

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 $\text{VCdim}(\mathcal{B}_n) \geq n + 1$

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 from p_n to c , thus we can construct the new $p'_n = p_n + \delta(c - p_n)$. $\delta > 0$. Thus we can construct a new ball with new $n+1$ points set $P'_N \cup P_P$. 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 .

For $r > r'$, 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 $B(c', r')$.

(b)

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

$$\|x - c\|^2 \leq r^2 \quad (1)$$

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

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

$$h = \begin{bmatrix} 1, \\ -2c_1, \\ -2c_2, \\ \cdot \\ -2c_n \end{bmatrix} \quad (3)$$

$$x' = \begin{bmatrix} \sum_{i=1}^n \|x_i\|^2 \\ x_1 \\ x_2 \\ \cdot \\ x_n \end{bmatrix} \quad (4)$$

$$b = \sum c_i^2 - r \quad (5)$$

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

We have already know that

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

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 $\text{VCdim}(\mathcal{B}_2) = 3$

Maximum Margin Multiple Kernel

1

(a)

$$\widehat{\mathcal{M}}_q = \{\boldsymbol{\mu} : \boldsymbol{\mu} \in \Delta_q, \widehat{\gamma}_{K_\mu} \geq \gamma_0\} \quad (7)$$

Δ_q is the set of μ ,

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

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

$$K = \sum_{k=1}^p \mu_k K_k \quad (9)$$

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

From the paper,

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

which equals to

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

Converting the optimization to convex optimization, we have

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

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

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

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)

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

We can transform the optimization problem to,

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

$$\beta = \frac{1}{C} \quad (16)$$

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

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

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

Then we can get the KKT,

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

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

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

By using the KKT(18),

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

$$P3 = \sum_{i,y} \alpha_{i,y} (\mathbf{w}_y \cdot \Phi(x_i)) \quad (22)$$

$$= \beta^{-1} \sum_{i,y} \alpha_{i,y} \left(\left[\sum_j (\delta_{y_j,y} - \alpha_{j,y}) \Phi(x_j) \right] \cdot \boldsymbol{\mu} \cdot \Phi(x_i) \right) \quad (23)$$

$$= \beta^{-1} \sum_{i,y} \alpha_{i,y} \boldsymbol{\mu} \cdot \left(\sum_j (\delta_{y_j,y} - \alpha_{j,y}) \Phi(x_j) \cdot \Phi(x_i) \right) \quad (24)$$

$$= \beta^{-1} \sum_{i,y} \alpha_{i,y} \boldsymbol{\mu} \cdot \left(\sum_j (\delta_{y_j,y} - \alpha_{j,y}) K(x_i, x_j) \right) \quad (25)$$

$$= \beta^{-1} \sum_{i,j} K(x_i, x_j) \boldsymbol{\mu} \cdot \left(\sum_y \alpha_{i,y} (\delta_{y_j,y} - \alpha_{j,y}) \right) \quad (26)$$

$$P4 = \sum_{i,y} \alpha_{i,y} (\mathbf{w}_{y_i} \cdot \Phi(x_i)) \quad (27)$$

$$= \beta^{-1} \sum_{i,y} \alpha_{i,y} \left(\left[\sum_j (\delta_{y_j,y_i} - \alpha_{j,y_i}) \Phi(x_j) \right] \cdot \boldsymbol{\mu} \cdot \Phi(x_i) \right) \quad (28)$$

$$= \beta^{-1} \sum_{i,y} \alpha_{i,y} \boldsymbol{\mu} \cdot \left(\sum_j (\delta_{y_j,y_i} - \alpha_{j,y_i}) \Phi(x_j) \cdot \Phi(x_i) \right) \quad (29)$$

$$= \beta^{-1} \sum_{i,y} \alpha_{i,y} \boldsymbol{\mu} \cdot \left(\sum_j (\delta_{y_j,y_i} - \alpha_{j,y_i}) K(x_i, x_j) \right) \quad (30)$$

$$= \beta^{-1} \sum_{i,j} K(x_i, x_j) \boldsymbol{\mu} \cdot (\delta_{y_j,y_i} - \alpha_{j,y_i}) \underbrace{\left(\sum_y \alpha_{i,y} \right)}_{=1} \quad (31)$$

$$= \beta^{-1} \sum_{i,j} K(x_i, x_j) \boldsymbol{\mu} \sum_y \delta_{y_j,y} (\delta_{y_j,y} - \alpha_{j,y}) \quad (32)$$

