Homework 2

Fondations of Machine Learning

Xiang Pan (xp2030)

March 17, 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 $(\mathcal{B}_n) \geq n+1$

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 constuct the new $p'_n = p_n + \delta(c - p_n)$. $\delta > 0$. 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 $\mathbb{R}^n$.

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 $x \in B(c, r)$

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

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

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)

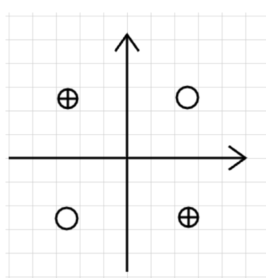
Show that VCdim

$$(\mathcal{B}_2) = 3. \tag{6}$$

We have already know that

$$3 \leq (\mathcal{B}_2) \leq 4 \tag{7}$$

But for four points,



This case can not find a ball to shatter it.
Thus we have VCdim $(\mathcal{B}_2) = 3$

Maximum Margin Multiple Kernel

1

(a)

$$\widehat{\mathcal{M}}_q = \left\{ \boldsymbol{\mu} : \boldsymbol{\mu} \in \Delta_q, \widehat{\gamma}_{K_\mu} \geq \gamma_0 \right\} \tag{8}$$

Δ_q is the set of μ,

$$\Delta_q = \{ \boldsymbol{\mu} : \boldsymbol{\mu} \geq 0, \|\boldsymbol{\mu}\|_q = 1 \} \text{ with } q \geq 1 \tag{9}$$

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

$$K = \sum_{k=1}^{p} \mu_k K_k \tag{10}$$

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

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

which equals to

$$\max_{\boldsymbol{\mu} \in \Delta_q} \frac{1}{m} \sum_{i=1}^{m} \min_{y \neq y_i} \boldsymbol{\mu} \cdot \boldsymbol{\eta} \left(x_i, y_i, y \right) \tag{12}$$

.

Converting the optimization to convex optimization, we have

$$\max_{\substack{\mu \in \Delta_q \\ \gamma}} \sum_{i=1}^{m} \gamma_i \text{ s.t. } \forall i \in [1, m], \forall y \neq y_i, \boldsymbol{\mu} \cdot \boldsymbol{\eta} \left(x_i, y_i, y \right) \geq \gamma_i \tag{13}$$

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

$$\widehat{\gamma}_{K_\mu} = \frac{1}{m} \sum_{i=1}^{m} \min_{y \neq y_i} \boldsymbol{\mu} \cdot \boldsymbol{\eta} \left(x_i, y_i, y \right) \tag{14}$$

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(13).

(b)

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

We can trasnform the optimization problem to,

$$
\begin{aligned}
&\min_{\boldsymbol{\mu}\in\widehat{\mathcal{M}}_q}\ \min_{\mathbf{w},\boldsymbol{\xi}} \frac{1}{2}\sum_{y=1}^{c}\frac{\left\|\mathbf{w}_y\right\|^2}{\mu_k}+C\sum_{i=1}^{m}\xi_i,\\
&\text{subject to: } \forall i\in[1,m], \xi_i\geq 0,\\
&\qquad\qquad\left(\mathbf{w}_{y_i}\cdot\Phi\left(x_i\right)-\mathbf{w}_y\cdot\Phi\left(x_i\right)\right)+\delta_{y_i,y}-1+\xi_i\geq 0
\end{aligned}
\tag{16}
$$

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

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

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

Then we can get the KKT,

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

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

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

By using the KKT(18),

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

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

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

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

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

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

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

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

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

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

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

$$= C \sum_{i,j} K(x_i, x_j) \boldsymbol{\mu} \sum_y \delta_{y_j,y} (\delta_{y_j,y} - \alpha_{j,y}) \tag{32}$$

$$
\begin{aligned}
P1 &= \frac{1}{2}\sum_{y=1}^{c}\left\|\mathbf{w}_y\cdot\frac{1}{\boldsymbol{\mu}}\right\|^2 && (33)\\
&= \frac{1}{2}\sum_{y=1}^{c}(\mathbf{w}_y\cdot\frac{1}{\boldsymbol{\mu}})\cdot(\mathbf{w}_y\cdot\frac{1}{\boldsymbol{\mu}}) && (34)\\
&= \frac{1}{2}\sum_{y=1}^{c}\left[\sum_i\left(\delta_{y_i,y}-\alpha_{i,y}\right)\Phi(x_i)\right]\cdot\left[\sum_j\left(\delta_{y_j,y}-\alpha_{j,y}\right)\Phi(x_j)\right] && (35)\\
&= \frac{1}{2}\sum_{i,j}K(x_i,x_j)\sum_y(\delta_{y_i,y}-\alpha_{i,y})(\delta_{y_j,y}-\alpha_{j,y}) && (36)
\end{aligned}
$$

