i) + \sum_{i:y_i\neq y}\left(-\alpha_{i,y}\right)\Phi(x_i)\right]\boldsymbol{\mu} \tag{42}$$

Thus we get the following dual problem,

$$\begin{aligned}
&\min_{\boldsymbol{\mu}\in\widehat{M}_q}\max_{\boldsymbol{\alpha}\in\mathbb{R}^{m\times c}} -\frac{C}{2}\sum_{i,j}K(x_i,x_j)\sum_y(\delta_{y_i,y}-\alpha_{i,y})(\delta_{y_j,y}-\alpha_{j,y})\boldsymbol{\mu} + \sum_i \alpha_i\cdot\mathbf{1}_{y_i}\\
&\text{subject to: } \forall i\in[1,m], \boldsymbol{\alpha}_i \geq 0 \wedge \boldsymbol{\alpha}_i\cdot\mathbf{1} = 1
\end{aligned} \tag{43}$$

L is concave in $\boldsymbol{\alpha}$.We define the new dual variable $\hat{\boldsymbol{\alpha}}$ as $\boldsymbol{\delta}-\boldsymbol{\alpha}$:

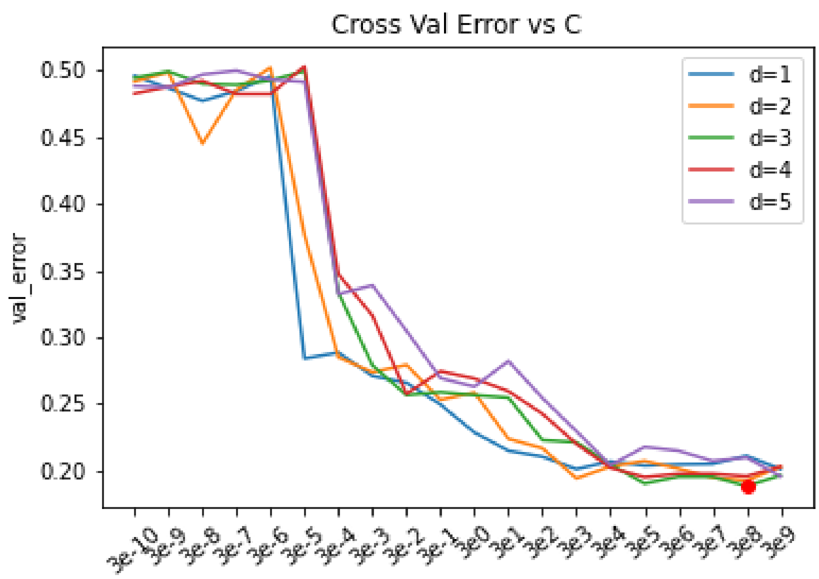
$$\hat{\alpha}_{i,y} = \delta_{y_i,y} - \alpha_{i,y} \tag{44}$$

$$\begin{aligned}
&\min_{\boldsymbol{\mu}\in\widehat{M}_q}\max_{\hat{\boldsymbol{\alpha}}\in\mathbb{R}^{m\times c}} \sum_{i=1}^m \hat{\boldsymbol{\alpha}}_i\cdot\mathbf{e}_{y_i} - \frac{C}{2}\sum_{i,j=1}^m\left(\hat{\boldsymbol{\alpha}}_i\cdot\hat{\boldsymbol{\alpha}}_j\right)\sum_{k=1}^p \mu_k K_k\left(x_i,x_j\right)\\
&\text{subject to: } \forall i\in[1,m], \hat{\boldsymbol{\alpha}}_i \leq \mathbf{e}_{y_i} \wedge \hat{\boldsymbol{\alpha}}_i\cdot\mathbf{1} = 0
\end{aligned} \tag{45}$$