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

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

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

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

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

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

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

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

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

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

$$+ \sum_{i,y} \alpha_{i,y} \delta_{y_i,y}$$

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

$$+ \sum_{i,y} \alpha_{i,y} \delta_{y_i,y}$$

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

$$= -\frac{C}{2} \sum_{i,j} K(x_i, x_j) \sum_y (\delta_{y_i,y} - \alpha_{i,y}) (\delta_{y_j,y} - \alpha_{j,y}) \boldsymbol{\mu} + \sum_i \alpha_i \cdot \mathbf{1}_{y_i} \quad (\text{Using } C = \frac{1}{\beta}) \quad (40)$$

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

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

Thus we get the following dual problem,

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

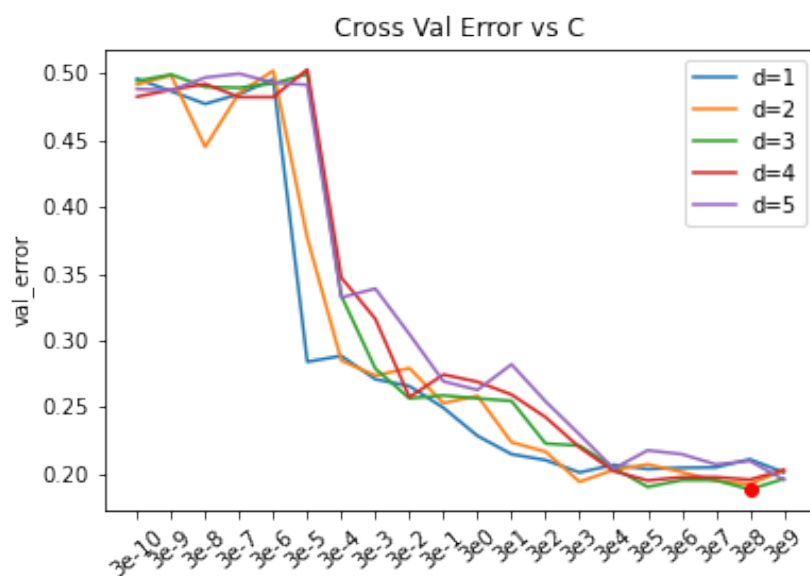
L is concave in $\boldsymbol{\alpha}$. We define the new dual variable $\hat{\boldsymbol{\alpha}}$ as $\boldsymbol{\delta} - \boldsymbol{\alpha}$:

