Indian Institute of Technology Kharagpur

Department of Computer Science and Engineering

CS60050 : Machine Learning

Spring 2021 | Short Test 2 (ALL Q&A) | Marks : 20

Date : 15-Mar-2021 (Monday) | Time : 8:30pm-9:00pm (30 min)

---

Question-1: Which of the following are possible growth functions $m_H(N)$ for some hypothesis set (N = number of training points/examples)? [Marks = 3]

Choose ALL the correct options from the following.

(i) $m_H(N) = 1 + N$

(ii) $m_H(N) = 1 + N + N(N-1)/2$

(iii) $m_H(N) = 1 + N + N(N-1)(N-2)$

(iv) $m_H(N) = 2^N$

(v) $m_H(N) = 2^{\lfloor\sqrt{N}\rfloor}$

(vi) $m_H(N) = 2^{\lfloor N/2\rfloor}$

Note: $x = \lfloor n\rfloor$ (called the floor of n) is the highest integer value with $x \le n$

Answer-1: (i) , (ii) , (iv)

Explanation:

We have only two cases for the growth function (let VC-dimension = d): either $d = \infty$ (infinite) and $m_H(N) = 2^N$ for all N, or d is finite and $m_H(N) \le N^d + 1$.

(i) If $m_H(N) = 1 + N$, we have $d = 1$ (as, $m_H(2) = 3 < 2^2$). So, $m_H(N) \le N^1 + 1$ for all N, which is obviously the case here. In conclusion, $m_H(N) = 1 + N$ is a possible growth function.

(ii) If $m_H(N) = 1 + N + N(N-1)/2$, we have $d = 2$ (as, $m_H(3) = 7 < 2^3$). So, $m_H(N) \le N^2 + 1$ for all N, which is also the case as $N \ge 1$. In conclusion, $m_H(N) = 1 + N + N(N-1)/2$ is a possible growth function.

(iii) If $m_H(N) = 1 + N + N(N-1)(N-2)$, we have $d = 1$ (as, $m_H(2) = 3 < 2^2$). Consequently, it must be the case that $m_H(N) \le N^1 + 1$ for all N, which is not true (for N = 3 for example). In conclusion, $m_H(N) = 1 + N + N(N-1)(N-2)$ is NOT a possible growth function.

(iv) Obviously $m_H(N) = 2^N$ is a possible growth function when $d = \infty$ (infinity).

(v) If $m_H(N) = 2^{\lfloor\sqrt{N}\rfloor}$, we have $d = 1$ (as, $m_H(2) = 2 < 2^2$). Consequently, it must be the case that $m_H(N) \le N^1 + 1$ for all N, which is not true (for N = 25 for example). In conclusion, $m_H(N) = 2\lfloor\sqrt{N}\rfloor$ is NOT a possible growth function.

(vi) If $mH(N) = 2^{\lfloor N/2\rfloor}$, we have $d = 0$ (as, $mH(1) = 1 < 2^1$). Consequently, it must be the case that $m_H(N) \le N^0 + 1 = 2$ for all N, which is not true (for N = 4 for example). In conclusion, $m_H(N) = 2\lfloor N/2\rfloor$ is NOT a possible growth function.

---

Question-2: Suppose $m_H(N) = N + 1$. Determine the Generalization Bound ($\Omega$) for Eout with at least 90% probability (confidence) when the number of training examples are 10000. [Marks = 2]

(In case of Real numbers as answer, write the approximated value upto THREE decimal places after point.)

Answer-2: $\Omega = 0.1042782$

Explanation:
Here, $1-\delta = 0.9$, $N = 100$, and $m\_H(N) = N + 1$.
We know that, $E_{out} \le E_{in} + \Omega$,
where Generalization Bound, $\Omega = \sqrt{(8/N)\ln(4.m\_H(2N)/\delta)}$.
So, $\Omega = \sqrt{(8/10000)\ln(4.(2.10000+1)/0.1)} = 0.1042782$.

