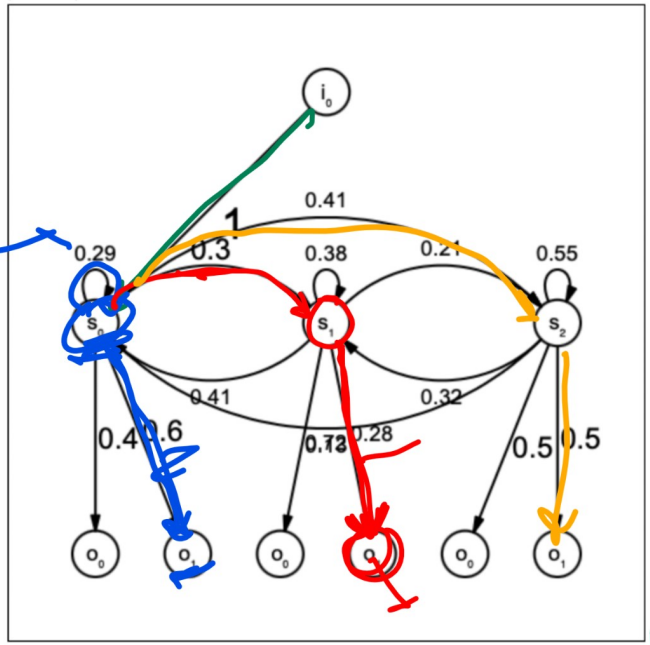
Question 10

MTQ10

• [5 points] Andy is a three-month old baby. He can be happy (state 0), hungry (state 1), or having a wet diaper (state 2). Initially when he wakes up from his nap at 1pm, he is happy. If he is happy, there is a 0.29 chance that he will remain happy one hour later, a 0.3 chance to be hungry by then, and a 0.41 chance to have a wet diaper. Similarly, if he is hungry, one hour later he will be happy with 0.41 chance, hungry with 0.38 chance, and wet diaper with 0.21 chance. If he has a wet diaper, one hour later he will be happy with 0.13 chance, hungry with 0.32 chance, and wet diaper with 0.55 chance. He can smile (observation 0) or cry (observation 1). When he is happy, he smiles 0.4 of the time and cries 0.6 of the time; when he is hungry, he smiles 0.28 of the time and cries 0.5 of the time; when he has a wet diaper, he smiles of the time and cries of the time.

What is the probability that the particular observed sequence cry, cry (or Y_1, Y_2 = 1, 1) happens (in the first two periods)?



$X_1 \to X_2 \to X_3 \to \cdots$

$\downarrow \quad \downarrow$

$Y_1 \quad Y_2$

$\Pr\{Y_1 = 1, Y_2 = 1\}$ need to add X_1, X_2

$= \Pr\{Y_1 = 1, Y_2 = 1, X_1 = 0, X_2 = 0\} +$

$X_2 = 1\} +$

$X_2 = 2\}$

BN $=$ $\Pr\{Y_1 = 1 \mid X_1 = 0\} \cdot \Pr\{Y_2 = 1 \mid X_2 = 0\}$

$\Pr\{X_1 = 0\} \cdot \Pr\{X_2 = 0 \mid X_1 = 0\}$

$0.6 \cdot 0.6 \cdot 1 \cdot 0.29$

$= 0.6 \cdot 0.28 \cdot 1 \cdot 0.3$

$= 0.6 \cdot 0.5 \cdot 1 \cdot 0.41$

sum

Question 4

X2 Q4

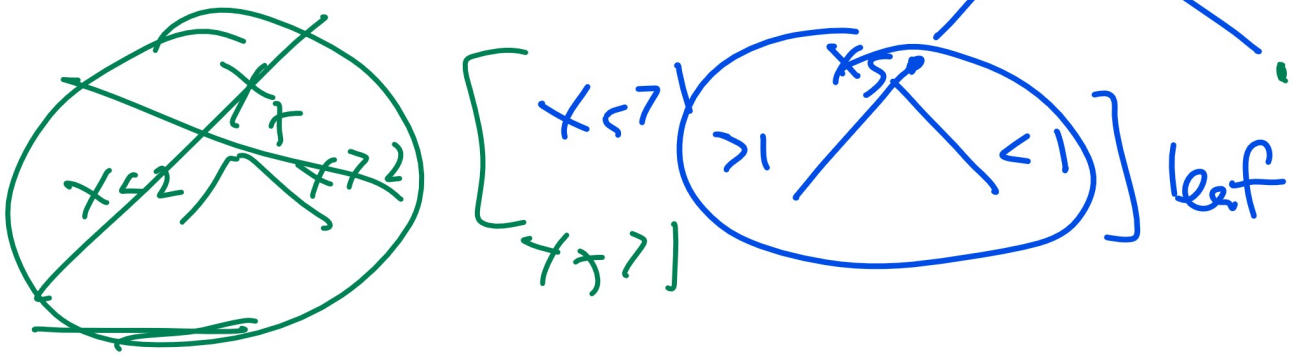
• [4 points] In a problem where each example has 10 real-valued attributes (i.e. features), where each attribute can be split at 3 possible thresholds (i.e. binary splits), to select the best attribute for a decision tree node at depth 6, where the root is at depth 0, how many conditional entropies must be calculated (at most)?

