

csc671-HW3

wma8

November 2019

1 Problem 1

1.1 a:

$$k(x, z) = \alpha k_1(x, z) + \beta k_2(x, z), \text{ for } \alpha, \beta \geq 0$$

(a) $k(x, z) = \alpha k_1(x, z) + \beta k_2(x, z)$, for $\alpha, \beta \geq 0$

We can prove it's kernel by getting its semidefinite \rightarrow

$$\forall C \in \mathbb{R}^{m \times n}, C^T k(x, z) C \geq C^T (\alpha k_1 + \beta k_2) C = \alpha C^T k_1 C + \beta C^T k_2 C$$

Since $\alpha \geq 0$, $\beta \geq 0$, and k_1, k_2 are valid kernels, so we can make sure $C^T k_1 C$, $C^T k_2 C$ are ≥ 0 , so that the whole terms are P.S.D. so $k(x, z)$ is a valid kernel

Problem a

1.2 b:

$$(b) k(x, z) = k_1(x, z)k_2(z)$$

We can prove it's kernel by using inner product

K.

$$k_1(x, z) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(z)$$

$$k_2(x, z) = \sum_{j=1}^{\infty} \lambda_j \psi_j(x) \psi_j(z)$$

$$\begin{aligned} \text{so. } k_1(x, z)k_2(z) &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \lambda_i \lambda_j \psi_i(x) \phi_i(x) \phi_j(z) \psi_j(z) \\ &= \langle \gamma(x), \gamma(z) \rangle_{H_{\text{new}}} \end{aligned}$$

$$\text{where } \gamma(x) = [\sqrt{\lambda_1} \psi_1(x) \phi_1(x)]$$

Problem b

1.3 c:

$$(c) K(x, z) = \underline{|f(x)f(z)|}$$

We can prove it's kernel by taking its semidefinite

(S.D.)

assume $x \in \{x_1, x_2, \dots, x_n\} \subset \mathbb{R}$

$f \in \{f(x_1), f(x_2), \dots, f(x_n)\} \subset \mathbb{R}^n$

$$\begin{aligned} \text{so } k(x, z) &= f^T f, \Rightarrow \forall C \in \mathbb{R}^{n \times n}, C^T K C = C^T f^T f C = (Cf)^T C \geq 0 \\ \text{so L.S. P.S.D} &\Rightarrow \text{it's a valid kernel} \end{aligned}$$

Problem c

1.4 d:

(d) $k(x, z) = f(k_1(x, z))$ where $f(x)$ is poly with positive coeff
So

$$f(k(x, z)) = \{2_1 k(x, z) + 2_2 k(x, z)^2 + 2_3 k(x, z)^3 \dots\}$$

We prove that product of kernel with a positive α is a valid kernel in (b)
and the sum of two valid kernels with positive α is valid by (a), so
that $f(k(x, z))$ is positive in recursive manner.

Problem d

1.5 e:

(e) $k(x, z) = \exp(-\gamma \|x - z\|^2)$

so

$$\exp[-\gamma(x^2 + z^2 - 2xz)] = \underbrace{\exp(-\gamma x^2)}_{\text{exp}(-\gamma \cdot \text{inner product})} \underbrace{\exp(-\gamma z^2)}_{\text{exp}(\gamma \cdot \text{inner product})} \underbrace{\exp(\gamma \cdot 2xz)}_{\text{exp}(\gamma \cdot \text{inner product})}$$

\Rightarrow in the term $\exp(-\gamma x^2) \exp(-\gamma z^2)$ is an inner product, so it's a valid kernel

for $\exp(\gamma \cdot 2xz) = \lim_{i \rightarrow \infty} (1 + 2\gamma xz + \frac{(2\gamma xz)^2}{2} + \dots + \frac{(2\gamma xz)^i}{i!})$, and it's basically a

poly with positive coeff, and xz is an inner product, so $\exp(\gamma \cdot 2xz)$ is

a valid kernel proved by k(d), so $k(x, z)$ is product of two valid kernel,

so it's a valid kernel proved by (b).

