# CIS 419/519: Homework 5

{Shruti Sinha}

12/12/2018(2 late days taken)

Although the solutions are my own, I consulted with the following people while working on this homework: {Hemanth, Andres}

1. (a) i. B)SVM: index of examples in training data
   ii. D)Soft SVM, slack variable is squared

   (b) i. A) O(logN)
   ii. Considering that X is a subset of the form X $a+bi$: i= 0,1,2.., we know a and b both can take values from 0,1,2...N. Therefore,we know that the concept class is of $N^2$.
   We also know that order of magnitude of VC dimension is :
   $VC_{dim} = log|C|$, where C is the concept class.
   $|C| = N^2$
   Therefore, $VC_{dim} = O(log(N^2))$
   $VC_{dim} = O(2log(N)) = O(logN)$
   Therefore we choose A) as our answer.
   iii. B) Simpler classifiers are sufficient for this.

   (c) i. B)Halt after some number of iterations
   ii. It will halt after some iterations.
   The description says, that "Adaboost is consistent with m examples".This implies that the algorithm would stop making mistakes and terminate after m examples.
   Now we know that the training error of the hypothesis h generated by the Adaboost algorithm is bounded by :$e^{-2\gamma^2 T}$
   Thus,

$$e^{-2\gamma^2 T} < \frac{1}{m} \tag{1}$$

   Pluggin in $m = e^{14}$ and $\gamma = \frac{1}{4}$

$$e^{-2\times(\frac{1}{4})^2 T} < \frac{1}{e^{14}} \tag{2}$$

$$e^{\frac{-1}{8}T} < \frac{1}{e^{14}} \tag{3}$$

cross multiplying,

$$e^{\frac{-1}{8}T+14} < 1 \tag{4}$$

Taking log on both sides,

$$14 - \frac{1}{8}T < 0 \tag{5}$$

$$14 < \frac{1}{8}T \tag{6}$$

$$T > 112 \tag{7}$$

Thus from the above equation we can see, that if T is greater than 112, the Adaboost algorithm would terminate.
Hence we pick B.

2. (a) C)15, the largest number of independent parameters for these 4 variables is $2^N - 1$, here N = 4.
   $2^4 - 1 = 15$

   (b) The joint probability distribution is,
   $P(A, B, C, D) = P(A) \times P(B|A) \times P(C|A) \times P(D|B, C)$

   (c) 1, 7, 9, 14
   1)P(A=1)
   7)$P(B = 1|A = i)$
   9)$P(C = 1|A = i)$
   14)$P(D = 1|B = i, C = j)$

   (d) $P(A = 1) = \frac{3}{5}$
   $P(B = 1|A = 1) = \frac{1}{3}$
   $P(C = 1|A = 1) = \frac{2}{3}$
   $P(D = 1|B = 1, C = 1) = \frac{2}{3}$
   $P(D = 1|B = 1, C = 0) = \frac{1}{2}$
   $P(D = 1|B = 0, C = 1) = \frac{1}{3}$
   $P(D = 1|B = 0, C = 0) = \frac{1}{2}$

   $P(A = 0) = \frac{2}{5}$
   $P(B = 1|A = 0) = \frac{3}{4}$
   $P(C = 1|A = 0) = \frac{2}{4}$

   (e) The likelihood is, Probability of example 1 x Probability of example 2

   $L = \prod_{i=1}^{2} P(A_i, B_i, C_i, D_i) = P(A_1 = 1)P(B_1 = 0|A_1 = 1)P(C_1 = 1|A_1 = 1)P(D_1 = 0|B_1 = 0, C_1 = 1) \times P(A_2 = 0)P(B_2 = 1|A_2 = 0)P(C_2 = 0|A_2 = 0)P(D_2 = 1|B_2 = 1, C_2 = 0)$

   $$L = \prod_{i=1}^{2} P(A_i, B_i, C_i, D_i) = \frac{3}{5} \times (1 - \frac{1}{3}) \times \frac{2}{3} \times (1 - \frac{1}{3}) \times (1 - \frac{3}{5}) \times \frac{3}{4} \times (1 - \frac{2}{4}) \times \frac{1}{2}$$

$$L = \frac{3}{5} \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{5} \times \frac{3}{4} \times \frac{2}{4} \times \frac{1}{2} = \frac{1}{75}$$

(f) $P(D = 1|A = 1) =$
$P(D = 1|B = 1, C = 1)P(B = 1|A = 1)P(C = 1|A = 1)$
$+ P(D = 1|B = 1, C = 0)P(B = 1|A = 1)P(C = 0|A = 1)$
$+ P(D = 1|B = 0, C = 1)P(B = 0|A = 1)P(C = 1|A = 1)$
$+ P(D = 1|B = 0, C = 0)P(B = 0|A = 1)P(C = 0|A = 1)$

$$P(D = 1|A = 1) = \frac{2}{3}\times\frac{1}{3}\times\frac{2}{3}+\frac{1}{2}\times\frac{1}{3}\times\frac{1}{3}+\frac{1}{3}\times\frac{2}{3}\times\frac{2}{3}+\frac{1}{2}\times\frac{2}{3}\times\frac{1}{3} = \frac{25}{54} = 0.463$$

3. (a) C)(0,1)
Since, $\sigma_x = \frac{1}{1+e^{-x}}$, the range for this function is (0,1). Hence, the output can have range of (0,1).