$$
\begin{aligned}
P3-P4 &= \sum_{i,j}K(x_i,x_j)\boldsymbol{\mu}\sum_y\alpha_{i,y}\left(\delta_{y_j,y}-\alpha_{j,y}\right) && (37)\\
&\quad - C\sum_{i,j}K(x_i,x_j)\boldsymbol{\mu}\sum_y\delta_{y_j,y}(\delta_{y_j,y}-\alpha_{j,y}) && (38)
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{L}(\mathbf{w},\boldsymbol{\xi},\boldsymbol{\alpha}) &= \underbrace{\frac{1}{2}\sum_{y=1}^{c}\left\|\mathbf{w}_y\cdot\frac{1}{\boldsymbol{\mu}}\right\|^2}_{=P1} + \underbrace{\sum_{i,y}\alpha_{i,y}\delta_{y_i,y}}_{=P2} + \underbrace{\sum_{i,y}\alpha_{i,y}\left(\mathbf{w}_y\cdot\Phi\left(x_i\right)\right)}_{=P3} - \underbrace{\sum_{i,y}\alpha_{i,y}\left(\mathbf{w}_{y_i}\cdot\Phi\left(x_i\right)\right)}_{=P4}\\
&= \frac{1}{2}\sum_{i,j}K(x_i,x_j)\sum_y(\delta_{y_i,y}-\alpha_{i,y})(\delta_{y_j,y}-\alpha_{j,y}) \qquad + \sum_{i,y}\alpha_{i,y}\left(\mathbf{w}_y\cdot\frac{1}{\boldsymbol{\mu}}\right.\\
&+ \sum_{i,j}K(x_i,x_j)\boldsymbol{\mu}\sum_y\alpha_{i,y}\left(\delta_{y_j,y}-\alpha_{j,y}\right)\\
&- C\sum_{i,j}K(x_i,x_j)\boldsymbol{\mu}\sum_y\delta_{y_j,y}(\delta_{y_j,y}-\alpha_{j,y})\\
&+ \sum_{i,y}\alpha_{i,y}\delta_{y_i,y}\\
&= -\frac{C}{2}\sum_{i,j=1}^{m}\left(\boldsymbol{\alpha}_i\cdot\boldsymbol{\alpha}_j\right)\boldsymbol{\mu}\cdot K(x_i,x_j) + \sum_{i=1}^{m}\boldsymbol{\alpha}_i\cdot\mathbf{e}_{y_i}\\
&= \sum_{i=1}^{m}\boldsymbol{\alpha}_i\cdot\mathbf{e}_{y_i} - \frac{C}{2}\sum_{i,j=1}^{m}\left(\boldsymbol{\alpha}_i\cdot\boldsymbol{\alpha}_j\right)\sum_{k=1}^{p}\mu_kK_k\left(x_i,x_j\right)
\end{aligned}
\tag{39}
$$

Thus we get the following dual problem,

$$
\begin{aligned}
&\min_{\boldsymbol{\mu}\in\widehat{M}_q}\max_{\boldsymbol{\alpha}\in\mathbb{R}^{m\times c}}\sum_{i=1}^{m}\boldsymbol{\alpha}_i\cdot\mathbf{e}_{y_i} - \frac{C}{2}\sum_{i,j=1}^{m}\left(\boldsymbol{\alpha}_i\cdot\boldsymbol{\alpha}_j\right)\sum_{k=1}^{p}\mu_kK_k\left(x_i,x_j\right)\\
&\text{subject to: }\forall i\in[1,m],\boldsymbol{\alpha}_i\le\mathbf{e}_{y_i}\wedge\boldsymbol{\alpha}_i\cdot\mathbf{1}=0
\end{aligned}
\tag{40}
$$