Problem e

1.6 f:

$$(+) \quad k(x, z) = \exp(-\gamma \|x - z\|_2^2) = \phi(x)^T \phi(z)$$

$$\exp(-\gamma \|x - z\|_2^2) = \exp[-\gamma (x_1^2 + z_1^2 - 2x_1 z_1)] = \exp(-\gamma x^2) \exp(-\gamma z^2) \exp(\gamma x z)$$

$$= \exp(-\gamma x^2) \exp(-\gamma z^2) \cdot \sum_{i=1}^n (\gamma x_i z_i)^2 / n$$

$$= \left(\exp(-\gamma x^2) \cdot \sum_{i=1}^n ((\sqrt{\gamma} x_i)^2) / n \right) \left(\exp(-\gamma z^2) \cdot \exp(\gamma x z) \right)_H \Rightarrow \text{inner product exists}$$

\Rightarrow proved

Problem f

1.7 g:

(g) Kernel is $\exp(-r\|x-z\|^2)$, for all training point, we set $\hat{y} = \text{sign}(w^T x + b)$ as classifier

So the Lagrangian function for these hard Margin SVM,

$$L(w, b, \alpha) = \frac{1}{2} w^T w + \sum_{i=1}^N \alpha_i (1 - y_i (w^T x_i + b))$$

As we apply KKT condition, $\frac{\partial L}{\partial w} = 0$, $\frac{\partial L}{\partial b} = 0$, $\frac{\partial L}{\partial \alpha} = 0$

$$\Rightarrow \text{we can get } w^T x = \sum \alpha_i y_i k(x_i, x)$$

\Rightarrow classifier becomes $\text{sign}(\sum \alpha_i y_i k(x_i, x) + b)$

$$\text{as } r \rightarrow \infty, \exp(-r\|x-z\|^2) = \begin{cases} 0 & \text{if } (x \neq z) \\ 1 & \text{if } (x = z) \end{cases}$$

so $\hat{y} = \text{sign}(\sum \alpha_i y_i + b)$ (since only one $x_i = x$)

$$\Rightarrow b = y_i - \sum \alpha_i y_i k(x_i, x_i) \Rightarrow b = y_i - \alpha_i y_i \text{ (according to previous theorem)}$$

$$\Rightarrow \hat{y} = \text{sign}(\sum \alpha_i y_i + y_i - \alpha_i y_i) = \text{sign}(y_i)$$

Therefore, for $r \rightarrow \infty$, $\hat{y} = \text{sign}(y_i)$ for $\forall \{x_i, y_i\}$, and it can achieve 0 error in training data.

Problem g

2 Problem 2

2.1 a:

(a) prove that H can shatter at least $d+1$ points, which means the H can produce any pre-specified $\{f_i\}$ for $d+1$ points's output.

$$H = \{f | f(x) = \text{sign}(w^T x + b), b \in R\}$$

Let's make $d+1$ points represented by matrix:

$$X = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{bmatrix} \in \mathbb{R}^{d+1 \times d+1}$$

The first coordinate of each x_i is 1, so the bias term can take its effect. Besides the first row, W^T can assign +/- sign for any point by giving weight. For the first row, the bias term can assign (+/-) directly without giving weight for the extra dimension.

$$\text{so } \text{dvc}(H) \geq d+1$$

Now we need to prove $\text{dvc}(H) \leq d+1$

Assume we have $d+2$ points in $(d+1)$ dimension space. There exists $x_j = \sum_{i=1}^d a_i x_i$ such that not all a_i are 0. Set $y_j = -1$ and $y_i = \text{sign}(a_i) = \text{sign}(w^T x_i)$. If $a_i = 0$, y_i can be arbitrary.

So $\text{Sign}(w^T x_i) = \text{Sign}(\sum_{j=1}^d a_j w^T x_j) > 0 \neq y_j$ so there exists an assignments of $d+2$ points not achievable by H . So $\text{dvc}(H) \leq d+1 \Rightarrow \text{dvc}(H) = d+1$

Problem a

2.2 b:

Since the VC-dimension of SVM kernel is the kernel dimension.

For each $\Phi(x)$, the dimension will be $\binom{p+d-1}{d}$, so due to the Hilbert Space of kernel, the kernel dimension should be $\binom{\sum_{i=0}^d p+di-1}{d} \rightarrow \binom{p+d}{d}$, so the kernel's dimension is $\binom{p+d}{d}$, and since we have proved that the *VCdimension* equals to the terms in the *function* in (2a). Therefore, the VC dimension is $\binom{p+d}{d}$.

2.3 c:

We have proved RBF is capable of classifying any number of data points, so it basically means RBD is capable of shattering infinite amount of data points; Therefore, the VC-dimension for RBF is ∞ .

