COMS 4771 Machine Learning (Fall 2022)
Problem Set #2

Xinhao Li - xl2778@columbia.edu

2022/10/23

Problem 1

(i)

A simple example could be that: Let's say we have a square. The two points at one diagonal have the same label and rest two points at the other diagonal have the different label. This case is clearly not linear separable. For the target function φ and set S, they can be expressed as:

$$\varphi = (x_1 \wedge x_2) \vee (\bar{x_1} \wedge \bar{x_2}) \tag{1}$$

$$S = \{((0\ 0)^T, 1), ((0\ 1)^T, -1), ((1\ 0)^T, -1), ((1\ 1)^T, 1)\} \tag{2}$$

(ii)

This question asks us to find a way to compute $K(\vec{a}, \vec{b})$ in $O(d)$ time. We can use the Kernel Trick which is the dot product between two data points in kernel space can be computed relatively quick. By looking at the features after the kernel transformation, for d = 2 case, the feature transformation function looks like:

$$\phi(x)_{d=2} = (1 + x_1 + \bar{x_1})(1 + x_2 + \bar{x_2}) \tag{3}$$

Then we can formally define ϕ as:

$$\phi(\vec{x}) = \prod_{k=1}^{d}(1 + x_k + \bar{x_k}) \tag{4}$$

Therefore, $K(\vec{a}, \vec{b})$ becomes:

$$\begin{aligned} K(\vec{a}, \vec{b}) &= \phi(\vec{a}) \cdot \phi(\vec{b}) \\ &= \prod_{k=1}^{d}(1 + a_k + \bar{a_k}) \cdot \prod_{k=1}^{d}(1 + b_k + \bar{b_k}) \\ &= \prod_{k=1}^{d}(1 + a_k + \bar{a_k}) \cdot (1 + b_k + \bar{b_k}) \\ &= \prod_{k=1}^{d}(1 + a_k b_k + \bar{a_k}\bar{b_k}) \end{aligned} \tag{5}$$

Since:

$$1 + a_k b_k + \bar{a_k}\bar{b_k} := \begin{cases} 2 & \text{if } a_k = b_k \\ 1 & \text{otherwise} \end{cases} \tag{6}$$

We will have:

$$\begin{aligned} K(\vec{a}, \vec{b}) &= \prod_{k=1}^{d}(1 + a_k b_k + \bar{a_k}\bar{b_k}) \\ &= \prod_{k=1}^{d} 2^{a_k b_k + \bar{a_k}\bar{b_k}} \\ &= 2^{\sum_{k=1}^{d} 2^{a_k b_k + \bar{a_k}\bar{b_k}}} \\ &= 2^{\vec{a}\cdot\vec{b} + \bar{\vec{a}}\cdot\bar{\vec{b}}} \end{aligned} \tag{7}$$

Thus, equation 7 can be computed in $O(d)$ time.

(iii)

Based on the definition of w^*, we know that $w_1^* = -0.5$. For the rest values of w^*, they are determined by both $\Phi(x)$ and $\varphi(x)$, which is $\forall i > 1$, $w_i^* = 1$ iff $\Phi(x_i)$ is one of the conjunctions of φ. This means if the conjunction a in φ appear in $\Phi(x)$, then $w^*_{i:\Phi(x_i)=a} = 1$. We know the label of one point is determined by:

$$sign(w^* \cdot \Phi(x)) \tag{8}$$

This means to determine the sign of one point, the only valid multiplication between w^* and $\Phi(x)$ will be:

$$\sum\Big(\text{conjunctions in } \varphi\Big) - 0.5 \tag{9}$$

We know that φ is a DNF formula, so if the label of one point is 0, this means that all conjunctions in φ should be 0. Then even though in w^*, some elements have value 1, but $1 \times 0 = 0$. Therefore:

$$sign(w^* \cdot \Phi(x)) = sign(-0.5) = 0 \tag{10}$$

If the label of one point is 1, this means that there must be one conjunction in φ has the value of 1. Then in w^*, there could be several elements having value 1, but at least one element will have its associated conjunction also has value 1. Therefore:

$$sign(w^* \cdot \Phi(x)) = sign(\geq 0.5) = 1 \tag{11}$$

In conclusion, w^* linearly separates $\Phi(S)$.