• Answer: . Calculate

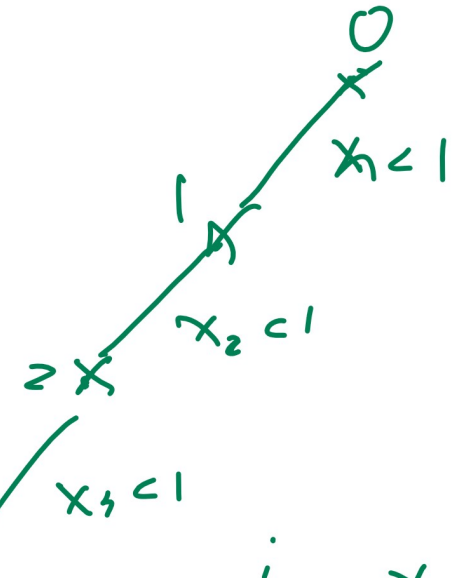
P2

↳ $x_1 \ldots x_{10}$ each can 0, 1, 2, 3

↳ 3 possible splits.



6 splits



[30 possible splits 6 splits are used

x_4<2
x_1<1

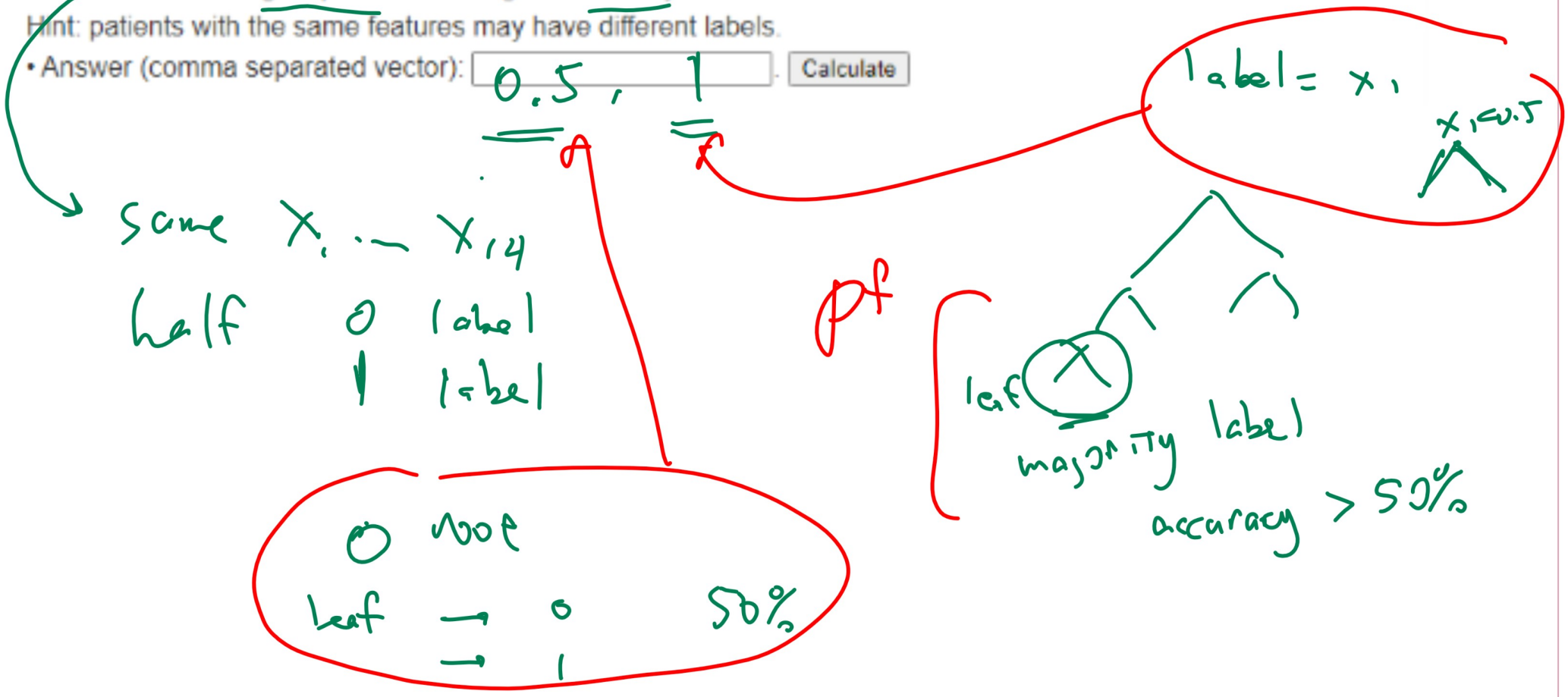
⟹ 24 remaining splits

Question 5

X2 Q5

• [3 points] A hospital trains a decision tree to predict if any given patient has technophobia or not. The training