Machine Learning -- Assignment 2

Omkar Nibandhe

Question 1

Test hypothesis h has errors r = 240 on sample S of n =800.

$$error_S(h) = \frac{r}{n}$$

$$= \frac{240}{800}$$

$$\boldsymbol{error_s(h) = 0.3}$$

The standard deviation for $error_S(h) = \sqrt{\frac{r*(1-r)}{n}}$

$$= \sqrt{\frac{0.3*0.7}{800}}$$

$$Standard\ deviation\ for\ \boldsymbol{error_s(h)} = \ \boldsymbol{0.0162}$$

Example in class had $n = 40$ and $r = \ 12$

$$error_S(h) = \frac{12}{40} = 0.3$$

Thus, standard deviation for $error_S(h) = \sqrt{\frac{0.3*0.7}{40}} = 0.0725$

As no other information is given, most probable value of $error_D(h)$ is $error_S(h)$. this also assumes standard unbiased estimator for $error_D(h)$ and data sample S is independent of discrete-valued hypothesis h.

Question 2

Sample S contains n = 100.

Incorrect classification, r = 100 – correct classification.

$$r = 100 - 79.$$

Error, r = 21.

95% confidence Implies, $N\% = 95, Z_{0.95} = 1.96$

Standard deviation $= \sqrt{\frac{error_S(\mathrm{h})*(1-error_S(\mathrm{h}))}{n}}$

$$= \sqrt{\frac{(0.21)*(1-0.21)}{100}} = \boldsymbol{0.0407}$$

$$error_D(h) = \ error_S(h) \ \mp \ Z_N.\sqrt{\frac{error_S(\mathrm{h})*(1-error_S(\mathrm{h}))}{n}}$$

$= 0.21 \mp (1.96) * (0.0407)$

$\boldsymbol{error_D(h) = 0.21 \mp 0.0798}$

Upper bound = 0.21+0.0798 = 0.130

Lower bound = 0.21 – 0.0798 = 0.2898

Question 3

Range of interval, L = 0.3 and U = 0.6

Therefore, midpoint, M = $(L + U)/2 = 0.45 = P$

And (1 - p) = 1 – 0.45 = 0.55

Lower bound, $L = p - z\sqrt{\frac{p(1-p)}{n}}$ upper bound, $U = p + z\sqrt{\frac{p(1-p)}{n}}$

Confidence interval width is $U - L = 2z\sqrt{\frac{p(1-p)}{n}}$

95% of two level confidence, we use $Z_{0.975}$ = 1.96

Minimum number of examples $n \geq \frac{4*z^2*p*(1-p)}{(U-L)^2}$

$$n \geq \frac{4*(1.96)^2*0.45*0.55}{(0.1)^2}$$

$$n \geq 380.3$$

Thus minimum number of examples needed to collect should be **greater than or equal to 380**.

Question 4

$r = 10 , n = 75$

90% 2-sided true error:

$error_D(h) = error_S(h) \mp Z_N.\sqrt{\frac{error_S(\mathrm{h})*(1-error_S(\mathrm{h}))}{n}}$

$= \frac{10}{75} \mp (1.64) *.\sqrt{\frac{(0.133)*(0.867)}{75}}$

$= 0.133 \mp (1.64) * 0.0392$

$\boldsymbol{error_D(h) = 0.133 \mp 0.06429}$

95% 1-sided true error:

Machine Learning -- Assignment 2

Omkar Nibandhe

$error_D(h) = error_s(h) + Z_{.90}.\sqrt{\frac{error_s(\mathrm{h})*(1-error_s(\mathrm{h}))}{n}}$ ...consider only Upper bound

$= 0.133 + (1.64) * 0.0392$

$\boldsymbol{error_D(h)} = 0.133 + 0.064$ = 0.197

80% 1-sided true error:

$error_D(h) = error_s(h) + Z_{.80}.\sqrt{\frac{error_s(\mathrm{h})*(1-error_s(\mathrm{h}))}{n}}$ ...consider only Upper bound

$= 0.133 + (1.28) * 0.0392$

$\boldsymbol{error_D(h) = 0.133 + 0.0501}$ **= 0.1831**

Question 5

$\boldsymbol{V_c(H) = 3}$ $\boldsymbol{V_c}$ Dimension for linear separator in 2-dimensional.