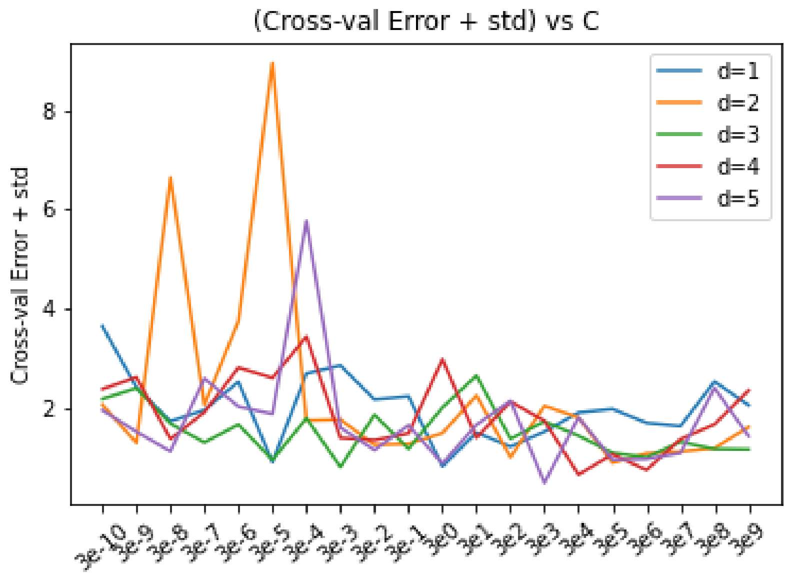
SVMs hand-on

The code is available at `https://github.com/Xiang-Pan/NYU_FML/blob/master/HW2/c.ipynb`.

3



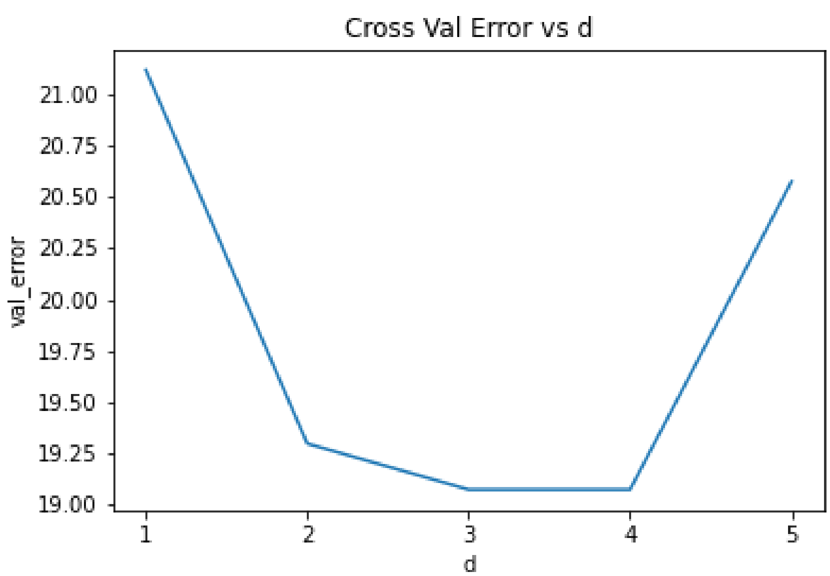
The red point is the optimizal combination. The best pair is d = 3, and C = 6561.0. (C*, d*) = (6561.0, 3). The error is 18.8197%.



(Cross-val Error + std) vs C
Cross-val Error + std
d=1
d=2
d=3
d=4
d=5
3e-10
3e-9
3e-8
3e-7
3e-6
3e-5
3e-4
3e-3
3e-2
3e-1
3e0
3e1
3e2
3e3
3e4
3e5
3e6
3e7
3e8
3e9

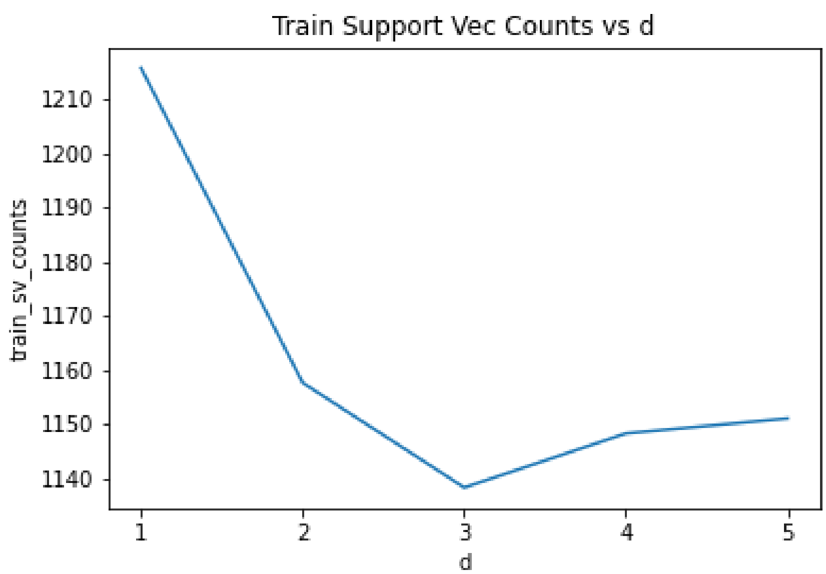
4

$(C^*, d^*) = (6561.0, 3)$.