set consists of 80000 patients. There are 14 features. The labels are binary. The decision tree is not pruned. What are the smallest and largest possible training set accuracy of the decision tree? Enter two numbers between 0 and 1.
Hint: patients with the same features may have different labels.

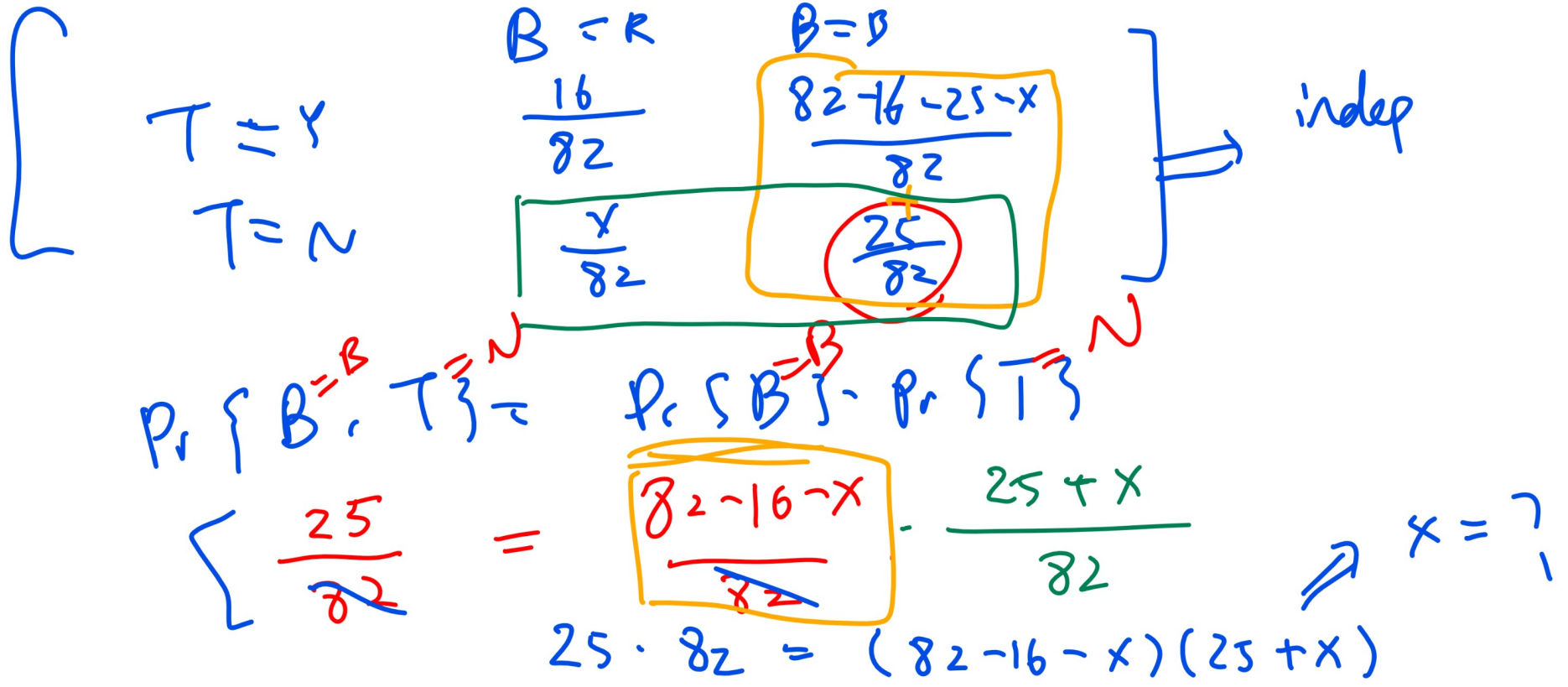
• Answer (comma separated vector): 0.5, 1 Calculate



