csc671-HW3

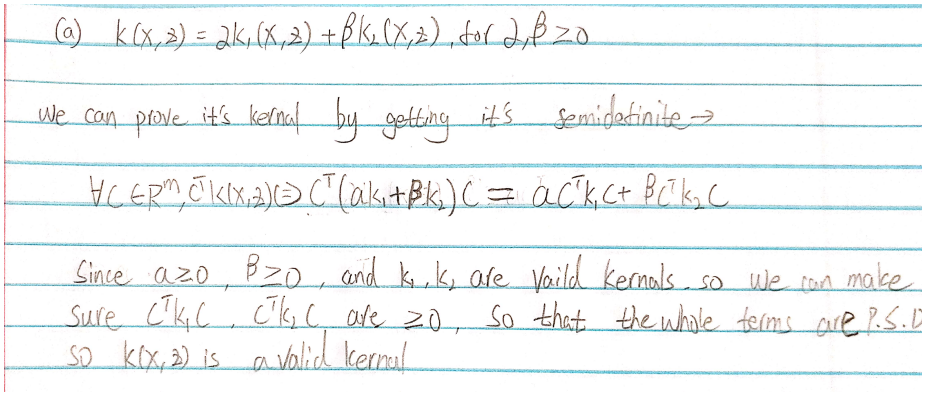
wma8

November 2019

1 Problem 1

1.1 a:

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



Problem a

1.2 b:

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

We can prove it's kernal by using inner product

$$k_1(x,z) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(z)$$
$$k_2(x,z) = \sum_{j=1}^{\infty} \lambda_j \phi_j(x)\phi_j(z)$$

So. $k_1(x,z)k_2(x,z) = \sum_{i}^{\infty}\sum_{j}^{\infty} \lambda_i\lambda_j \psi_i(x)\phi_j(x)\psi_i(z)\phi_j(z)$

$= \langle \gamma(x), \gamma(z) \rangle_{H_{new}}$

where $\gamma(x) = [\sqrt{\lambda_i\lambda_j}\psi_i(x)\phi_j(x)]$

Problem b

1.3 c:

(c) $K(x,z) = f(x)f(z)$

We can prove it's kernal by taking it's semidefinite

assume $x \in \{x_1, x_2 \cdots x_n\}$ $n \in R$

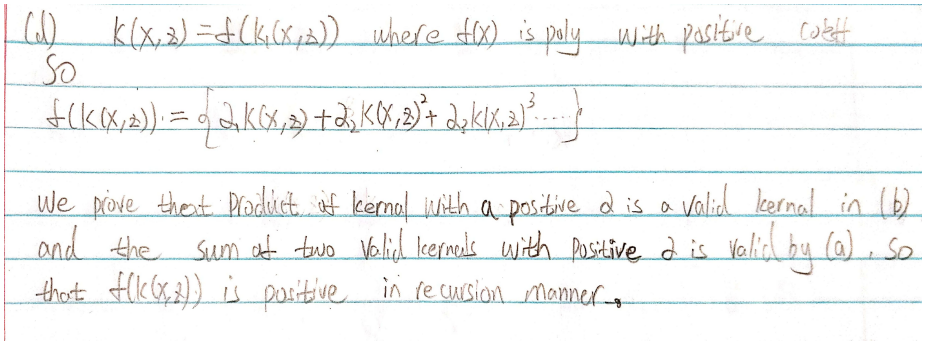
$f \in \{f(x_1), f(x_2) \cdots f(x_n)\}$ $n \in R$

So $k(x,z) = f^T f_2 \Rightarrow \forall c \in R^m$ $c^T k c = c^T f^T f c = (cf)^T cf \geq 0$

so H.S P.S.D $\Rightarrow$ it's a valid kernal

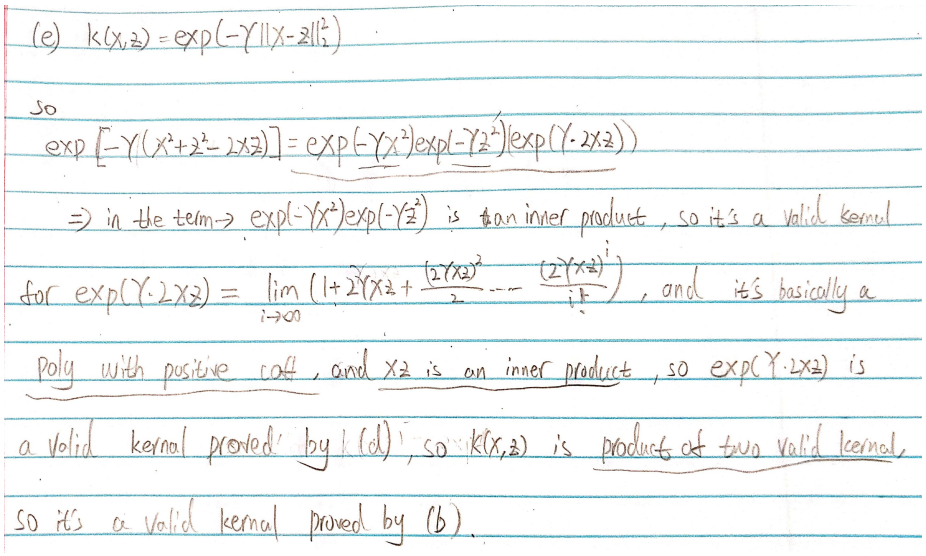
Problem c

1.4 d:



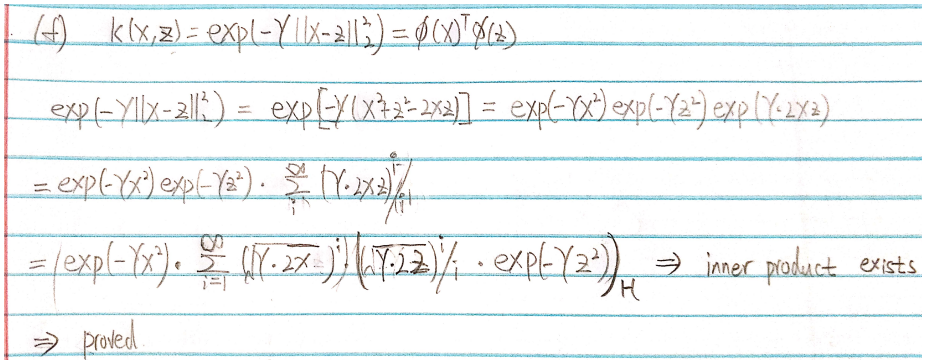
Problem d

1.5 e:



Problem e

1.6 f:



Problem f

1.7 g:

(g) Kernel is $\exp(-r\|x-z\|_2^2)$, for all training point. we set $\hat{y} = \text{sign}(w^T x + b)$ as classifier

So the Lagran 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, $\frac{\partial L}{\partial w} = 0$, $\frac{\partial L}{\partial b} = 0$, $\frac{\partial L}{\partial \alpha} = 0$

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

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

as $r \to \infty$, $\exp(-r\|x-z\|_2^2) = \begin{cases} 0 & \text{if } (x_i \neq x) \\ 1 & \text{if } (x_i = x) \end{cases}$

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

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

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

Therefore, for $\gamma \to \infty$, $\hat{y} = \text{sign}(y_i)$ 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 $\{-1, 1\}$ for d+1 points's output.

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

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

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

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.

so $d_{vc}(H) \geq d+1$

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

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

So $Sign(w^T x_j) = sign(\sum_{i \neq j} a_i w^T x_i) > 0 \neq y_j$ so there exists an assignment of d+2 points not achievable by H. so $d_{vc}(H) \leq d+1 \Rightarrow d_{vc}(H) = d+1$

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 $\binom{\sum_{di=0}^{d} p+di-1}{d} \rightarrow \binom{p+d}{d}$, 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 classify 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,$R(true) \leq R(empirical)+\sqrt{\frac{logM+log(1/\delta)}{2n}}$, 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:

Upper Bound = R(empirical) + $\sqrt{\frac{log3^p+log(1/\delta)}{2n}}$

2.5 dii:

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

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

3 Problem 3

3.1 a:

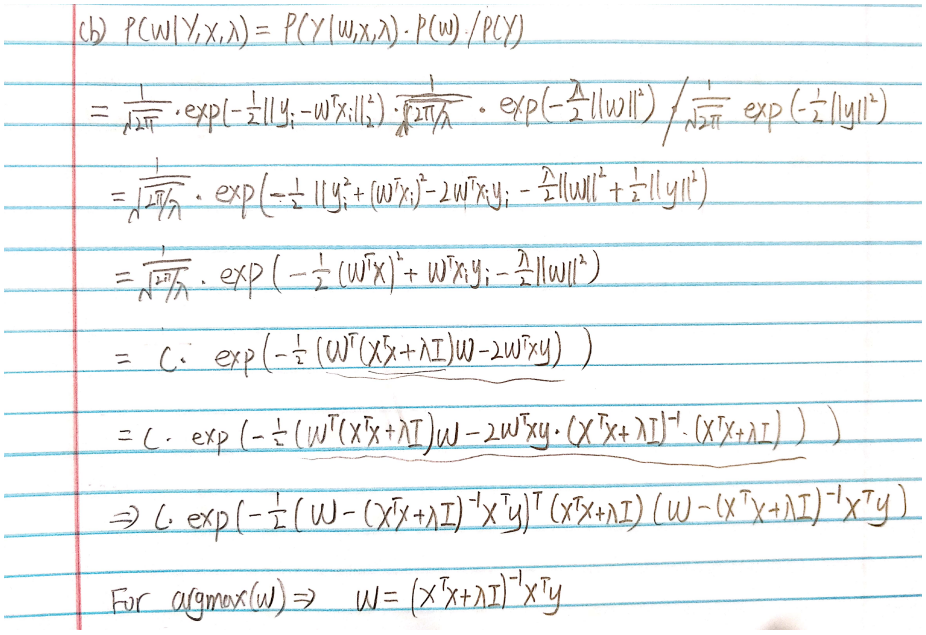
For

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

$\Pr(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \Pr(y_1...y_n|_1....x_n, \mathbf{w})$

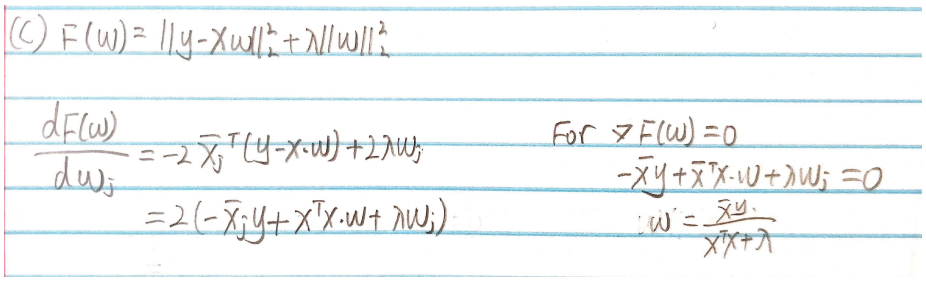
$= \prod_{i=1}^n \Pr(y_i|x_i, W) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} exp(-\frac{(w^T x_i - y_i)^2}{2})$

3.2 b:



Problem b

3.3 c:



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 $argmin\ ||y - x(w + \mu)||_2^2 + \lambda||w + \mu||_2^2$
(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

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

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