Machine Learning, Spring 2023
Homework 4

Due on 23:59 April 27, 2023

1 Understanding VC dimension (50 points)

In this part, you need to complete some mathematical proofs about VC dimension. Suppose the hypothesis set

$$\mathcal{H} = \{f(x, \alpha) = \text{sign } (\sin(\alpha x))|, \alpha \in \Re\}$$

where x and f are feature and label, respectively.

- Show that $\mathcal{H}$ cannot shatter the points $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4$. (20 points)

  (Key: Mathematically, you need to show that there exists y_1, y_2, y_3, y_4, for any $\alpha \in \Re$, $f(x_i) \neq y_i, i = 1, 2, 3, 4$, for example, $+1, +1, -1, +1$)

- Show that the VC dimension of $\mathcal{H}$ is ∞. (Note the difference between it and the first question) (30 points)

  (Key: Mathematically, you have to prove that for any label sets $y_1, \cdots, y_m, m \in \mathbb{N}$, there exists $\alpha \in \Re$ and $x_i, i = 1, 2, \cdots, m$ such that $f(x; \alpha)$ can generate this set of labels. Consider the points $x_i = 10^{-i}$...)

1. To show that $\mathcal{H}$ cannot shatter the points $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4$, we need to find a set of labels y_1, y_2, y_3, y_4 such that for any $\alpha \in \mathbb{R}$, we have $f(x_i, \alpha) \neq y_i$ for $i = 1, 2, 3, 4$.

Let $y_1 = y_2 = 1$ and $y_3 = y_4 = -1$. Then, for any $\alpha \in \mathbb{R}$, we have:

$$f(x_1, \alpha) = \text{sign}(\sin \alpha), \quad f(x_2, \alpha) = \text{sign}(\sin 2\alpha), \quad f(x_3, \alpha) = \text{sign}(\sin 3\alpha), \quad f(x_4, \alpha) = \text{sign}(\sin 4\alpha).$$

We can see that for any choice of α, there will always be at least two points with the same sign, so we cannot assign the labels y_1, y_2, y_3, y_4 to the points x_1, x_2, x_3, x_4, respectively. Therefore, $\mathcal{H}$ cannot shatter the points $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4$.

2. To show that the VC dimension of $\mathcal{H}$ is ∞, we need to show that for any $m \in \mathbb{N}$, there exist $x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_m$ such that $\mathcal{H}$ can shatter $\{x_1, x_2, \ldots, x_m\}$ with labels $\{y_1, y_2, \ldots, y_m\}$.

Let $x_i = 10^{-i}$ for $i = 1, 2, \ldots, m$ and let $y_1 = y_2 = 1$ and $y_3 = y_4 = \cdots = y_m = -1$. We claim that for any m, $\mathcal{H}$ can shatter $\{x_1, x_2, \ldots, x_m\}$ with labels $\{y_1, y_2, \ldots, y_m\}$. To see this, let $\alpha = \frac{\pi}{2}$, then we have

$$f(x_i, \alpha) = \text{sign}(\sin \alpha x_i) = \begin{cases} 1, & \text{if } i = 1, 2, \\ -1, & \text{if } i > 2. \end{cases}$$

Therefore, we can assign the label y_i to the point x_i for $i = 1, 2, \ldots, m$ using the function $f(x, \alpha)$ with $\alpha = \frac{\pi}{2}$. Since we have shown that $\mathcal{H}$ can shatter any set of m points, we conclude that the VC dimension of $\mathcal{H}$ is infinite.

2 Bias-variance decomposition (50 points)

When there is noise in the data, $E_{out}(g^{(\mathcal{D})}) = \mathbb{E}_{\mathbf{x},y}\left[(g^{(\mathcal{D})}(\mathbf{x}) - y(\mathbf{x}))^2\right]$, where $y(\mathbf{x}) = f(\mathbf{x}) + \epsilon$. If ϵ is a zero-mean noise random variable with variance σ^2, show that the bias-variance decomposition becomes

$$\mathbb{E}_{\mathcal{D}}\left[E_{out}(g^{(\mathcal{D})})\right] = \sigma^2 + \text{bias} + \text{var}$$

The bias-variance decomposition states that the expected out-of-sample error of a model can be decomposed into three parts: bias, variance, and irreducible error. When there is noise in the data, the expected out-of-sample error becomes

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[E_{out}(g^{(\mathcal{D})})\right] &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{x},y}\left[(g^{(\mathcal{D})}(\mathbf{x}) - y(\mathbf{x}))^2\right]\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})])^2\right]\right] + \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[(\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})] - y(\mathbf{x}))^2\right]\right] \\
&= \mathrm{Var}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})] + \mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})])^2\right] + \sigma^2 \\
&= \mathrm{Var}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})] + \mathrm{Bias}^2[g^{(\mathcal{D})}(\mathbf{x})] + \sigma^2,
\end{aligned}$$

where the first line follows from the definition of E_{out}, the second line follows from the law of total expectation, the third line uses the definition of bias and variance, and the fourth line applies the bias-variance decomposition to the second term. Note that $\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})]$ is the expected value of $g^{(\mathcal{D})}(\mathbf{x})$ over all possible training sets $\mathcal{D}$, and $\mathrm{Bias}[g^{(\mathcal{D})}(\mathbf{x})]$ is the difference between $\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})]$ and $f(\mathbf{x})$. Therefore, we have shown that the expected out-of-sample error of a model with noisy data can be decomposed into bias, variance, and irreducible error, where the irreducible error is given by the variance of the noise term.