Question 2

• [4 points] There are 82 parrots. They have either a red beak or a black beak. They can either talk or not. Complete the two cells in the following table so that the mutual information (i.e. information gain) between "Beak" and "Talk" is 0:

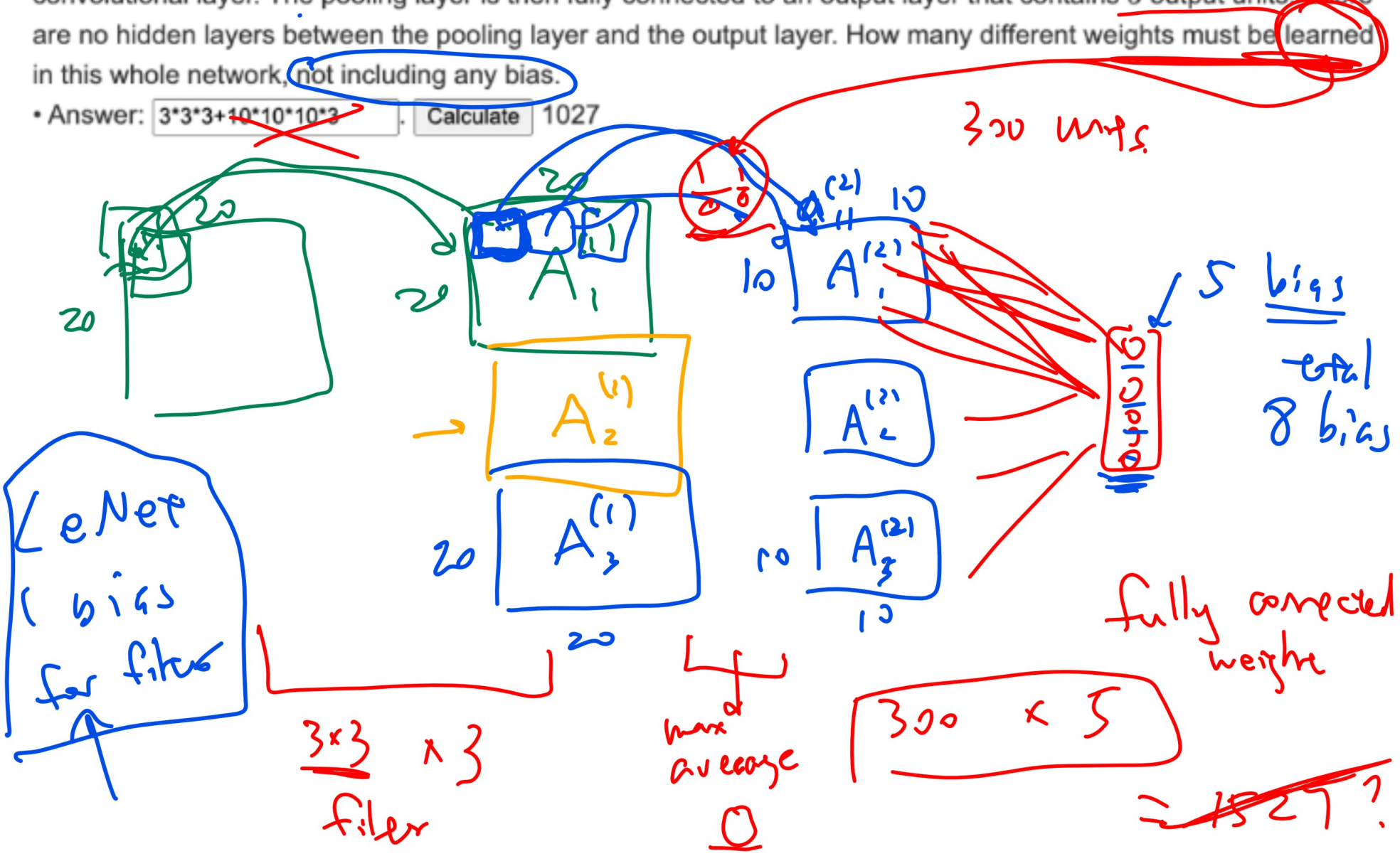
Number of parrots	Beak	Talk
16	Red	Yes
X	Red	No
82 − 16 − 25 − X	Black	Yes
25	Black	No