To find the lower bound for the margin γ, the definition of γ is given by:

$$\min_{(\Phi^{(x_i)}, y^{(i)}) \in \Phi(S)} y_i \big(\frac{w^*}{||w^*||} \cdot \Phi(x^{(i)}) \big) = \min_{(\Phi^{(x_i)}, y^{(i)}) \in \Phi(S)} \frac{y_i \ w^* \cdot \Phi(x^{(i)})}{||w^*||} \tag{12}$$

To minimize equation 12, we can either maximize $||w^*||$ or minimize $y_i \ w^* \cdot \Phi(x^{(i)})$. There is no room to maximize $||w^*||$ since the lower bound should depend on s, the number of conjunctions in φ. So, $||w^*||$ always equals to:

$$||w^*|| = \sqrt{(-0.5)^2 + s} \ , where \ \sum_i^s 1^2 = s \tag{13}$$

Then to minimize $y_i \ w^* \cdot \Phi(x^{(i)})$, when the label is 0, we have

$$y_i \ w^* \cdot \Phi(x^{(i)}) = -1 * (1 \times -0.5) = 0.5 \tag{14}$$

since no conjunctions have value 1.
When the label is 1, the way to minimize is to only have one conjunction has value 1, therefore, we have:

$$y_i \ w^* \cdot \Phi(x^{(i)}) = 1 * (1 \times -0.5 + 1 \times 1) = 0.5 \tag{15}$$

In conclusion, the lower bound for the margin γ is:

$$\gamma = \frac{0.5}{\sqrt{(-0.5)^2 + s}} \tag{16}$$

(iv)

To find an upper bound on the radius R of the dataset $\Phi(S)$, we need to think about what are included in the dataset. $\Phi(S)$ includes all possible combinations of x, $\bar{x}$, and 1 and they are conjunctions. Since they are conjunctions, if there is one 0 in the conjunctions, then this whole term is 0. so even the whole $\Phi(S)$ space has 3^d elements, but the maximum bound is not 3^d. Then the most extreme case would be, we have all x in value 1, then all $\bar{x}$ are in value 0. In this case, any element which contains $\bar{x}$ will be 0 and will not be counted into the radius of R. Then the previous 3^d elements will be reduced to 2^d elements, since now we will only consider all possible combinations of x and 1. Therefore, the upper bound on the radius R is:

$$R = \sqrt{\sum_i^{2^d} 1} = 2^{\frac{d}{2}} \tag{17}$$

(v)

Since we have n data points and each dot product takes $O(d)$ time, so in total the running time is: $O(nd)$. To derive the mistake bound, as it has been proved in the lecture:

$$T \leq (\frac{R}{\gamma})^2 \tag{18}$$

Then substitute R and γ into the above equation, we have:

$$T \leq \frac{(0.25 + s)2^d}{0.25} \leq (1 + 4s)2^d \tag{19}$$

Therefore, the mistake bound is: $O(s2^d)$.

Problem 2

As the hint given in the question, let $(\tilde{x_1}, \tilde{y_1})$, ... , $(\tilde{x_m}, \tilde{y_m})$ be another i.i.d. random sample, independent of (x_1, x_2), ... , (x_n, x_n), but having the same distribution as (x, y). Then

$$R(\tilde{w}) = E[\frac{1}{m}\sum_{i=1}^{m}(\tilde{w} \cdot \tilde{x_i} - \tilde{y_i})^2], \ \hat{w} \in \Re^d \tag{20}$$

So here, basically, we want to prove that:

$$E[\frac{1}{n}\sum_{i=1}^{n}(\hat{w} \cdot x_i - y_i)^2] \leq E[\frac{1}{m}\sum_{i=1}^{m}(\hat{w} \cdot \tilde{x_i} - \tilde{y_i})^2] \tag{21}$$

First, we know that the expected test error is independent of the number of test points based on the linearity of expectation. And we have:

$$E[\frac{1}{m}\sum_{i=1}^{m}(\tilde{w} \cdot \tilde{x_i} - \tilde{y_i})^2] = E[(\tilde{w} \cdot \tilde{x_1} - \tilde{y_1})^2] \tag{22}$$

Then we also have:

$$E[\frac{1}{n}\sum_{i=1}^{n}(\hat{w} \cdot x_i - y_i)^2] = E[(\hat{w} \cdot x_1 - y_1)^2] \tag{23}$$

Based on the assumption that $(\tilde{x_1}, \tilde{y_1})$ is independent of (x_1, x_2), we will have:

$$E[\frac{1}{n}\sum_{i=1}^{n}(\hat{w} \cdot x_i - y_i)^2] = E[\frac{1}{m}\sum_{i=1}^{m}(\tilde{w} \cdot \tilde{x_i} - \tilde{y_i})^2] \tag{24}$$