(b) A and D
a) Correct,
$\lim_{y\to 0^+} L(y = 0, y) = \lim_{y\to 0^+} -0 \ln 0 - (1 - 0)\ln(1 - 0) = 0$
which is equal to 0
b)Incorrect,
$\lim_{y\to 0^+} L(y = 0, y) = \lim_{y\to 0^+} -0 \ln 0 - (1 - 0)\ln(1 - 0) = 0$
which is not equal to $\infty$
c)Incorrect,
$\lim_{y\to 0^+} L(y = 1, y) = \lim_{y\to 0^+} -1 \ln 0 - (1 - 1)\ln(1 - 0)$
which is not equal to 0
d)Correct,
$\lim_{y\to 1} L(y = 0, y) = \lim_{y\to 0^+} -0 \ln 1 - (1-0)\ln(1-1) = -1 \times -\infty = \infty$
which is equal to $\infty$

(c) B) $Err = \frac{1}{m}\sum_{i=1}^{m} -y^{(i)} \ln(NN(x_1^{(i)}, x_2^{(i)})) - (1-y^{(i)}) \ln(1-NN(x_1^{(i)}, x_2^{(i)}))$

(d)

$$\frac{\partial_E}{\partial_{w1}} = \frac{\partial_E}{\partial_o} \times \frac{\partial_o}{\partial_{net_{56}}} \times \frac{\partial_{net_{56}}}{\partial_{o_{13}}} \times \frac{\partial_{o_{13}}}{\partial_{net_{13}}} \times \frac{\partial_{net_{13}}}{\partial_{w1}}$$

$$\frac{\partial net_{13}}{\partial w_1} = \frac{\partial_{(w_1x_1+w_3x_2)}}{\partial_{w1}} = x_1$$

$$\frac{\partial_{o_{13}}}{\partial net_{13}} == \sigma(net_{13} + b_{13})(1 - \sigma(net_{13} + b_{13})) = o_{13}(1 - o_{13})$$

$$\frac{\partial net_{56}}{\partial o_{13}} = w_5$$

$$\frac{\partial o}{\partial net_{56}} == \sigma(net_{56} + b_{56})(1 - \sigma(net_{56} + b_{56})) = o(1 - o)$$

$$\frac{\partial E}{\partial o} = \frac{\partial \frac{1}{m}\sum_{i=1}^{m} -y^i \ln o - (1 - y^i)\ln(1 - o)}{\partial_o} = \frac{1}{m}\sum_{i=1}^{m}(\frac{-y^i}{o} + \frac{1 - y^i}{1 - o}) = \frac{1}{m}\sum_{i=1}^{m}(\frac{o - y^i}{o(1 - o)})$$

$$\frac{\partial_E}{\partial_{w1}} = \frac{1}{m}\sum_{i=1}^{m}(\frac{o-y^i}{o(1-o)}) \times o(1-o) \times w_5 \times o_{13}(1-o_{13}) \times x_1$$

$$\frac{\partial_E}{\partial w_1} = \frac{1}{m}\sum_{i=1}^{m}(o-y^i) \times w_5 \times o_{13}(1-o_{13}) \times x_1$$

$$\frac{\partial_E}{\partial_{w_5}} = \frac{\partial_E}{\partial_o} \times \frac{\partial_o}{\partial_{net_{56}}} \times \frac{\partial_{net_{56}}}{\partial_{w_5}}$$

$$\frac{\partial net_{56}}{\partial w_5} = \frac{\partial w_5 o_{13} + w_6 o_{24}}{\partial w_5} = o_{13}$$

$$\frac{\partial_E}{\partial_{w_5}} = \frac{1}{m}\sum_{i=1}^{m}(\frac{o-y^i}{o(1-o)}) \times o(1-o) \times o_{13}$$

$$\frac{\partial_E}{\partial_{w_5}} = \frac{1}{m}\sum_{i=1}^{m}(o-y^i) \times o_{13}$$

(e) The update rule is given as:
$\Delta w_1 = -R \times \frac{1}{m}\sum_{i=1}^{m}(o-y^i) \times w_5 \times o_{13}(1-o_{13}) \times x_1$
$\Delta w_5 = -R \times \frac{1}{m}\sum_{i=1}^{m}(o-y^i) \times o_{13}$

Thus,with R as the learning rate, we can write the final update rule as:
$w_{1t+1} = w_{1t} - R \times \frac{1}{m}\sum_{i=1}^{m}(o-y^i) \times w_5 \times o_{13}(1-o_{13}) \times x_1$
$w_{5t+1} = w_{5t} - R \times \frac{1}{m}\sum_{i=1}^{m}(o-y^i) \times o_{13}$

(f) B)It is a non linear function of the inputs

4. (a) B) $Pr(x_i, x_j|y) = Pr(x_i|y)Pr(x_j|y)$
Naive Bayes independent assumption