2.4 di:

According to Occham's Razor. $\forall f \in F, R(true) \leq R(empirical) + \sqrt{\frac{\log M + \log(1/\delta)}{2n}}$, where M is the number of classifiers. For k literals in the conjunction, we can have 3^k classifiers because each classifier can either choose x or $\neg x$ or not pick that term. Therefore:

$$\text{Upper Bound} = R(\text{empirical}) + \sqrt{\frac{\log 3^p + \log(1/\delta)}{2n}}$$

2.5 dii:

For function class \mathcal{F} to be at most p' literals, which means we can have $\sum_{k=1}^{k \leq p} 2^k$ classifiers for each specific literals combination among k , and we can have $\binom{p}{k}$ choices for classifiers for each $p \leq k$. Therefore:

$$\text{Upper Bound} = R(\text{empirical}) + \sqrt{\frac{\log \sum_{k=1}^{k \leq p} 2^k \binom{p}{k} + \log(1/\delta)}{2n}}$$

3 Problem 3

3.1 a:

For

$$y_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i, \text{ where } \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \frac{1}{\lambda} \mathbf{I}) \text{ and noise } \epsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

$$\Pr(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \Pr(y_1 \dots y_n | x_1 \dots x_n, \mathbf{w})$$

$$= \prod_{i=1}^n \Pr(y_i | x_i, W) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\mathbf{w}^T \mathbf{x}_i - y_i)^2}{2}\right)$$

3.2 b:

$$(b) P(w|Y, X, \lambda) = P(Y|w, X, \lambda) \cdot P(w) / P(Y)$$

$$= \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{1}{2} \|y_i - w^T x_i\|_2^2\right) \cdot \frac{1}{\sqrt{2\pi}\lambda} \cdot \exp\left(-\frac{\lambda}{2} \|w\|^2\right) / \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \|y\|_2^2\right)$$

$$= \frac{1}{\sqrt{2\pi}\lambda} \cdot \exp\left(-\frac{1}{2} \|y_i^2 + (w^T x_i)^2 - 2w^T x_i y_i - \frac{\lambda}{2} \|w\|^2 + \frac{1}{2} \|y\|_2^2\right)$$

$$= \frac{1}{\sqrt{2\pi}\lambda} \cdot \exp\left(-\frac{1}{2} (w^T x_i^2 + w^T x_i y_i - \frac{\lambda}{2} \|w\|^2)\right)$$

$$= C \cdot \exp\left(-\frac{1}{2} (w^T (x_i^T + \lambda I) w - 2w^T x_i y_i)\right)$$

$$= C \cdot \exp\left(-\frac{1}{2} (w^T (x^T x + \lambda I) w - 2w^T x y \cdot (x^T x + \lambda I)^{-1} \cdot (x^T x + \lambda I))\right)$$

$$\Rightarrow C \cdot \exp\left(-\frac{1}{2} (w - (x^T x + \lambda I)^{-1} x^T y)^T (x^T x + \lambda I) (w - (x^T x + \lambda I)^{-1} x^T y)\right)$$

$$\text{For } \arg\max(w) \Rightarrow w = (x^T x + \lambda I)^{-1} x^T y$$

Problem b

3.3 c:

$$(c) F(w) = \|y - Xw\|_2^2 + \lambda \|w\|_2^2$$

$$\begin{aligned} \frac{dF(w)}{dw_j} &= -2 \bar{x}_j^T (y - X \cdot w) + 2\lambda w_j \\ &= 2(-\bar{x}_j^T y + \bar{x}_j^T X \cdot w + \lambda w_j) \end{aligned}$$

$$\begin{aligned} \text{For } \nabla F(w) = 0 \\ -\bar{x}_j^T y + \bar{x}_j^T X \cdot w + \lambda w_j = 0 \\ w_j = \frac{\bar{x}_j^T y}{\bar{x}_j^T X + \lambda} \end{aligned}$$

Problem c

3.4 d:

From (b) (c), we realize that the w^* that minimizes L^2 Loss with L^2 regularization term equals to the w that maximizes the A Posteriori expression.

3.5 e:

- (i) We need to set the w to be $\arg\min \|y - x(w + \mu)\|_2^2 + \lambda \|w + \mu\|_2^2$
- (ii) We need the regularization term to add additional penalty and avoiding

over-fitting; otherwise the final Regression model might have high dimension and likely to over-fit the data.

4 Problem 4

4.1 a:

(i)

The

$$\min_{\mathbf{c}_1, \dots, \mathbf{c}_K} \sum_{i=1}^N \min_{k \in \{1, \dots, K\}} \|\mathbf{x}_i - \mathbf{c}_k\|_1,$$

try to minimizes the Manhattan Distance between points and the cluster centers.

While the k-means try to minimize the Euclidean distances between the points and the cluster centers.

4.2 b:

Check the ipynb file

4.3 c:

Check the ipynb file