Shirui Ye
CS542
PS 4
Cover

I tried to use Latex and work on Overleaf. However, I found out it took me 1 hr and 15min just for first question, since It took a lot of time to type in the math symbols and doing so really interrupted my logic. In this case, to give my best effort, I decide to hand write this assignment. Thanks for the understanding.

Shimi Ye

PS4

CS542

1.(a) Bishop 6.2

① Show w can be written as a linear combination of vectors $t_n\phi(x_n)$

② denote coefficients by α_n, derive formulation of perceptron and the predictive function in terms of α_n

③ Show vector $\phi(x)$ enters only in the form of the kernel function $k(x,x')=\phi(x)^T\phi(x')$

To do ①, we can initiate with $w=0$, then it increases by $t_n\phi(x_n)$. In this case, the final weight vector gets a linear combination which contains vector $t_n\phi(x_n)$. Thus we can write w in a linear combination of $t_n\phi(x_n)$ as following: $w=\sum_{n=1}^{N}\alpha_n t_n\phi(x_n)$

② In above equation, α_n is the how many times of w use n to update traing process, In this case, the predictions according to perceptron are below:

$y(x)=\text{sign}(w^T\phi(x))$ $\quad y(x)=\text{sign}(\sum_{n=1}^{N}\alpha_n t_n k(x_n,x))$

$\Downarrow$

$\sum_{n=1}^{N}\alpha_n t_n\phi(x_n)$

③ Above is in kernel function form, we can write similar form of learning algorithm of perceptron as $\alpha_n\to\alpha_n+1$ so that patterns satisfy form $t_n(w^T\phi(x_n))\geqslant 0$, use $w=\sum_{n=1}^{N}\alpha_n t_n\phi(x_n)$ when $\alpha_n\geqslant 0$, then we get $t_n(\sum_{m=1}^{N}k(x_m,x_n))\geqslant 0$ which is in the kernel function form

Shimi Ye

PS 4

CS542

1. (b) Bishop 7.3

We want to show, unrelated to the dimensionality of the data space, a data set consisting of 2 data points, one from each class, is enough to get the location of the maximum margin hyperplane.

To show above, we suppose that we have a data set which consisting of 2 points $d_1 \in C_+ (t_1 = 1)$ and $d_2 \in C_- (t_2 = -1)$, then max margin hyperplane is gotten by solving $\arg\min_{w,b} \frac{1}{2}\|w\|^2$ which has boundary of $\begin{cases} w^T d_1 + b = 1 \\ w^T d_2 + b = -1 \end{cases}$

Now, we apply Lagrange multipliers η, λ and do below;

$$\arg\min_{w,b} \{ \eta(w^T d_2 + b + 1) + \lambda(w^T d_1 + b - 1) + \frac{1}{2}\|w\|^2 \}$$

First, we take $\frac{dw}{d}$ $\arg\min_{w,b} \{\eta(w^T d_2 + b + 1) + \lambda(w^T d_1 + b - 1) + \frac{1}{2}\|w\|^2\}$ and then $\frac{db}{d}$

we get $= \eta d_2 + \lambda d_1 + w$ we set this to equal 0

second take $\eta d_2 + \lambda d_1 + w = 0$ and get $\eta + \lambda = 0$

we can obtain $\lambda = -\eta$ now, combine with $\eta d_2 + \lambda d_1 + w = 0$

we get $\lambda(d_1 - d_2) = w$

Finally, we add $\begin{cases} w^T d_1 + b = 1 \\ w^T d_2 + b = -1 \end{cases}$ together, get $w^T (d_1 + d_2) + 2b = 0$

$$\Downarrow$$

$$-w^T (d_1 + d_2) = 2b$$

Now, we apply $\lambda(d_1 - d_2) = w$ and we can get below

$$b = -\frac{1}{2}\lambda (d_1 - d_2)^T (d_1 + d_2) = -\frac{1}{2}(d_1^T d_1 - d_2^T d_2)$$

According to Lagrange, we know λ is undetermined, so it shows inherent indeterminacy of w, b.

Thus, we showed a data set consisting of 2 points, one from each class is enough to get the location of the maximum margin hyperplane and is unrelated to the dimensionality of the data space

Shirui Ye
PS4
CS542

1. (C) Bishop 7.4

We need to show that ρ of the margin for max-margin hyperplane $\frac{1}{\rho^2} = \sum_{n=1}^{N} a_n$ → given by 7.10 and has constraint 7.11, 7.12

$a_n \geq 0, n=1,\dots,N$ $\quad \sum_{n=1}^{N} a_n t_n = 0$

According to $t_n(w^T\phi(x_n)+b)=1$ and the margin value $\rho = \frac{1}{\|w\|}$ we get $\frac{1}{\rho^2} = \|w\|^2$

Then, according to $a_n\{t_n y(x_n) - 1\} = 0$, when at max margin solution,

in $L(w,b,a) = \frac{1}{2}\|w\|^2 - \underbrace{\sum_{n=1}^{N} a_n\{t_n(w^T\phi(x_n)+b)-1\}}_{\text{This get eliminated}}$

Then, apply $w = \sum_{n=1}^{N} a_n t_n \phi(x_n)$, so 7.10 dual can be expressed as

$\frac{1}{2}\|w\|^2 = \sum_{n}^{N} a_n - \frac{1}{2}\|w\|^2$ which follows the question conclusion.