(b) B)$Pr(y, x_1, x_2, x_3) = Pr(y)Pr(x_1|y)Pr(x_2|y)Pr(x_3|y)$

(c) 4, 6, 7

(d) We predict y=1, iff
$\frac{P(X_1=1|Y=1)^{X_1}P(X_1=0|Y=1)^{1-X_1}P(X_2=1|Y=1)^{X_2}P(X_2=0|Y=1)^{1-X_2}P(X_3=1|Y=1)^{X_3}P(X_3=0|Y=1)^{1-X_3}P(Y}{P(X_1=1|Y=0)^{X_1}P(X_1=0|Y=0)^{1-X_1}P(X_2=1|Y=0)^{X_2}P(X_2=0|Y=0)^{1-X_2}P(X_3=1|Y=0)^{X_3}P(X_3=0|Y=0)^{1-X_3}P(Y}$
1

$$\frac{[p_1^{x_1}(1-p_1)^{(1-x_1)}][p_2^{x_2}(1-p_2)^{(1-x_2)}][p_3^{x_3}(1-p_3)^{(1-x_3)}]\beta}{[t_1^{x_1}(1-t_1)^{(1-x_1)}][t_2^{x_2}(1-t_2)^{(1-x_2)}][t_3^{x_3}(1-t_3)^{(1-x_3)}](1-\beta)} > 1$$

If the above expression is greater than 1, we predict y =1 , else we predict y=0

(e) i. i.$P(X_1 = 1|Y = 0) = \frac{2+1}{2+2} = \frac{3}{4}$
ii.$P(Y = 1|X_2 = 1) = \frac{2+1}{3+2} = \frac{3}{5}$
iii.$P(X_3 = 1|Y = 1) = \frac{0+1}{6+2} = \frac{1}{8}$

ii. i.$P(X_3 = 1, Y = 1) = P(Y = 1)P(X_3 = 1|Y = 1) = \frac{6+1}{8+2} \times \frac{1}{8} = \frac{7}{80}$
ii.$P(X_3 = 1) = P(X_3 = 1|Y = 1)P(Y = 1) + P(X_3 = 1|Y = 0)P(Y = 0) = \frac{1}{8} \times \frac{7}{10} + \frac{2}{4} \times \frac{3}{10} = \frac{19}{80}$

5. (a) Two directed trees $T_i$ and $T_j$ obtained from the same undirected tree T,over variables $x_1, x_2...x_n$ are equivalent if the joint probability distributions they represent are the same. In other words,

$$P_{Ti}(x_1.x_2...x_n) = P_{Tj}(x_1, x_2...x_n)$$

This implies that for every event E over $x_1, x_2...x_n$ $P_{Ti}(E) = P_{Tj}(E)$

(b) Let $T_i$ and $T_j$ be two directed trees obtained from same undirected tree T by choosing two different roots $x_i$ and $x_j$ where $(i \neq j)$ Denoting $x = x_1, x_2...x_n$, we want to show that $P_{Ti}(x_1, x_2...x_n) = P_{Tj}(x_1, x_2...x_n)$.
Due to the tree structure, there exists a unique path between $x_i$ and $x_j$, which is denoted by $P_{ij}$. All nodes in T are denoted by N. Using induction we can show the proof as follows:

Assuming $||P_{ij}|| = 1$, then

$$P_{Ti}(x_1, x_2...x_n) = P(x_i) \prod_{k \epsilon N, k \neq i} P(x_k|Parent(x_k))$$

$$= P(x_i)P(x_j|x_i) \prod_{k \epsilon N, k \neq i, k \neq j} P(x_k|Parent(x_k))$$

$$= P(x_i, x_j) \prod_{k \epsilon N, k \neq i, k \neq j} P(x_k|Parent(x_k))$$

$$= P(x_j)P(x_i|x_j) \prod_{k \epsilon N, k \neq i, k \neq j} P(x_k|Parent(x_k))$$

$$= P(x_j) \prod_{k \epsilon N, k \neq j} P(x_k|Parent(x_k))$$

$$= P(T_j)(x_1, x_2...x_n)$$

Now, considering if $P_{Ti}(x_1, x_2...x_n) = P_{Tj}(x_1, x_2...x_n)$ is true for Ti and Tj when $||P_{ij}|| = 1$, then when $||P_{ij}|| = 1 + 1$, we can find an intermediate node $x_m$ such that,

$||P_{im}|| = 1$ and $||P_{mj}|| = 1$.
Thus we have,

$$P_{Ti}(x_1, x_2...x_n) = P_{Tm}(x_1, x_2...x_n)$$

$$P_{Tm}(x_1, x_2...x_n) = P_{Tj}(x_1, x_2...x_n)$$

This gives

$$P_{Ti}(x_1, x_2...x_n) = P_{Tj}(x_1, x_2...x_n)$$

Thus using both the cases we can see that $P_{Ti}(x_1, x_2...x_n) = P_{Tj}(x_1, x_2...x_n)$, this expression is always true.