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

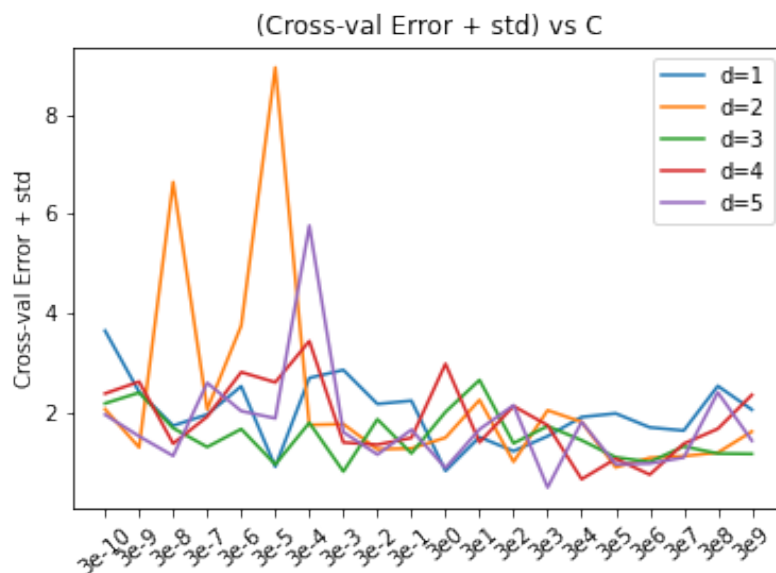
$$\begin{aligned} \min_{\boldsymbol{\mu} \in \widehat{M}_q} \max_{\hat{\boldsymbol{\alpha}} \in \mathbb{R}^{m \times c}} & \sum_{i=1}^m \hat{\boldsymbol{\alpha}}_i \cdot \mathbf{e}_{y_i} - \frac{C}{2} \sum_{i, j=1}^m (\hat{\boldsymbol{\alpha}}_i \cdot \hat{\boldsymbol{\alpha}}_j) \sum_{k=1}^p \mu_k K_k(x_i,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

SVMs hand-on

3

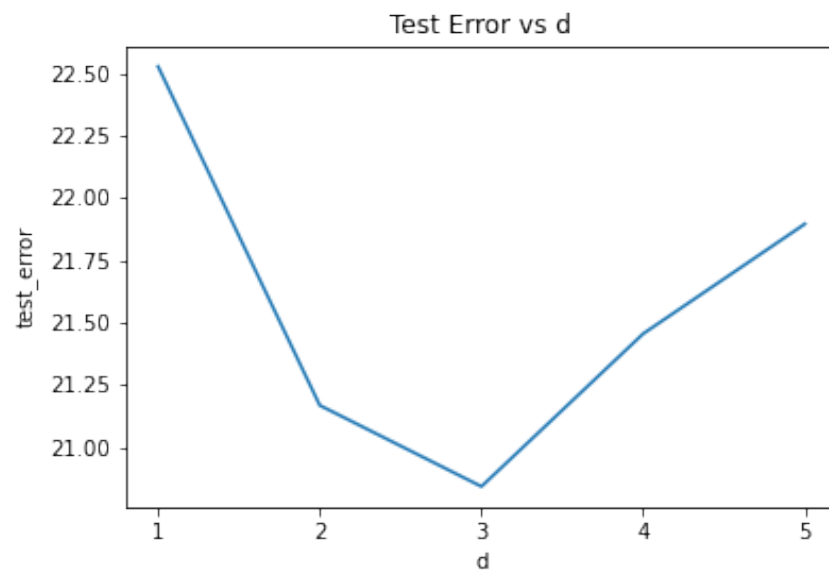
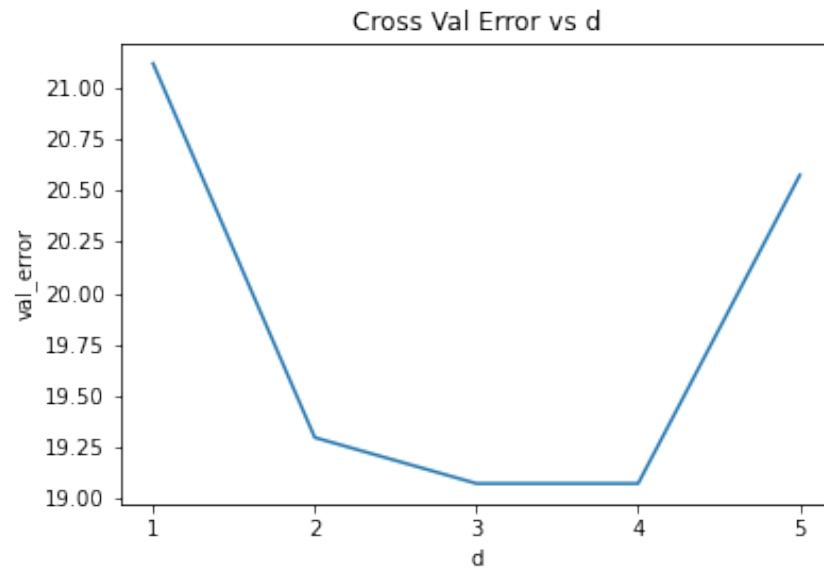