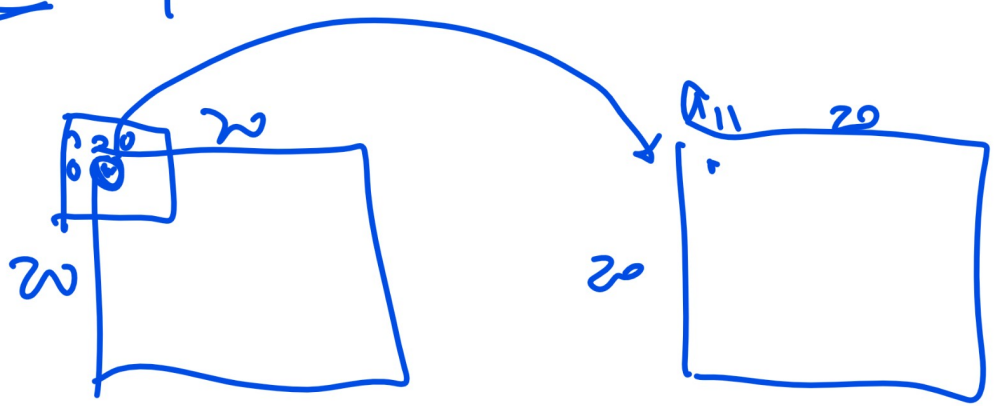
Question 10

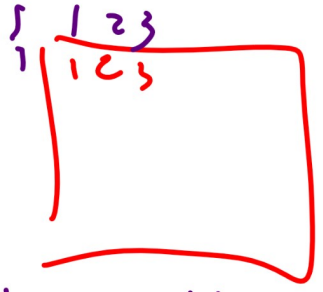
• [4 points] A convolutional neural network has input image of size 20 x 20 that is connected to a convolutional layer that uses a 3 x 3 filter, zero padding of the image, and a stride of 1. There are 3 activation maps. (Here, zero-padding implies that these activation maps have the same size as the input images.) The convolutional layer is then connected to a pooling layer that uses 2 x 2 max pooling, a stride of 2 (non-overlapping, no padding) of the convolutional layer. The pooling layer is then fully connected to an output layer that contains 5 output units. There are no hidden layers between the pooling layer and the output layer. How many different weights must be learned in this whole network, not including any bias.

