CSE – 426 Homework 5

Griffin Kent

1) Let h and f be two hypotheses that map $\boldsymbol{x}$ to $\{0,1\}$. Given the data points $\boldsymbol{x}\sim\mathcal{D}$ for some data distribution $\mathcal{D}$, prove the following:

$$\mathbb{P}_{x\sim\mathcal{D}}\big(h(\boldsymbol{x}) \neq f(\boldsymbol{x})\big) = \mathbb{E}_{x\sim\mathcal{D}}\big[1[h(\boldsymbol{x}) \neq f(\boldsymbol{x})]\big].$$

That is, the probability that h and f do not agree is equal to the expectation of the indicator function $1[h(\boldsymbol{x}) \neq f(\boldsymbol{x})]$.

[*Hints: you can assume that there are only finitely many $\boldsymbol{x}$ and thus $\mathcal{D}$ is a probability mass function. Also use the fact that the indicator function is a binary random variable.*]

(Proof): Since the indicator function $1[h(\boldsymbol{x}) \neq f(\boldsymbol{x})]$ is a binary random variable, it has a Bernoulli distribution taking on a value of 1 if $h(\boldsymbol{x}) \neq f(\boldsymbol{x})$ and 0 if $h(\boldsymbol{x}) = f(\boldsymbol{x})$. The Bernoulli distribution is defined as follows:

$$f(y;p) = p^y q^{1-y}$$

With the expectation

$$\mathbb{E}(y;p) = \sum_{y=0}^{1} y p^y q^{1-y} = p$$

Where p is the probability of y taking on the value of 1 and $q = 1 - p$.

For legibility, let $g(\boldsymbol{x}) \triangleq 1[h(\boldsymbol{x}) \neq f(\boldsymbol{x})]$ represent the binary indicator variable. Thus, the expected value of the indicator function g can be written as

$$\mathbb{E}_{x\sim\mathcal{D}}[g(\boldsymbol{x})]$$

$$= \sum_{g(\boldsymbol{x})=0}^{1} (g(\boldsymbol{x}))\mathbb{P}_{x\sim\mathcal{D}}(g(\boldsymbol{x}) = 1)^{g(\boldsymbol{x})}\mathbb{P}_{x\sim\mathcal{D}}(g(\boldsymbol{x}) = 0)^{1-g(\boldsymbol{x})}$$

$$= \mathbb{P}_{x\sim\mathcal{D}}(g(\boldsymbol{x}) = 1)$$

$$= \mathbb{P}_{x\sim\mathcal{D}}\big(h(\boldsymbol{x}) \neq f(\boldsymbol{x})\big).$$

Completing the proof.■

2) Let x be sampled uniformly from the interval $[-2,2]$. The data are labeled by the function $f(x) = 1[x \geq 0]$. Given three training examples $x^{(1)} = -1$, $y^{(1)} = 0$, $x^{(2)} = 1$, $y^{(2)} = 1$, $x^{(3)} = 1.5$, and $y^{(3)} = 1$, find (without proof) the linear classifier of the form of $h_a(x) = 1[x \geq a]$ with the maximum margin. Then show that the generalization error of this classifier is zero.

[*Hints: since the coefficient of x in $h_a(x)$ is fixed at 1, you just need to pick a value for the parameter a, so that the minimum functional margins of these training examples are maximized.*]

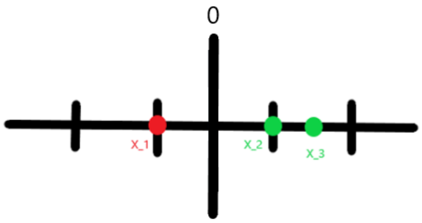
(Solution): The uniform distribution density function over the interval $[-2,2]$ is defined as follows:

$$g(x) = \frac{1}{2-(-2)} = \frac{1}{4}.$$

Likewise, the generalization error of a classifier h given a distribution $\mathcal{D}$ is defined as

$$L_{\mathcal{D},f}(h) = \mathbb{P}_{x\sim\mathcal{D}}\big(h(\boldsymbol{x}) \neq f(\boldsymbol{x})\big) = \mathbb{E}_{x\sim\mathcal{D}}\big[1[h(\boldsymbol{x}) \neq f(\boldsymbol{x})]\big].$$