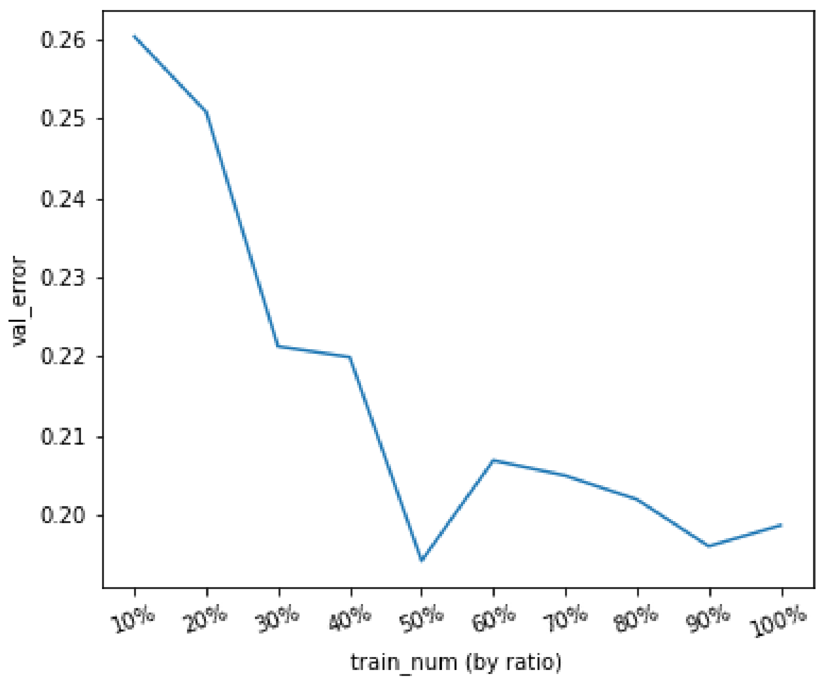


Cross Val Error vs d
val_error
d

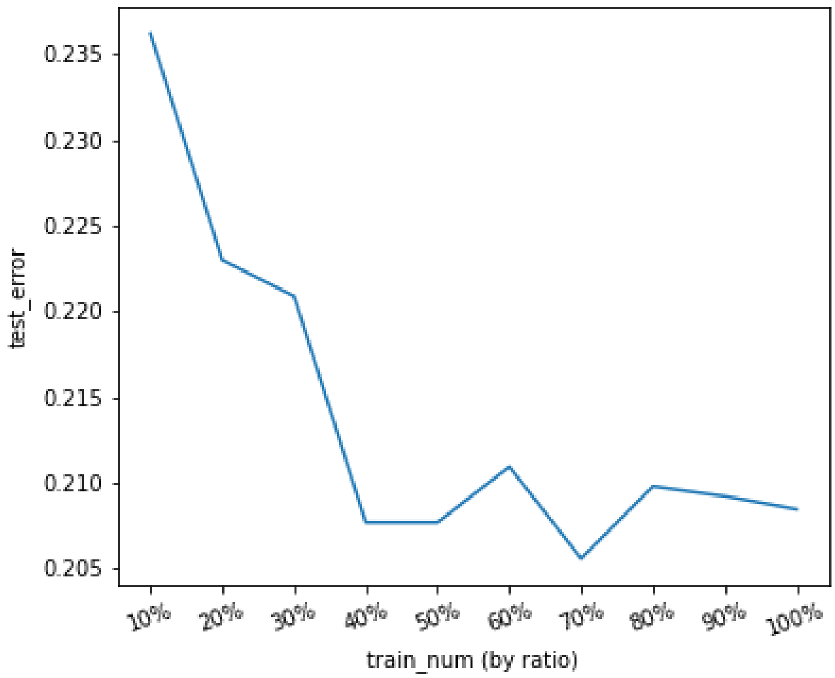

Test Error vs d
test_error
22.50
22.25
22.00
21.75
21.50
21.25
21.00
1
2
3
4
5
d


Train Support Vec Counts vs d
train_sv_counts
1210
1200
1190
1180
1170
1160
1150
1140
1
2
3
4
5
d

