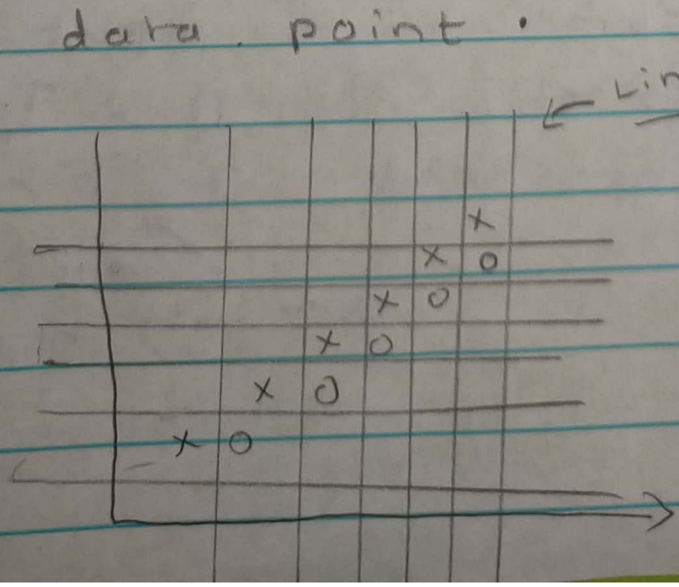
☆ Fall 2019 Midterm

Q1] a] False. Information gain approach of decision tree is a greedy approach which means that if a node is chosen, then we cannot back track. This in turn results in Tree height / number of nodes larger than the optimal solution. Greedy will give a solution which is highly unlikely to be the optimal.

b] False. Even if the data is linearly seperable the size of tree can be the size of the dataset. Dataset size ≥ poly-nomial in d features.
Consider a situation where ~~the~~ each leaf node is exactly one data point.



← Linearly seperable data

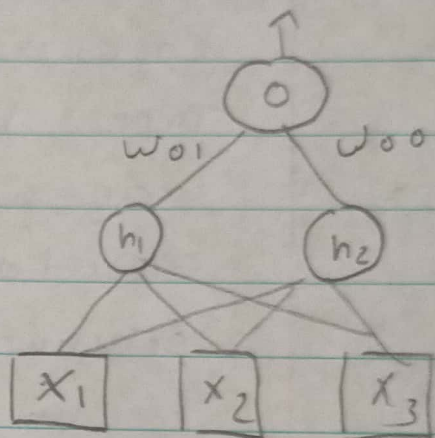
decision boundaries ∝ dataset size

☆ consider this example

Q 2] The table can be written as

$\Rightarrow \neg((x_1 \wedge x_3 \wedge \neg x_2) \vee (\neg x_1 \wedge \neg x_2 \wedge x_3))$

we can represent this by 3 nodes (2 hidden, 1 output)



* $h_1 = x_1 \wedge x_3 \wedge - x_2$

  $w_0 = 0.5 - 3 = -2.5$

  $w_1 = 1, w_2 = 1, w_3 = -1$

* $h_2 = -x_1 \wedge -x_2 \wedge x_3$

  $w_0 = -2.5$

  $w_1 = -1, w_2 = -1, w_3 = 1$

* $o = -(h_1 \vee h_2)$

  $= -h_1 \wedge -h_2$

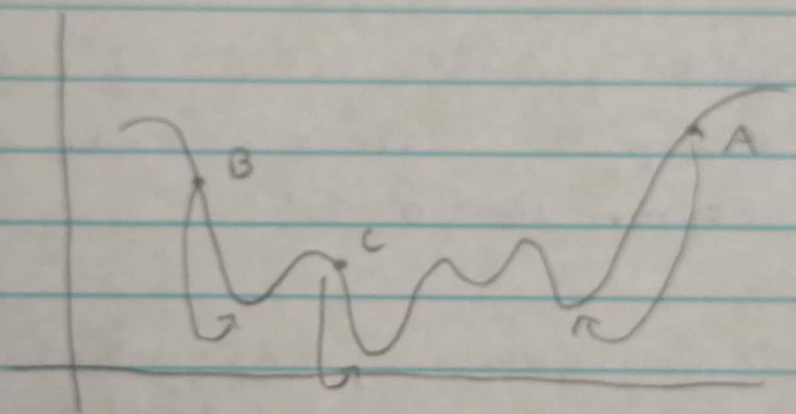
  $w_{00} = -1.5$

  $w_{01} = -1, w_2 = -1$

* Assume that input is -1, 1 instead of 0 & 1
* Threshold if $o > 0, +1$
  else -1
  same with h1, h2