Since the data are scalars, we have the following graphic:



The linear classifier of the form $h_a(x) = 1[x \geq a]$ that yields the maximum margin is found by maximizing the minimum margins. The functional margins of the three training examples are (For mathematical convenience, denote $y = 0$ as $y = -1$)

$$\hat{\gamma}^{(1)} = y^{(1)}\big(x^{(1)} + b\big) = -1(-1 + b) = 1 - b$$

$$\hat{\gamma}^{(2)} = y^{(2)}\big(x^{(2)} + b\big) = 1(1 + b) = 1 + b$$

$$\hat{\gamma}^{(3)} = y^{(3)}\big(x^{(3)} + b\big) = 1(1.5 + b) = 1.5 + b.$$

Then we can see that the minimum margin will be $\hat{\gamma}^{(1)} = 1 - b$ and thus we would like to maximize this. This will be maximized when $b = 0$. Thus, the linear classifier that maximizes the margin is $h_0(x) = 1[x \geq 0]$, where $a = b = 0$. Since this classifier is the exact same as the labeling function $f(x) = 1[x \geq 0]$, we can see that the generalization error of the hypothesis $h_0(x)$ is

$$L_{\mathcal{D},f}(h_0) = \mathbb{P}_{x\sim\mathcal{D}}\big(h_0(\boldsymbol{x}) \neq f(\boldsymbol{x})\big) = 0.$$

(Since $h_0(x)$ and $f(x)$ are the same classifier, they will never differ in their classifications. Thus, the generalization error will always be 0)

$$\blacksquare$$

3) Let $x, z \in \{1, \ldots, N\}$ where N is an integer greater than 1. Prove that the function $k(x, z) = \min\{x, z\} = \langle \psi(x), \psi(z) \rangle$ for some function ψ that maps from $\{1, \ldots, N\}$ to some higher dimensional Euclidean space $\mathbb{R}^n$. You need to determine n, which is related to N, and the exact mapping ψ. This will prove that $k(x, z)$ is indeed a kernel function.

[*Hints: $min\{1,2\} = 1$ and $min\{2,1\} = 1$ so the min function is symmetric in its two arguments.*]

(Proof): Recall that a kernel function is defined as the following:

$$k(\boldsymbol{x}, \boldsymbol{z}) = \langle \psi(\boldsymbol{x}), \psi(\boldsymbol{z}) \rangle = \psi(\boldsymbol{x})^T \psi(\boldsymbol{z}).$$

If we let $\psi(x)$ be a vector mapping such that $\psi: \mathbb{R} \rightarrow \mathbb{R}^N$, then we can use an indicator function to determine if the iterate $i \in \{1, \ldots, N\}$ is less than or equal to x. Therefore, we can define ψ as

$$\psi(x) = [1[1 \leq x], 1[2 \leq x], \ldots, 1[N \leq x]\ ]$$

Or

$$\psi(x) = \big[1[i \leq x]\big], \quad \forall i \in \{1, \ldots, N\}.$$

Therefore, we can see that the function $k(x, z) = \min\{x, z\}$ can be expressed as the inner product of two vectors $\langle \psi(x), \psi(z) \rangle$, where the mapping ψ is defined above. Therefore, $k(x, z)$ is a kernel function.$\blacksquare$

4) (Kernelizing Linear Regression) Given the MSE loss function of linear regression

$$L\left(\boldsymbol{w};\left\{\left(\boldsymbol{x}^{(i)},y^{(i)}\right)\right\}_{i=1}^{m}\right)=\frac{1}{2}\sum_{i=1}^{m}\left(y^{(i)}-\langle\boldsymbol{w},\boldsymbol{x}^{(i)}\rangle\right)^{2}$$

Let $\boldsymbol{w}=\sum_{j=1}^{m}\alpha_j\boldsymbol{x}^{(j)}$. Rewrite the loss function in terms of the linear kernel $\langle\boldsymbol{x}^{(i)},\boldsymbol{x}^{(j)}\rangle=K_{ij}$. [*Hints: refer to the derivation of the dual problem of SVM using the KKT condition* $\boldsymbol{w}=\sum_{j=1}^{m}\alpha_j\boldsymbol{x}^{(j)}$ *for some dual variables* $\alpha=[\alpha_1,\ldots,\alpha_m]$. *Note that* $K_{ij}=K_{ji}$.]