$\delta = (100 - 90)\% = 0.1 \; \epsilon = 0.05$

For Upper bound,

$m \geq \frac{1}{\epsilon}(4\, log_2\left(\frac{2}{\delta}\right) + 8\, V_c(h) log_2\left(\frac{13}{\epsilon}\right))$

$m \geq \frac{1}{0.05}(4\, log_2\left(\frac{2}{0.1}\right) + 8\, V_c(h) log_2\left(\frac{13}{0.05}\right))$

$m \geq \frac{1}{0.05}(4 * 4.321 + 8 * 3 * 8.022)$

$m \geq 20 * (17.28 + 192.53)$

$m \geq 4196.33$

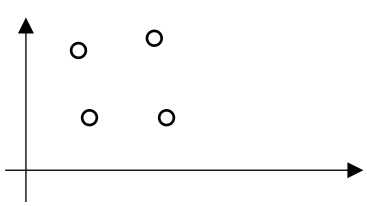
This bound does not seem to be realistic.

Because a hyperplane in 2 dimension is a line which has to be defined by 2 points. So if 90% confidence with at most 5% error will break the line. So this might be the reason to sound unrealistic.

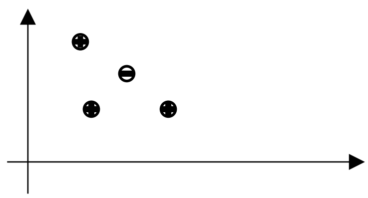
Question 6

(a) For rectangle,

Consider point in $XY - plane$ and $a < x < b$ and $c < y < d$



Consider that 3 points are positive and a point is negative in the following case. Here it is not possible to shatter the points in$XY - plane$.



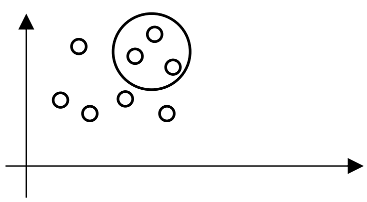
Hence, the $\boldsymbol{V_c(h) = 4}$

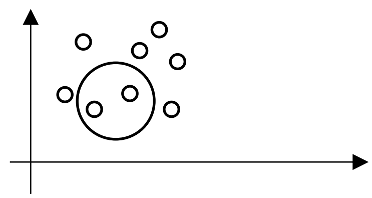
(b) For circle,

Points in circle are positive and outside is negative.

$H_c = circle\ in\ XY - plane$

$V_c dimension$ for circle in $XY - plane$ is at least 3, as 3 points make up non-degenerated triangle.





It is possible to shatter 1,2 or 3 positive points so,

$\boldsymbol{V_c = 3}$

Question 7

(a)

Region bounded by point (0,0) and (n,n) in the interval (0,100) implies that n = 101.

δ with probability 95% = 100 – 95 = 0.05

$\boldsymbol{\delta = 0.05}$

$|h| = (\frac{n(n+1)}{2})^2$

$\epsilon = 0.15$

Size of hypothesis m is calculated as:

$m \geq \frac{1}{\epsilon}\left[\ln|H| + \ln\left(\frac{1}{\delta}\right)\right]$

$m \geq \frac{1}{0.15}\left[\ln(\frac{n(n+1)}{2})^2 + \ln\left(\frac{1}{0.05}\right)\right]$

$m \geq \frac{1}{0.15}\left[\ln(\frac{101(101+1)}{2})^2 + \ln(20)\right]$

$m \geq \frac{1}{0.15}[\ln(5151)^2 + \ln(20)]$

$\boldsymbol{m \geq 133.93}$

(b)