Since $\hat{w}$ and $\tilde{w}$ denote the squared training error minimizing decision boundary based on samples (x_1, x_2), ... , (x_n, x_n) and $(\tilde{x_1}, \tilde{y_1})$, ... , $(\tilde{x_m}, \tilde{y_m})$, respectively, therefore, for any other w, the error should be larger. Therefore:

$$E[\frac{1}{m}\sum_{i=1}^{m}(\tilde{w} \cdot \tilde{x_i} - \tilde{y_i})^2] \leq E[\frac{1}{m}\sum_{i=1}^{m}(\hat{w} \cdot \tilde{x_i} - \tilde{y_i})^2] \tag{25}$$

By equation 24, we have:

$$E[\frac{1}{n}\sum_{i=1}^{n}(\hat{w} \cdot x_i - y_i)^2] \leq E[\frac{1}{m}\sum_{i=1}^{m}(\hat{w} \cdot \tilde{x_i} - \tilde{y_i})^2] \tag{26}$$

In conclusion,

$$E[\hat{R}(\hat{w})] \leq E[R(\hat{w})] \tag{27}$$

To compare, $E[\tilde{R}(\hat{w})]$, $E[\tilde{R}(\tilde{w})]$, and $E[\hat{R}(\hat{w})]$, again, since $\hat{w}$ and $\tilde{w}$ denote the squared training error minimizing decision boundaries based on samples (x_1, x_2), ... , (x_n, x_n) and $(\tilde{x_1}, \tilde{y_1})$, ... , $(\tilde{x_m}, \tilde{y_m})$, respectively and $(\tilde{x_1}, \tilde{y_1})$ is independent of (x_1, x_2) therefore:

$$E[\tilde{R}(\tilde{w})] = E[\hat{R}(\hat{w})] \leq E[\tilde{R}(\hat{w})] \tag{28}$$

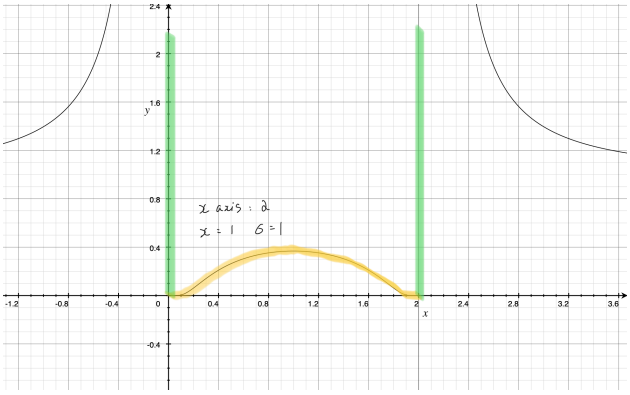

Figure 1: Visualizing the kernel transformation function.

Problem 3

(i)

As the hint given in the question, by visualizing the function, we have figure 1. Here. the x axis represents the α space. Both σ and x are 1, respectively. First, we know σ must be greater than 0 since $|\alpha - x|$ is always > 0. Because $x_1, x_2, ..., x_n$ are distinct points, let's assume

$$\sigma = \frac{1}{4} \min_{i \neq j} |x_i - x_j| \tag{29}$$

to make the interval totally disjoint. Then w can be defined as a function of α:

$$w(\alpha) = \sum_{k=1}^{n} y_k \cdot 1[|\alpha - x_k| < \sigma] \tag{30}$$

Therefore, the predicted class will be solved by the integral of the product:

$$sign\Big(w(\alpha) \cdot \Phi_\sigma(x_k)\Big) = y_k \int_{-\infty}^{+\infty} \Big[\exp(\frac{-1}{1 - \frac{|\alpha - x_k|}{\sigma}^2})\Big],\ satisfies\ |\alpha - x| < \sigma \tag{31}$$

From figure 1 we know that the integral is always greater than 0, therefore:

$$sign\Big(w(\alpha) \cdot \Phi_\sigma(x_k)\Big) = sign(y_k) \tag{32}$$

In conclusion, this mapping can linearly separate *any* binary labeling of the n points.