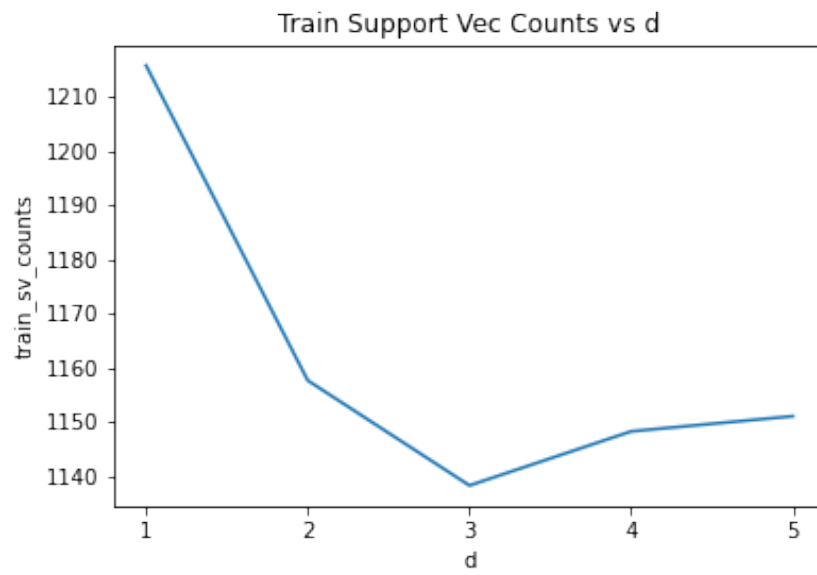
The red point is the optimal combination. The best pair is $d = 3$, and $C = 6561.0$. $(C^*, d^*) = (6561.0, 3)$. The error is 18.8197%.