Region bounded by point (0,100) a read values, $n = \infty$

$V_c\,(h) \leq 4\; if\; a < x < b : z = 1\; else\; z = 0$

$a < x < b : z = 0\; else\; z = 1$

$if\; c < y < d : z = 1\; else\; z = 0$

$c < y < d : z = 0\; else\; z = 1$

$\boldsymbol{\delta = 0.05}$

$\epsilon = 0.15$

$m \geq \frac{1}{\epsilon}\,(4\,log_2\left(\frac{2}{\delta}\right) + 8\,V_c(h) log_2\left(\frac{13}{\epsilon}\right))$

$m \geq \frac{1}{0.15}\,(4\,log_2\left(\frac{2}{0.05}\right) + 8 * 4 * log_2\left(\frac{13}{0.15}\right))$

$m \geq \frac{1}{0.15}\,(4 * 5.321 + 8 * 4 * 6.437)$

$m \geq \frac{1}{0.15}\,(227.26)$

$\boldsymbol{m \geq 1515.12}$

Question 8

(a)

The tree has depth 2 and 4 leaves in total.

So, syntactically distinct trees are $\boldsymbol{2^4 n(n-1)}$

$f(n) = 2^4 n(n-1)$

$H_{rd2} = f(x)$

```
            ( )
          /     \
       ( )       ( )
      /   \     /   \
   ( )    ( )  ( )    ( )
```

(b)

$$|h_{rd2}| = 2^4 n(n-1)$$

Upper bound for number of examples, m with $error = \epsilon$ & $confidence = \delta$

$$m \geq \frac{1}{\epsilon}\left[\ln|h_{rd2}| + \ln\left(\frac{1}{\delta}\right)\right]$$

$$m \geq \frac{1}{\epsilon}\left[\ln 2^4 n(n-1) + \ln\left(\frac{1}{\delta}\right)\right]$$

Question 9

(a)

$n = 100. N\% = 95\%$

$error_D(h)$ should be calculated to find the error.

$error_S(h) = \frac{r}{n} = \frac{0}{100}$ ...given

As $error_S(h) = 0$, standard deviation = 0 and true error is also 0.

This implies that it is difficult to calculate the true error with 95% probability as $error_D(h) = 0$

$$error_D(h) = error_S(h) \mp Z_N . \sqrt{\frac{error_S(\mathrm{h}) * (1 - error_S(\mathrm{h}))}{n}}$$

$$Z_N = 1.96$$

$$error_D(h) = error_S(h) \mp 1.96. \sqrt{\frac{error_S(\mathrm{h}) * (1 - error_S(\mathrm{h}))}{n}}$$

$$\boldsymbol{error_D(h) = 0}$$

$$p[error_D(h) > error_D(h) + \epsilon] \leq e^{-2m\epsilon^2}$$

$$p[error_D(h) > error_D(h) + \epsilon] \geq 95\%$$

Hence it is difficult to calculate the true error since$error_S(h) = 0$ it is difficult to find the interval.

Omkar Nibandhe

(b)

$$n = 100 \,\&\, r = 30$$

$$error_S(h) = \frac{r}{n} = 0.3$$

$$N\% = 90\% \; implies \; Z_{.90} = 1.96$$

$$error_D(h) = \; error_S(h) \; \mp \; Z_N . \sqrt{\frac{error_S(\mathrm{h})*(1-error_S(\mathrm{h}))}{n}}$$

$$error_D(h) = \; 0.3 \; \mp \; 1.96. \sqrt{\frac{0.3*0.7}{100}}$$

$$error_D(h) = \; 0.3 \; \mp \; 0.0898$$

Upper bound = 0.3898 & Lower bound = 0.2101

The interval in which this true error will fall is 0.2101 to 0.3898