(ii)

$$\begin{aligned} K(x, x^{'}) &= \Phi(x) \cdot \Phi(x^{'}) \\ &= \int_{|\alpha - x| < \sigma,\ |\alpha - x^{'}| < \sigma} \exp(\frac{-1}{1 - \frac{|\alpha - x|}{\sigma}^2}) \exp(\frac{-1}{1 - \frac{|\alpha - x^{'}|}{\sigma}^2}) d\alpha \\ &= \int_{|\alpha - x| < \sigma,\ |\alpha - x^{'}| < \sigma} \exp\Big(\frac{-1}{1 - \frac{|\alpha - x|}{\sigma}^2} + \frac{-1}{1 - \frac{|\alpha - x^{'}|}{\sigma}^2}\Big) d\alpha \end{aligned} \tag{33}$$

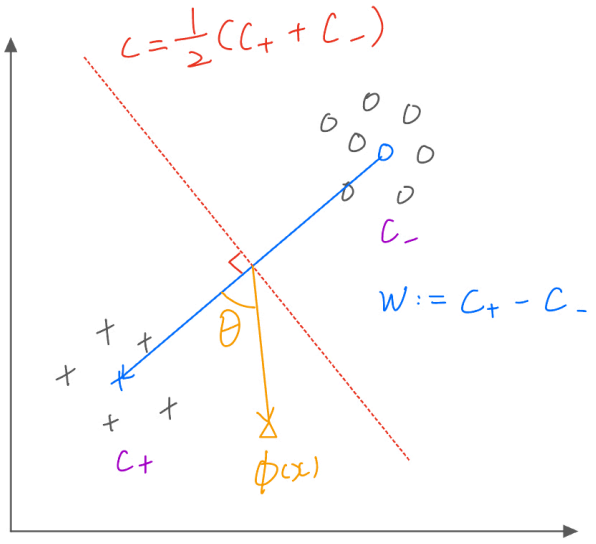

Figure 2: Classification problem in simplified space with w and c defined.

(iii)

(a)

Let's first define a new line which is perpendicular to w and passes the middle point of w.

$$c = \frac{1}{2}(c_+ + c_-) \tag{34}$$

Since $b = \frac{1}{2}(||c_-||^2 - ||c_+||^2)$, we can also have:

$$b = \frac{1}{2}(c_- + c_+)^T(c_- - c_+) \tag{35}$$

Then:

$$\begin{aligned} sign(\langle w, \Phi(x)\rangle + b) &= sign(\langle w, \Phi(x)\rangle + \frac{1}{2}(c_- + c_+)^T(c_- - c_+)) \\ &= sign(\langle w, \Phi(x)\rangle - w^T c) \\ &= sign(\langle w, \Phi(x) - c\rangle) \end{aligned} \tag{36}$$

The sign of the inner product is determined by the angle, therefore we have:

$$sign(\langle w, \Phi(x) - c\rangle) = sign(||w|| \cdot ||\Phi(x) - c|| \cdot \cos(\theta)) \tag{37}$$

We know that $||w||$ and $||\Phi(x) - c||$ terms will always be equal or greater than 0. Therefore, $sign(\langle w, \Phi(x)\rangle + b)$ purely determined by $\cos\theta$. The detailed illustration is shown in figure 2. The decision rule $h(x)$ is based on the distance between $\Phi(x)$ and c_-, c_+. If $\Phi(x)$ is closer to c_+, this means $||\Phi(x) - c_+|| \leq |\Phi(x) - c_-||$, which will give us a $+1$ label. Geometrically, as shown in figure 2, this condition is equal to $0 < \theta < \frac{\pi}{2}$. By applying $sign(||w|| \cdot ||\Phi(x) -$

$c||\cdot\cos(\theta))$, the classifier will also give us a $+1$ label. The opposite also holds true when $||\Phi(x)-c_+|| \geq |\Phi(x)-c_-||$ and it is equivalent to $\frac{\pi}{2} < \theta < \pi$.
Therefore, in conclusion:

$$h(x) = sign(\langle w, \Phi(x)\rangle + b)$$

(b)

$$\begin{aligned}
h(x) &= sign(\langle w, \Phi(x)\rangle + b) \\
&= sign\Big((c_+ - c_-)^T\Phi(x) + b\Big) \\
&= sign\Big(\big[\frac{1}{m_+}\sum_{i:y_i=1}\Phi(x_i) - \frac{1}{m_-}\sum_{i:y_i=-1}\Phi(x_i)\big]^T\Phi(x) + b\Big) \\
&= sign\Big(\big[\frac{1}{m_+}\sum_{i:y_i=1}\Phi(x_i)^T\Phi(x) - \frac{1}{m_-}\sum_{i:y_i=-1}\Phi(x_i)^T\Phi(x)\big] + b\Big) \\
&= sign\Big(\big[\frac{1}{m_+}\sum_{i:y_i=1}K(x_i,x) - \frac{1}{m_-}\sum_{i:y_i=-1}K(x_i,x)\big] + b\Big) \qquad (38)
\end{aligned}$$

Thus, $h(x)$ is efficiently computable.

Appendix

- Q2 uses this article as a reference.
- Q3(iii) uses video as a reference.