---

Question-3: For an hypothesis set (H) having break point 11, what is the minimum sample size (i.e. number of training points/examples) do you need (as prescribed by the generalization bound) to have at least 95% probability (confidence) that your generalization error is at most 0.05? [Marks = 2]
Choose the correct option from the following.
(i) 1000
(ii) $2.57251 \times 10^5$
(iii) $4.52957 \times 10^5$
(iv) $2^{10} + 1$

Answer-3: (iii) $4.52957 \times 10^5$

Explanation:
Note that, the generalization error is bounded by $\Omega = \sqrt{(8/N)\ln(4.m\_H(2N)/\delta)}$. So, it suffices to make $\sqrt{(8/N)\ln(4.m\_H(2N)/\delta)} \le \varepsilon$. It follows that, $N \ge \sqrt{(8/\varepsilon^2)\ln(4.m\_H(2N)/\delta)}$ suffices to obtain generalization error at most $\varepsilon$ (with probability/confidence at least $1-\delta$). This gives an implicit bound for the sample complexity N, since N appears on the both sides of the inequality. If we replace $m\_H(2N)$ by its polynomial upper bound based on VC-dimension (d), we get the final similar bound as,
This implies, $N \ge \sqrt{(8/\varepsilon^2)\ln(4.((2N)^d+1)/\delta)}$

So, as per above formula, we have the following implicit bound for the sample complexity N (with break point $k = 11$, so VC-dimension $d = 10$, $\varepsilon = 0.05$, and $1-\delta = 0.95$ implying $\delta = 0.95$),
$N \ge \sqrt{(8/(0.05)^2)\ln(4.((2N)^{10}+1)/(0.05))}$

To determine N, we will use an iterative process with an initial guess of $N = 1000$ in the RHS. We get
$N \ge \sqrt{(8/(0.05)^2)\ln(4.((2.1000)^{10}+1)/(0.05))} \approx 2.57251 \times 10^5$.

We then try the new value $N = 2.57251 \times 10^5$ in the RHS and iterate this process, rapidly converging to an estimate of $N \approx 4.52957 \times 10^5$.

---

Question-4: Consider a simplified learning scenario. Assume that, the input dimension is one. Assume that, the input variable x is uniformly distributed in the interval [-1, +1]. The data set consists of 2 points $\{x_1, x_2\}$ and assume that the target function is $y = f(x) = x^2$. Thus, the full data set is $D = \{ (x_1, x_1^2), (x_2, x_2^2) \}$. The learning algorithm returns the line fitting these two points as g (the hypothesis set, H, consists of functions of the form $h(x) = ax+b$). We are interested in the test performance (Eout) of our learning system with respect to the squared error measure, the bias and the variance.
Determine the following metrics. [Marks = 2 x 4 = 8]
(i) average hypothesis function $g'(x)$,
(ii) out-of-sample error (Eout),
(iii) bias (bias), and
(iv) variance (var).

(In case of Real numbers as answer, write the approximated value upto THREE decimal places after point.)

Answer-4: (i) 0 , (ii) 0.533 , (iii) 0.2 , (iv) 0.333

Explanation:
(i) We give the analytic expression for the average hypothesis function g'(x) below. We have,

$$\begin{aligned}
g(x) &= E_D[g(x)] \\
&= E_D[(y_1 - y_2)x/(x_1 - x_2) + (x_1y_2 - x_2y_1)/(x_1 - x_2)] \\
&= 1/4 \int_{-1}^{1}\int_{-1}^{1} (x_1^2 - x_2^2)/(x_1 - x_2)dx_1dx_2 \;.\; x \\
&\quad + 1/4 \int_{-1}^{1}\int_{-1}^{1} (x_1x_2^2 - x_2x_1^2)/(x_1 - x_2)dx_1dx_2 \\
&= 1/4 \int_{-1}^{1}\int_{-1}^{1} (x_1 + x_2)dx_1dx_2 \;.\; x \\
&\quad - 1/4 \int_{-1}^{1}\int_{-1}^{1} (x_1x_2)dx_1dx_2 \\
&= 1/4 \;.\; 0 - 1/4 \;.\; 0 = 0
\end{aligned}$$