SVMs hand-on

6

(a)

$$\begin{aligned}
&\min_{\boldsymbol{\alpha},b,\boldsymbol{\xi}} \frac{1}{2}\sum_{i=1}^{m}|\alpha_i| + C\sum_{i=1}^{m}\xi_i \\
\text{subject to } & y_i\left(\sum_{j=1}^{m}\alpha_j y_j K\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)+b\right) \geq 1-\xi_i, i\in[1,m] \\
&\xi_i,\alpha_i \geq 0, i\in[1,m].
\end{aligned} \tag{41}$$

We have the Lagrangian function:

$$\mathcal{L}(\boldsymbol{\alpha},b,\boldsymbol{\xi},\boldsymbol{\delta},\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{1}{2}|\boldsymbol{\alpha}|+C\sum_{i=1}^{m}\boldsymbol{\xi}_i-\sum_{i=1}^{m}\delta_i\left(y_i\left(\sum_{j=1}^{m}\alpha_j y_j K\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)+b\right)-1+\xi_i\right)-\sum_{i=1}^{m}\beta_i\xi_i-\sum_{i=1}^{m}\gamma_i\alpha_i \tag{42}$$

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:

$$\begin{aligned}
\nabla_{\alpha_j}\mathcal{L} = \frac{1}{2}sign(\alpha_j) - \sum_{i=1}^{m}\delta_i y_i\left(y_j K\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)\right)-\gamma_j = 0 \\
\implies \quad \frac{1}{2}sign(\alpha_j) = \sum_{i=1}^{m}\delta_i y_i y_j K\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)+\gamma_j
\end{aligned} \tag{43}$$

$$\nabla_b\mathcal{L} = -\sum_{i=1}^{m}\delta_i y_i = 0 \quad \implies \quad \sum_{i=1}^{m}\delta_i y_i = 0 \tag{44}$$

$$\nabla_{\xi_i}\mathcal{L} = C-\delta_i-\beta_i = 0 \quad \implies \quad \delta_i+\beta_i = C \tag{45}$$

$$\begin{aligned}
\forall i, \delta_i\left(y_i\left(\sum_{j=1}^{m}\alpha_j y_j K\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)+b\right)-1+\xi_i\right) = 0 \\
\implies \quad \delta_i = 0 \vee y_i\left(\sum_{j=1}^{m}\alpha_j y_j K\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)+b\right) = 1-\xi_i
\end{aligned} \tag{46}$$

$$\forall i, \beta_i\xi_i = 0 \quad \implies \quad \beta_i = 0 \vee \xi_i = 0 \tag{47}$$

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 (43) and apply the constraint (46):

$$\mathcal{L}(\boldsymbol{\alpha}, b, \boldsymbol{\xi}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2}|\boldsymbol{\alpha}| + C\sum_{i=1}^{m}\boldsymbol{\xi}_i - \sum_{i=1}^{m}\delta_i\left(y_i\left(\sum_{j=1}^{m}\alpha_j y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + b\right) - 1 + \xi_i\right) - \sum_{i=1}^{m}\beta_i\xi_i - \sum_{i=1}^{m}\gamma_i\alpha_i \tag{48}$$

$$\mathcal{L}(\boldsymbol{\alpha}, b, \boldsymbol{\xi}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2}\sum_{j=1}^{m} sign(\alpha_j)\cdot\alpha_j + \sum_{i=1}^{m}(\delta_i + \beta_i)\boldsymbol{\xi}_i - \sum_{i=1}^{m}\delta_i\left(y_i\left(\sum_{j=1}^{m}\alpha_j y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + b\right) - 1 + \xi_i\right) - \sum_{i=1}^{m} \tag{49}$$

$$= \frac{1}{2}\sum_{j=1}^{m} sign(\alpha_j)\cdot a_j - \sum_{i=1}^{m}\delta_i y_i\left(\sum_{j=1}^{m}\alpha_j y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + b - 1\right) - \sum_{1}^{m}\gamma_i\alpha_i \tag{50}$$

$$= \sum_{i=1}^{m}\left[\sum_{j=1}^{m}\delta_j y_i y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\right]\alpha_i - \sum_{i=1}^{m}\delta_i y_i\left(\sum_{j=1}^{m}\alpha_j y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + b - 1\right) \tag{51}$$

(b)

Derive the equivalent hinge loss minimization problem

References