5



0.26
0.25
0.24
0.23
0.22
0.21
0.20
val_error
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
train_num (by ratio)



6

(a)

$$\begin{aligned} &\min_{\boldsymbol{\alpha}, b, \boldsymbol{\xi}} \frac{1}{2} \sum_{i=1}^{m} |\alpha_i| + C \sum_{i=1}^{m} \xi_i \\ \text{subject to } & y_i \left(\sum_{j=1}^{m} \alpha_j y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) + b \right) \geq 1 - \xi_i, i \in [1, m] \\ & \xi_i, \alpha_i \geq 0, i \in [1, m]. \end{aligned} \tag{46}$$

We have the Lagrangian function:

$$\mathcal{L}(\boldsymbol{\alpha}, b, \boldsymbol{\xi}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2} \sum_{i=1}^{m} |\alpha_i| + C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \delta_i \left(y_i \left(\sum_{j=1}^{m} \alpha_j y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) + b \right) - 1 + \xi_i \right) \tag{47}$$

$$- \sum_{i=1}^{m} \beta_i \xi_i - \sum_{i=1}^{m} \gamma_i \alpha_i \tag{48}$$

The KKT conditions are obtained by setting the gradient of the Lagrangian with respect to the primal variables $\boldsymbol{\alpha}$, b, $\boldsymbol{\xi}$ to zero:

$$\begin{aligned} \nabla_{\alpha_j}\mathcal{L} = \frac{1}{2}sign(\alpha_j) - \sum_{i=1}^{m}\delta_i y_i \left(y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\right) - \gamma_j = 0 \\ \implies \quad \frac{1}{2}sign(\alpha_j) = \sum_{i=1}^{m}\delta_i y_i y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + \gamma_j \end{aligned} \tag{49}$$

$$\nabla_b\mathcal{L} = -\sum_{i=1}^{m}\delta_i y_i = 0 \quad \implies \quad \sum_{i=1}^{m}\delta_i y_i = 0 \tag{50}$$

$$= -\frac{C}{2}\sum_{i,j}K(x_i, x_j)\sum_{y}(\delta_{y_i,y} - \alpha_{i,y})(\delta_{y_j,y} - \alpha_{j,y})\boldsymbol{\mu} + \sum_{i,y}\alpha_{i,y}\delta_{y_i,y} \tag{51}$$

$$\nabla_{\xi_i}\mathcal{L} = C - \delta_i - \beta_i = 0 \quad \implies \quad \delta_i + \beta_i = C \tag{52}$$

$$\begin{aligned} \forall i, \delta_i\left(y_i\left(\sum_{j=1}^{m}\alpha_j y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + b\right) - 1 + \xi_i\right) = 0 \\ \implies \quad \delta_i = 0 \vee y_i\left(\sum_{j=1}^{m}\alpha_j y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + b\right) = 1 - \xi_i \end{aligned} \tag{53}$$

$$\forall i, \beta_i\xi_i = 0 \quad \implies \quad \beta_i = 0 \vee \xi_i = 0 \tag{54}$$

$$\Leftrightarrow \quad \delta_i = C \vee \xi_i = 0 \tag{55}$$

To derive the dual form of the constrained optimization, we plug into the Lagrangian the definition of $\boldsymbol{\alpha}$ in term of the dual variables (49) and apply the constraint (53):

$$\mathcal{L}(\boldsymbol{\alpha}, b, \boldsymbol{\xi}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2}\sum_{i=1}^{m}|\alpha_i| + C\sum_{i=1}^{m}\boldsymbol{\xi}_i - \sum_{i=1}^{m}\delta_i\left(y_i\left(\sum_{j=1}^{m}\alpha_j y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + b\right) - 1 + \xi_i\right) - \sum_{i=1}^{m}\beta_i\xi_i - \sum_{i=1}^{m}\gamma_i\alpha_i \tag{56}$$