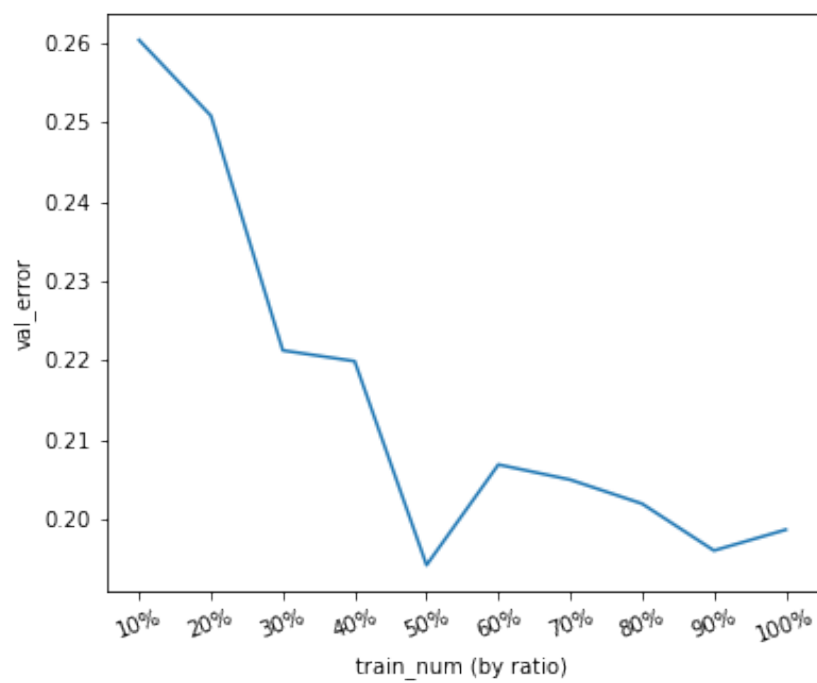
4

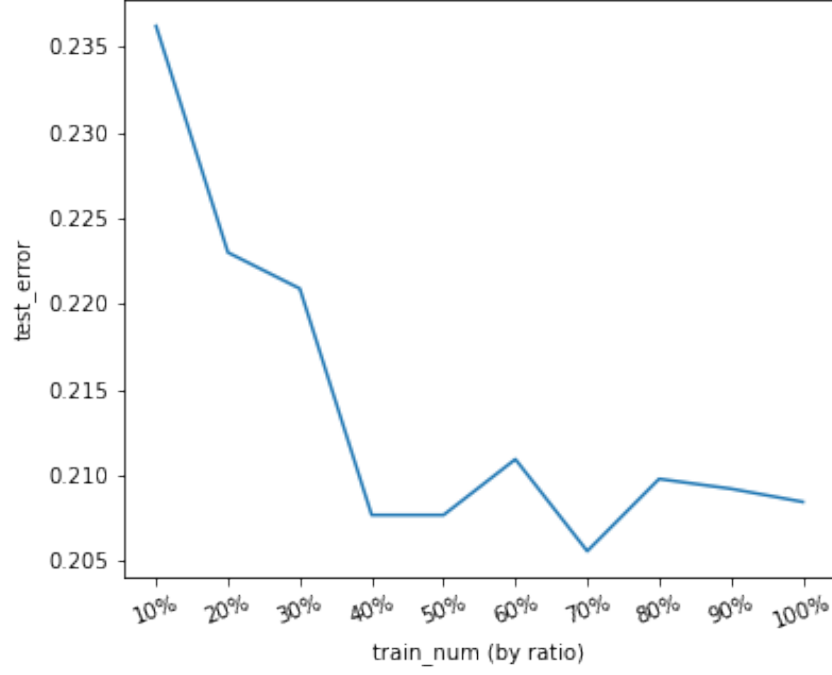
$(C^*, d^*) = (6561.0, 3)$.





5





6

(a)

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

We have the Lagrangian function:

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

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

The KKT conditions are obtained by setting the gradient of the Lagrangian with respect to the primal variables α , b , ξ to zero:

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

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

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

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

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

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

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

To derive the dual form of the constrained optimization, we plug into the Lagrangian the definition of α in term of the dual variables (49) and apply the constraint (53):

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

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

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

$$= \frac{1}{2} \sum_{j=1}^m \text{sign}(\alpha_j) \cdot \alpha_j - \sum_{i=1}^m \sum_{j=1}^m (\delta_i y_i \alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j) + b \delta_i y_i - \delta_i) - \sum_1^m \gamma_i \alpha_i \quad (59)$$

$$= \sum_{j=1}^m \left(\sum_{i=1}^m \delta_i y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) + \gamma_j \right) \cdot \alpha_j - \sum_{i=1}^m \sum_{j=1}^m (\delta_i y_i \alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j) + b \delta_i y_i - \delta_i) - \sum_1^m \gamma_i \alpha_i \quad (60)$$

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

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

$$= - \sum_{i=1}^m b \delta_i y_i - \delta_i \quad (63)$$

$$= \sum_{i=1}^m \delta_i \quad (64)$$

The dual problem is,

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

(b)

Derive the equivalent hinge loss minimization problem

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

The equivalent hinge loss optimization problem is:

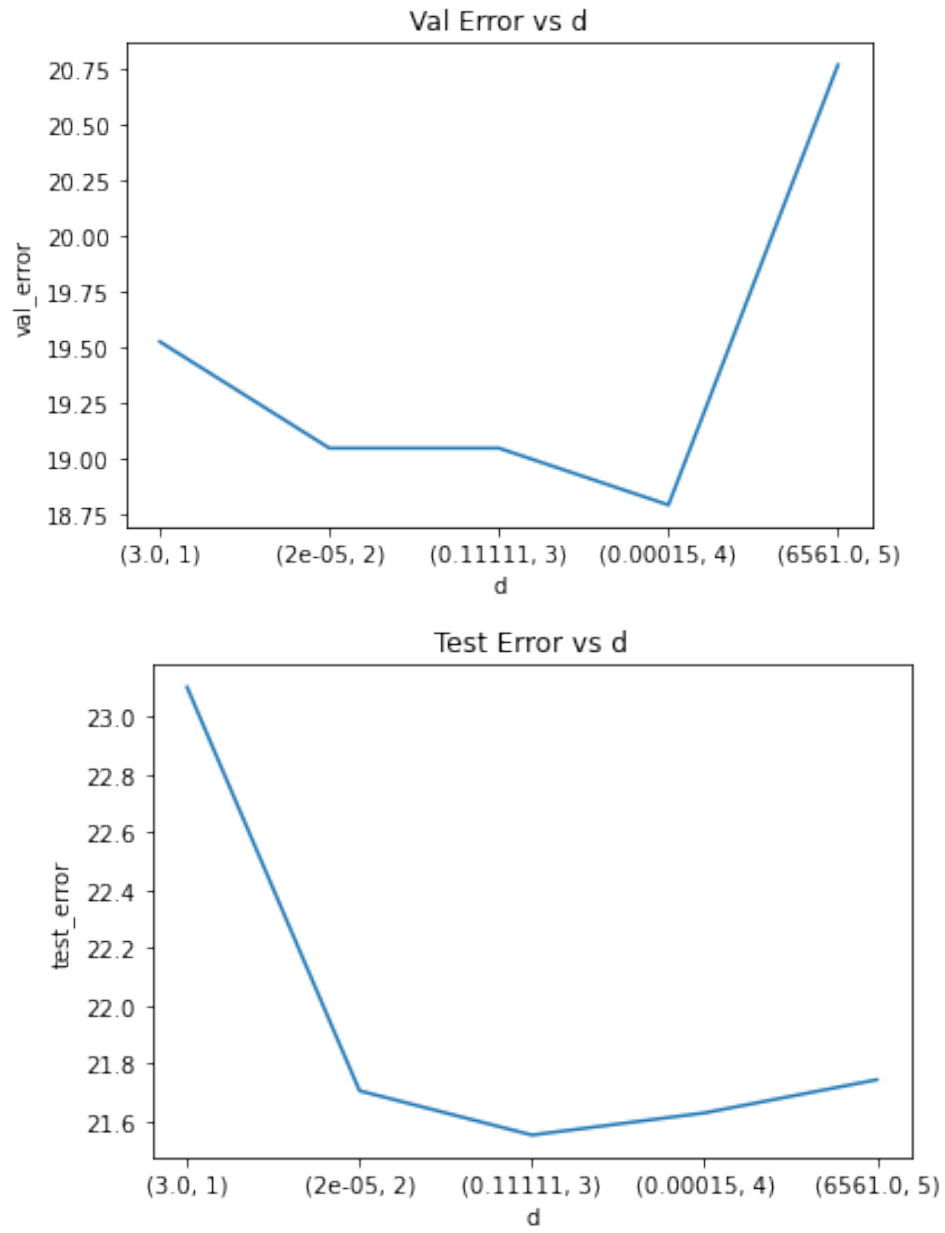
$$\begin{aligned}
& \min_{\alpha, b} \frac{1}{2} \sum_{i=1}^m |\alpha_i| + C \sum_{i=1}^m \left(1 - y_i \left(\sum_{j=1}^m \alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j) + b \right) \right)_+ \\
& \alpha_i \geq 0, i \in [1, m].
\end{aligned} \tag{67}$$

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

$\hat{y} = \left(\sum_{j=1}^m \alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j) + b \right)$ 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.