• Answer: 3*3*3+10*10*10*3 . Calculate 1027



Zero-padding



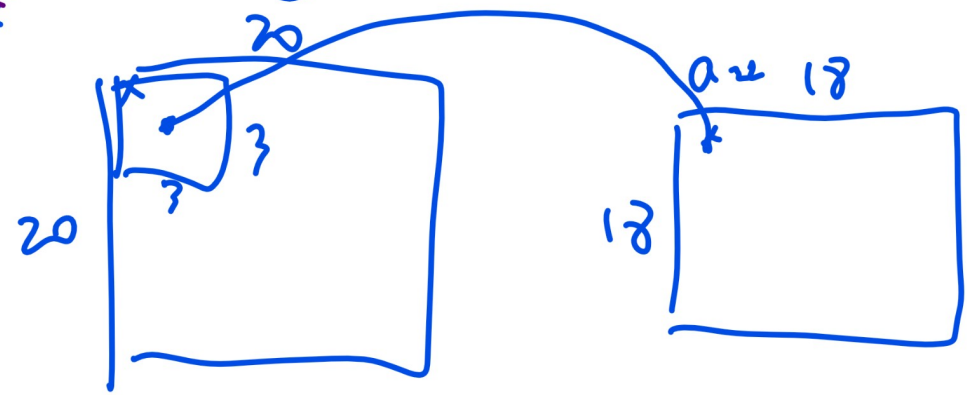


other padding

$m \times m$ image

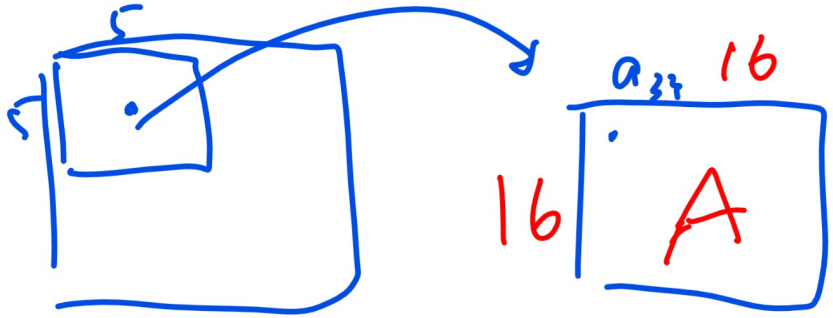
$\Rightarrow$ $m \times m$ activation

no-padding



filter $= (2k+1) \times (2k+1)$

no-padding

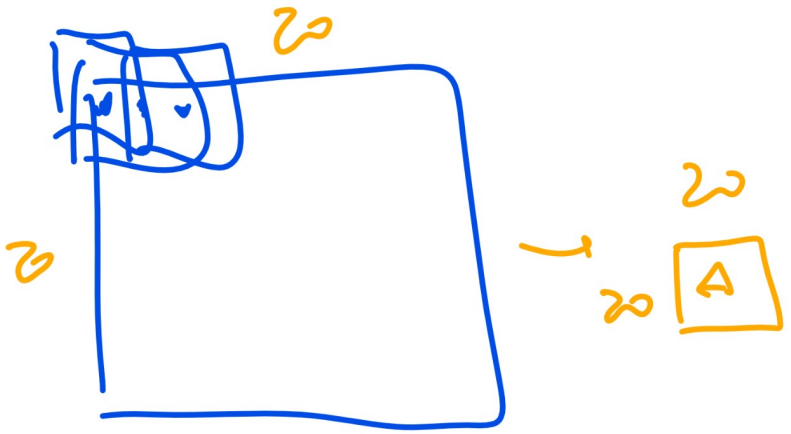


LeNet

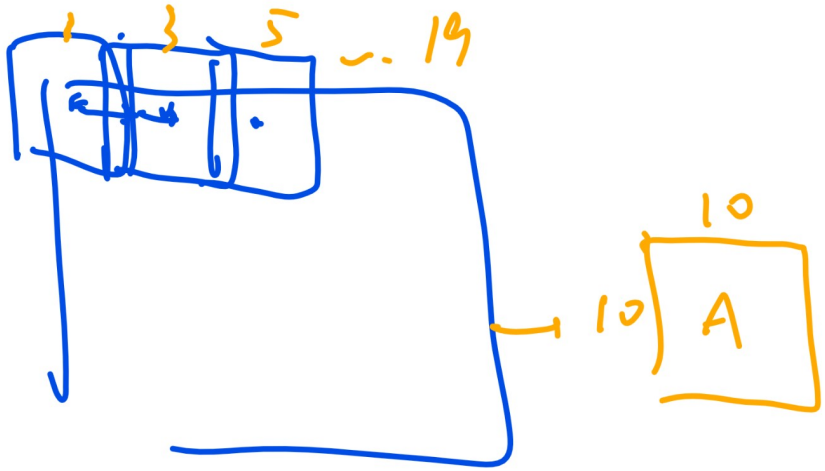
if size of activation $m \times m$

$(m-2k) \times (m-2k)$

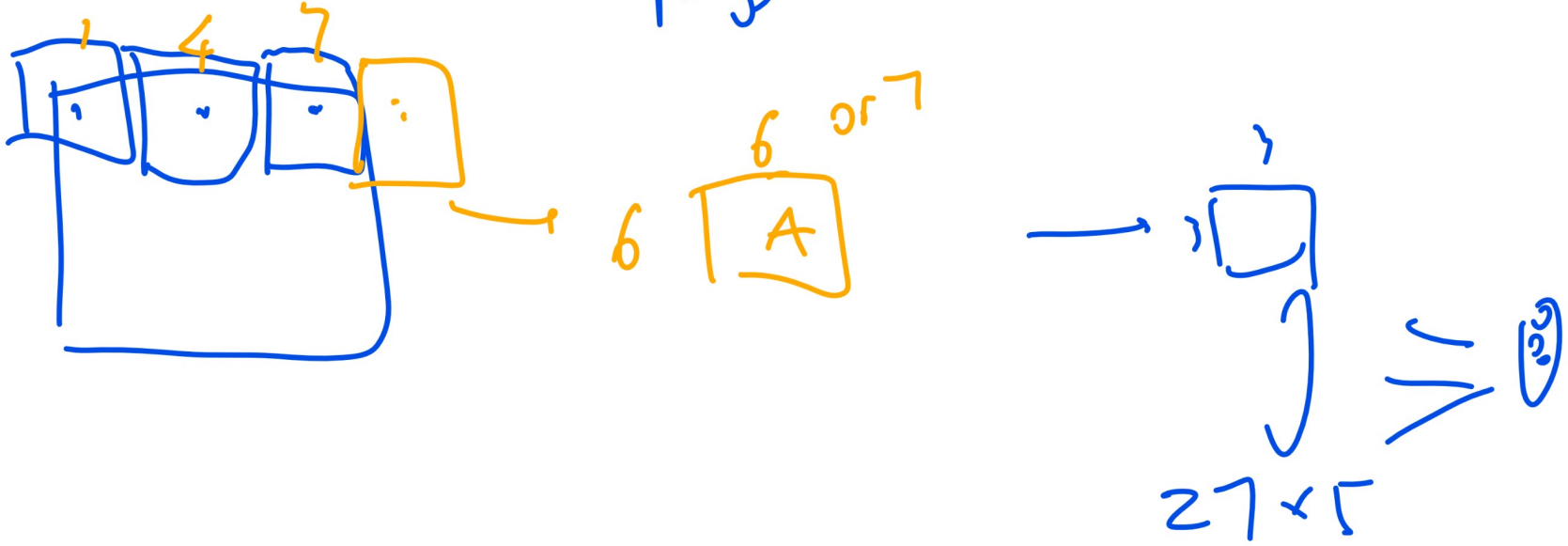
Stride 1



stride 2



stride 3 (non-overlapping)



X2 Q1 — auto grading

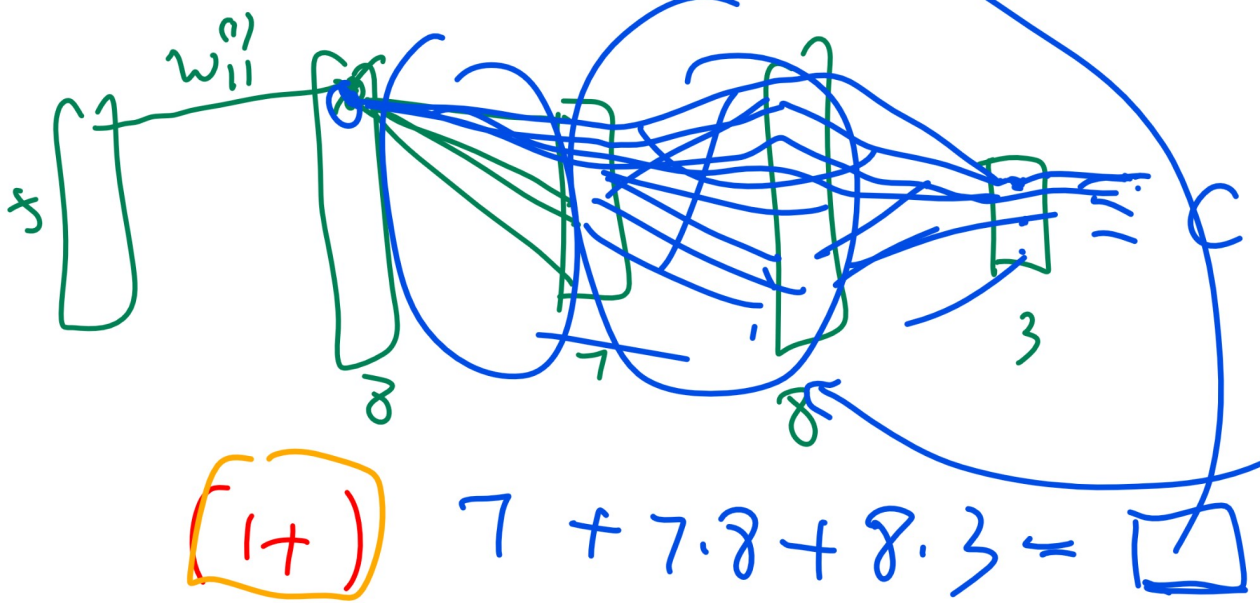
X621

Question 1

• [3 points] Suppose you are given a neural network with 3 hidden layers, 5 input units, 3 output units, and $[8 \quad 7 \quad 8]$ hidden units. In one backpropogation step when computing the gradient of the cost (for example, squared loss) with respect to $w_{11}^{(1)}$, the weight in layer 1 connecting input 1 and hidden unit 1, how many weights (including $w_{11}^{(1)}$ itself, and including biases) are used in the backpropogation step of $\frac{\partial C}{\partial w_{11}^{(1)}}$?

• Note: the backpropogation step assumes the activations in all layers are already known so do not count the weights and biases in the forward step computing the activations.