(ii) To compute $E_D[Eout]$, we will first determine Eout, we get,

$$\begin{aligned}
Eout &= E_x[(g(x) - f(x))^2] = E_x[(ax + b - x^2)] \\
&= E_x[x^4] - 2a \;.\; E_x[x^3] + (a^2 - 2b) \;.\; E_x[x^2] + 2ab \;.\; E_x[x] + b^2 \\
&= 1/4 \int_{-1}^{1} x^4dx - 2a \;.\; \int_{-1}^{1} x^3dx + (a^2 - 2b) \;.\; \int_{-1}^{1} x^2dx + 2ab \;.\; \int_{-1}^{1} x^2dx + b^2 \\
&= 1/5 + (a^2 - 2b)/3 + b^2
\end{aligned}$$

Then, we take the expectation with respect to D to get the test performance.
Since $x_1^2 = ax_1 + b$ and $x_2^2 = ax_2 + b$, which gives as solution $a = (x_1 + x_2)$ and $b = (-x_1x_2)$.
So, we replace a and b by $(x_1 + x_2)$ and $(-x_1x_2)$ respectively, we get,

$$\begin{aligned}
E_D[Eout] &= 1/5 + (1/3) \;.\; E_D[(x_1 + x_2)^2 + 2x_1x_2] + E_D[x_1^2x_2^2] \\
&= 1/5 + (1/3).(1/4)\int_{-1}^{1}\int_{-1}^{1} (x_1^2 + x_2^2 + 4x_1x_2)dx_1dx_2 \\
&\quad + 1/4 \int_{-1}^{1}\int_{-1}^{1} x_1^2x_2^2dx_1dx_2 \\
&= 1/5 + (1/3).(1/4).(8/3) + (1/4).(4/9) = 8/15
\end{aligned}$$

(iii) To compute bias, we first have, $bias(x) = (g'(x) - f(x))^2 = f(x)^2 = x^4$;
then we get, bias (bias) $= E_x[x^4] = 1/2 \int_{-1}^{1} x^4dx = 1/5$

(iv) Finally, we compute the variance, we first have,

$$\begin{aligned}
var(x) &= E_D[(g(x) - g'(x))^2] = E_D[a^2x^2 + 2abx + b^2] \\
&= E_D[a^2] \;.\; x^2 + 2 \;.\; E_D[ab] \;.\; x + E_D[b^2] \\
&= E_D[(x_1 + x_2)^2] \;.\; x^2 - 2 \;.\; E_D[(x_1 + x_2)x_1x_2] \;.\; x + E_D[x_1^2x_2^2] \\
&= E_D[x_1^2 + 2x_1x_2 + x_2^2] \;.\; x^2 - 2 \;.\; E_D[x_1^2x_2 + x_1x_2^2] \;.\; x + E_D[x_1^2x_2^2] \\
&= 1/4 \int_{-1}^{1}\int_{-1}^{1} (x_1^2 + 2x_1x_2 + x_2^2)dx_1dx_2 \;.\; x^2 \\
&\quad - 2/4 \int_{-1}^{1}\int_{-1}^{1} (x_1^2x_2 + x_1x_2^2)dx_1dx_2 \;.\; x + 1/4 \int_{-1}^{1}\int_{-1}^{1} x_1^2x_2^2dx_1dx_2 \\
&= (1/4).(4/3 + 0 + 4/3) \;.\; x^2 - 0.\; x + (1/4).(4/9) = 2x^2/3 + 1/9;
\end{aligned}$$

then we get, variance (var) $= E_x[2x^2/3 + 1/9]$
$= (2/3).(1/2) \int_{-1}^{1} x^2dx + 1/9 = 1/3$