(Solution): Given $\boldsymbol{w}=\sum_{j=1}^{m}\alpha_j\boldsymbol{x}^{(j)}$, we can rewrite the loss function as the following:

$$L\left(\boldsymbol{w};\left\{\left(\boldsymbol{x}^{(i)},y^{(i)}\right)\right\}_{i=1}^{m}\right)=L\left(\sum_{j=1}^{m}\alpha_j\boldsymbol{x}^{(j)};\left\{\left(\boldsymbol{x}^{(i)},y^{(i)}\right)\right\}_{i=1}^{m}\right)$$

$$=\frac{1}{2}\sum_{i=1}^{m}\left(y^{(i)}-\langle\sum_{j=1}^{m}\alpha_j\boldsymbol{x}^{(j)},\boldsymbol{x}^{(i)}\rangle\right)^{2}=\frac{1}{2}\sum_{i=1}^{m}\left(y^{(i)}-\sum_{j=1}^{m}\alpha_j\langle\boldsymbol{x}^{(j)},\boldsymbol{x}^{(i)}\rangle\right)^{2}$$

$$=\frac{1}{2}\sum_{i=1}^{m}\left(y^{(i)}-\sum_{j=1}^{m}\alpha_j K_{ij}\right)^{2}.$$

■

5) (Graduate Only) Continue working on Question 2, with the same data sampling distribution $\mathcal{D}$ over the interval $[-2,2]$, modify the labeling function $f(x)$, so that the same three training examples remain possible and the same, but the maximum margin classifier found in Question 2 will have a generalization error of $\frac{1}{4}$. Now suppose the labeling function is the same as that in Question 2, can you modify the data distribution $\mathcal{D}$ so that the maximum margin classifier found in Question 2 has a generalization error of $\frac{1}{4}$? If Yes, give a concrete such distribution; if not, give a rigorous proof.
[*Hints: you need to write down the generalization error and then manipulate* f *or* $\mathcal{D}$.]

(Solution): For the first question:

We have the same uniform distribution $\mathcal{D}$ over the interval $[-2,2]$ and maximum margin classifier $h_0(\boldsymbol{x})$. By the properties of the uniform distribution, we also know that the mean of $\mathcal{D}$ is defined as $\mathbb{E}_{\mathcal{D}}(x)=\frac{2+(-2)}{2}=0$. Notice that the interval $[-2,2]$ can be divided into quadrants defined by the five points $[-2,-1,0,1,2]$. Now, if we redefine the labeling function as $f(x)=1[x\geq-1]$, then the maximum margin classifier, $h_0(x)$, that was obtained in Question 2, will correctly classify three of the four quadrants defined by $[-2,-1,0,1,2]$ of any newly generated data; or in other words, it will disagree with the labeling function $\frac{1}{4}$ of the time. Therefore, the generalization error of the classifier $h_0(x)$ will be

$$L_{\mathcal{D},f}(h_0)=\mathbb{P}_{x\sim\mathcal{D}}\left(h_0(\boldsymbol{x})\neq f(\boldsymbol{x})\right)=\frac{1}{4}.$$

For the second question:

We have the same labeling function in Question 2, $f(x) = 1[x \geq 0]$, and maximum margin classifier $h_0(\boldsymbol{x})$. However, there is no way to modify the uniform distribution $\mathcal{D}$ such that the generalization error of the maximum margin classifier $h_0(\boldsymbol{x})$ will ever be $\frac{1}{4}$. This is because the generalization error is calculated strictly based on the misclassification rate of the labeling function $f(x)$ and the hypothesis $h(x)$. Since we know that $f(x)$ and $h(x) = h_0(\boldsymbol{x})$ are the same in this scenario, there will never be a misclassification, thus always yielding a generalization error of 0. Therefore, given $f(x) = 1[x \geq 0]$, $h_0(x) = 1[x \geq 0]$ can never have a generalization error of $\frac{1}{4}$.

■