b] True, assuming we are using a non-linear activation function

-> A NN with one hidden layer & a output will have local minimas & a global minima. Depending on where we start, the weights returned by the algorithm would be different.



c] i] Early stopping: Use a validation set to check accuracy during the training & stop it if the error doesn't improve for long or if it is increasing.

ii] Use bare minimum nodes that give a good performance on the training set

iii] Use bagging like approch to randomly select nodes in the NN & ignore them for that particular training iteration.

Q4] a] True

In Naive baye we can handle missing data by summing over all the probabilities of the possible values of K- lets

Say features are boolean.

$$P(X' | Y) = \underbrace{\sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{a=1}^{2} \cdots}_{K \text{ missing features}} \prod_{K+1}^{n} P(X | Y)$$

So this would take $\underline{\underline{2^K}}$ time to compute since each of the K feature $\in \{0, 1\}$

b] (Yet to reach, ans based on my basic understanding).

True,

Because as K incereases we rely on more & more points from the training set to determine the class which is bais. Variance decreases because prediction is becoming more stable.

Because lets say K = dataset size then we just predic the class with more occurnaces which is high bias.

c] $Pr(y) = \frac{\theta^y e^{3.5\theta}}{y!}$

Likelihood,

$$L(D|\theta) = \prod_{k=1}^{n} \frac{\theta^{y_k} e^{3.5\theta}}{y_k!}$$

Log Likelihood,

$$\Rightarrow \sum_{k=1}^{n} \left[y_k \ln(\theta) + 3.5\theta - \log(y_k!) \right]$$

MLE,

$$\frac{\partial L(D|\theta)}{\partial \theta} = 0$$

$$\Rightarrow \sum_{k=1}^{n} \left[\frac{y_k}{\theta} + 3.5 \right] = 0$$

$$\Rightarrow \sum_{k=1}^{n} \frac{y_k}{\theta} = -3.5n$$

$$\Rightarrow \boxed{\theta = -\frac{\sum_{k=1}^{n} y_k}{3.5n}}$$

Scanned with CamScanner

Q5] a] Ada boost will choose X_1 because $X_1 > 1$ is better at seperating data.
There seem to be no value of θ_2 that can out perform $X_1, \theta_1 = 1$

if $\theta_1 = 1$

-ve		+ve	
1	5	1	3
Incorrect	Correct	Incorect	Correct

$$\text{error} = \frac{\frac{1}{10} \times 2}{1} = \frac{1}{5}$$

$$\alpha_m = \ln\left(\frac{4/5}{1/5}\right) = \ln(4)$$

$$\text{weight}_{\text{correct}} = 1/10$$

$$\text{weight}_{\text{wrong prediction}} = \frac{1}{10} \times e^{\ln 4} = \frac{4}{10} = \frac{2}{5}$$

Next Page →

x_1	x_2	y	w
0	8	−	1/10
1	4	−	1/10
3	7	+	1/10
−2	1	−	1/10
−1	13	−	1/10
9	11	−	4/10
12	7	+	1/10
−7	−1	−	1/10
−3	12	+	4/10
5	9	+	1/10

if $\theta_1 = 1$, Error $= \dfrac{\frac{8}{10}}{\frac{16}{10}} = \frac{1}{2}$

if $\theta_2 = 6$, Error $= \dfrac{\frac{3}{10}}{\frac{16}{10}} = \frac{3}{16}$

☆ Therefore we choose x_2, $\theta_2 = 6$