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Question-5: Consider the feature transform $z = [L_0(x), L_1(x), L_2(x)]^t$ with Legendre polynomials and the linear model $h(x) = w^tz$. For the regularized hypothesis with $w = [+1, -1, +1]^t$, what is h(x) explicitly as a function of x? [Marks = 2]
(Notation: $[..]^t$ denotes transpose of the matrix [..])
Choose the correct option from the following.
(i) $1 - x$
(ii) $(3/2)x^2 - x + 1/2$
(iii) $3x^2 - x$
(iv) $(5/2)x^3 - (3/2)x^2 - (1/2)x + 1/2$

Answer-5: (ii) $(3/2)x^2 - x + 1/2$
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Explanation:
$L_0(x) = 1$, $L_1(x) = x$, $L_2(x) = (1/2).(3x^2 - 1)$

We may write $h(x) = [\ +1\ -1\ +1\ ] \begin{bmatrix} L_0(x) \\ L_1(x) \\ L_2(x) \end{bmatrix} = L_0(x) - L_1(x) + L_2(x)$

$= 1 - x + (1/2).(3x^2 - 1)$

$= (3/2)x^2 - x + 1/2$

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Question-6: You have a data set with 100 data points. You have 100 models each with VC dimension 10. You set aside 25 data points for validation. You select the model which produced minimum validation error of 0.25. What is the bound on the out-of-sample error for this selected function/model? [Marks = 2]
Choose the correct option from the following.
(i) $E_{out}(g_m^*) \leq 0.25 + \sqrt{(1/50).\ln(200/\delta)}$ with probability $\geq (1-\delta)$
(ii) $E_{out}(g_m^*) \leq 0.25 + \sqrt{(1/25).\ln(100/\delta)}$ with probability $\geq (1-\delta)$
(iii) $E_{out}(g_m^*) \leq 0.25 + \sqrt{\ln(100)/25}$
(iv) $E_{out}(g_m^*) \leq 0.25 + \sqrt{\ln(200)/50}$

Answer-6: (i) $E_{out}(g_m^*) \leq 0.25 + \sqrt{(1/50).\ln(200/\delta)}$ with probability $\geq (1-\delta)$

Explanation:
Here, we have a data set with $N = 100$ points and a validation set of $K = 25$ points. We consider $M = 100$ models $H_1, H_2, \ldots, H_{100}$ each with VC-dimension $d = 10$. In the first case, each model $H_m$ gives birth to a final hypothesis $g_m^-$ generated on the $N - K = 75$ training points; from these hypotheses, we select the one with the minimum validation error $g_m^{-*}$ of 0.25. We know that,
$E_{out}(g_m^*) \leq E_{out}(g_m^{-*}) \leq E_{val}(g_m^{-*}) + \sqrt{(1/2K)\ln(2M/\delta)}$ with probability $\geq (1-\delta)$
where $g_m^*$ is the chosen final hypothesis trained on the entire data set, since we selected our final hypothesis $g_m^{-*}$ from a finite hypothesis set $H_{val} = \{g_1^-, g_2^-, \ldots, g_{100}^-\}$. So, a bound on the out-of-sample error is given by,
$E_{out}(g_m^{-*}) \leq E_{val}(g_m^{-*}) + \sqrt{(1/2K)\ln(2M/\delta)}$
$= 0.25 + \sqrt{(1/50).\ln(200/\delta)}$ with probability $\geq (1-\delta)$
implies, $E_{out}(g_m^*) \leq 0.25 + \sqrt{(1/50).\ln(200/\delta)}$ with probability $\geq (1-\delta)$

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Question-7: Regarding bias and variance, which of the following statements are TRUE? (Here 'high' and 'low' are relative to the ideal model.) [Marks = 1]
Choose ALL the correct options from the following.
(i) Models which overfit have a high bias.
(ii) Models which overfit have a low bias.
(iii) Models which underfit have a high variance.
(iv) Models which underfit have a low variance.

Answer-7: (ii) and (iv)

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~