$$\mathcal{L}(\boldsymbol{\alpha}, b, \boldsymbol{\xi}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2}\sum_{j=1}^{m} sign(\alpha_j)\cdot\alpha_j + \sum_{i=1}^{m}(\delta_i + \beta_i)\boldsymbol{\xi}_i \tag{57}$$

$$- \sum_{i=1}^{m}\delta_i\left(y_i\left(\sum_{j=1}^{m}\alpha_j y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + b\right) - 1 + \xi_i\right) - \sum_{i=1}^{m}\gamma_i\alpha_i \tag{58}$$

$$= \frac{1}{2}\sum_{j=1}^{m} sign(\alpha_j)\cdot a_j - \sum_{i=1}^{m}\sum_{j=1}^{m}\left(\delta_i y_i \alpha_j y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + b\delta_i y_i - \delta_i\right) - \sum_{1}^{m}\gamma_i\alpha_i \tag{59}$$

$$= \sum_{j=1}^{m}\left(\sum_{i=1}^{m}\delta_i y_i y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + \gamma_j\right)\cdot a_j - \sum_{i=1}^{m}\sum_{j=1}^{m}\left(\delta_i y_i \alpha_j y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + b\delta_i y_i - \delta_i\right) - \sum_{1}^{m}\gamma_i\alpha_i \tag{60}$$

$$= \sum_{j=1}^{m}\left(\sum_{i=1}^{m}\delta_i y_i y_j a_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\right) - \sum_{i=1}^{m}\sum_{j=1}^{m}\left(\delta_i y_i \alpha_j y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + b\delta_i y_i - \delta_i\right) \tag{61}$$

$$= \sum_{i=1}^{m}\left(\sum_{j=1}^{m}\delta_i y_i y_j a_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\right) - \sum_{i=1}^{m}\sum_{j=1}^{m}\left(\delta_i y_i \alpha_j y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + b\delta_i y_i - \delta_i\right) \tag{62}$$

$$= -\sum_{i=1}^{m} b\delta_i y_i - \delta_i \tag{63}$$

$$= \sum_{i=1}^{m}\delta_i \tag{64}$$

The dual problem is,

$$\begin{aligned} &\min_{\boldsymbol{\delta}} \sum_{i=1}^{m}\delta_i \\ &\text{subject to: } \forall i : \delta_i = C \vee \xi_i = 0 \\ &\forall i : \delta_i = 0 \vee y_i\left(\sum_{j=1}^{m}\alpha_j y_j \Phi(x_i)^T\Phi(x_j) + b\right) = 1 - \xi_i \\ &\forall i : \alpha_i \geq 0 \\ &\forall i : \xi_i \geq 0 \end{aligned} \tag{65}$$

(b)

Derive the equivalent hinge loss minimization problem

$$\begin{aligned} &\min_{\boldsymbol{\alpha}, b, \boldsymbol{\xi}} \frac{1}{2} \sum_{i=1}^{m} |\alpha_i| + C \sum_{i=1}^{m} \xi_i \\ \text{subject to } y_i &\left(\sum_{j=1}^{m} \alpha_j y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + b \right) \geq 1 - \xi_i, i \in [1, m] \\ &\xi_i, \alpha_i \geq 0, i \in [1, m]. \end{aligned} \tag{66}$$

The equivalent hinge loss optimization problem is:

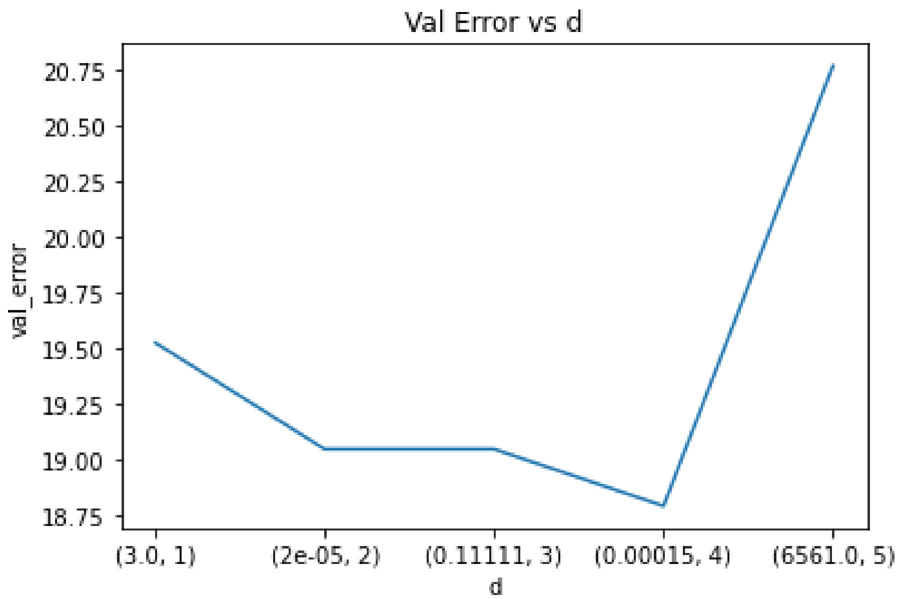
$$\begin{aligned} \min_{\boldsymbol{\alpha}, b} \frac{1}{2} \sum_{i=1}^{m} |\alpha_i| + C \sum_{i=1}^{m} \left(1 - y_i \left(\sum_{j=1}^{m} \alpha_j y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + b \right) \right)_+ \\ \alpha_i \geq 0, i \in [1, m]. \end{aligned} \tag{67}$$

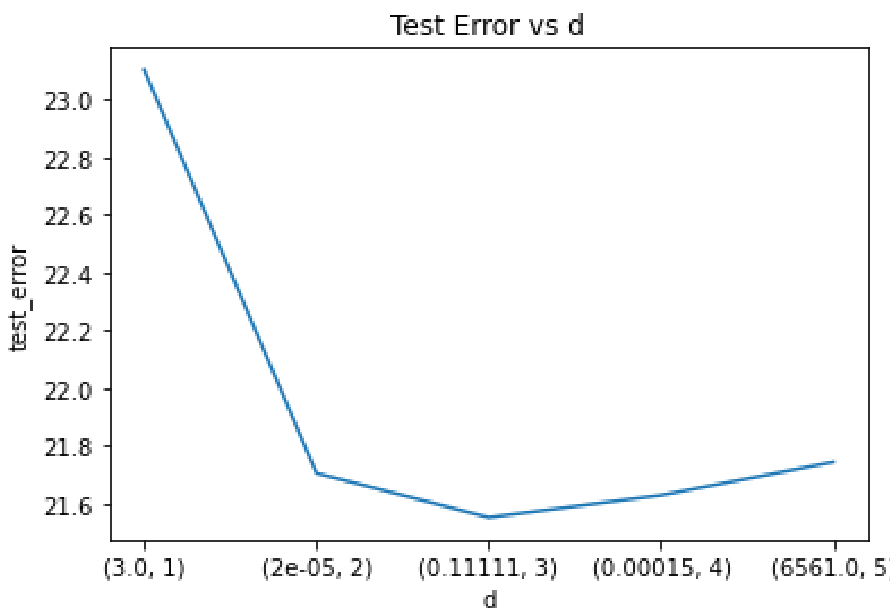
$$\begin{aligned} \min_{\boldsymbol{\alpha}, b} \frac{1}{2} \sum_{i=1}^{m} |\alpha_i| + C \sum_{i=1}^{m} \left(1 - \left(\sum_{j=1}^{m} \alpha_j y_i \Phi(x_i)^T y_j \Phi(x_j) + b y_i \right) \right)_+ \\ \alpha_i \geq 0, i \in [1, m]. \end{aligned} \tag{68}$$

$\hat{y} = \left(\sum_{j=1}^{m} \alpha_j y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + b \right)$ which is obtained by kernel trick, thus the equivalent problem is trying to get a trade-off between the sparsity of α and the error.

Comparing to the original svm: The original svm is trying to get a trade-off between the max margin and the error, and $\hat{y}$ is from definition $wx + b$.

(c)





We use the tuple (C, d) as the x-axis index, d is from 1 to 5, and C is selected based on the best validation error.

References

[1] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Multi-class classification with maximum margin multiple kernel. In *International Conference on Machine Learning*, pages 46–54. PMLR, 2013.

[2] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292, 2001.