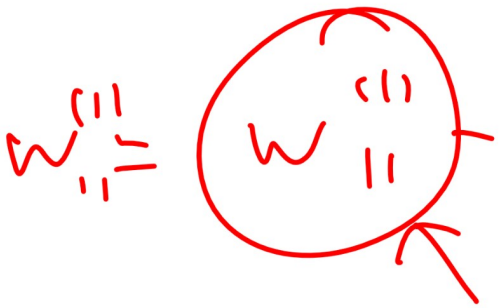
• Answer: 1+7+56+24+7+8+3 . Calculate 106



$\frac{\partial \mathcal{L}}{\partial w_{11}^{(1)}} = \sum_{i=1}^{3} \frac{\partial \mathcal{L}}{\partial a} \frac{\partial a}{\partial w_{11}^{(1)}}$

$\sum_{i=1}^{8} \frac{\partial a^{(3)}}{\partial a^{(2)}} \frac{\partial a^{(3)}}{\partial w_{11}^{(1)}}$

$a(1-a)\,w$

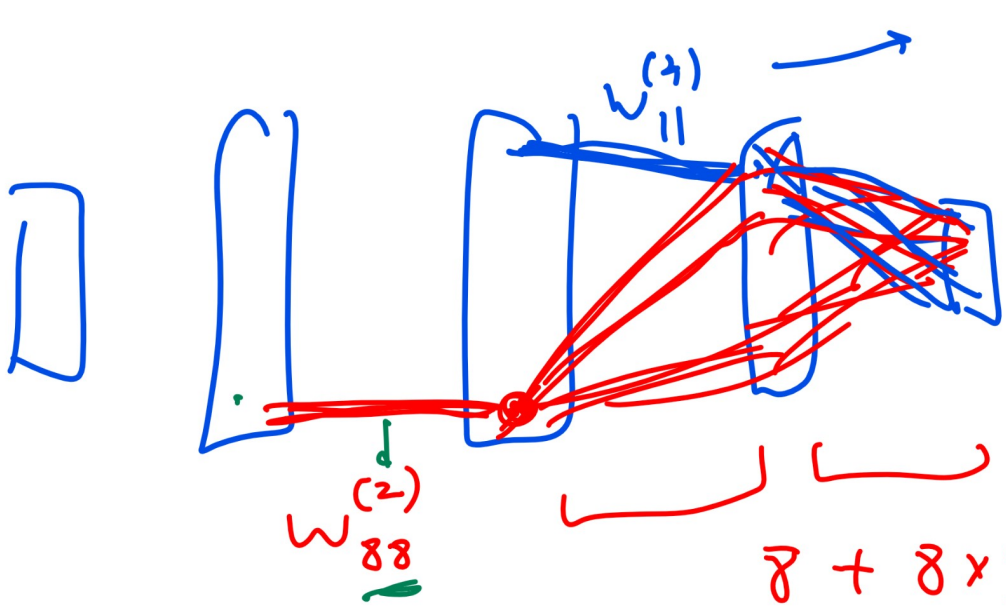


$$\alpha \frac{\partial L}{\partial w_{11}^{(1)}}$$

$$\sum_{7} \frac{\partial a^{(3)}}{\partial a^{(2)}} \frac{\partial a^{(2)}}{\partial w_{11}^{(1)}}$$

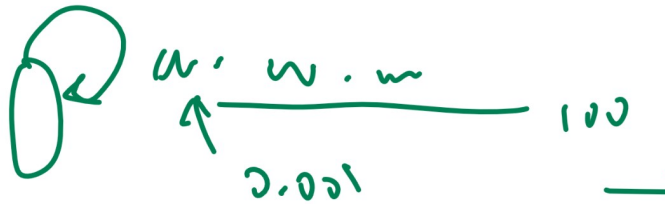
$$\sum_{8} \frac{\partial a^{(2)}}{\partial a^{(1)}} \cdot \frac{\partial a^{(1)}}{\partial w_{11}^{(1)}}$$

$$\frac{\partial a}{\partial x} = w$$



$8 + 8 \times 3 = 32$

RNN →



→ 0

vanishing gradient.

Question 3

• [3 points] You have a joint probability table over $k = 6$ random variables $X_1, X_2, \ldots, X_k$, where each variable takes $m = 4$ possible values: $1, 2, \ldots, m$. To compute the probability that $X_1 = 4$, how many cells in the table do you need to access (at most)?

• Answer: Calculate

$X_1, X_2, X_3 = 0, 1$ $\quad 2 \times 2 = 4$

$Pr\{X_1 = 1\} = Pr\{X_1=1, X_2=0, X_3=0\}$
$+ Pr\{X_1=1, X_2=0, X_3=1\}$
$+ Pr\{X_1=1, X_2=1, X_3=0\}$
$+ Pr\{X_1=1, X_2=1, X_3=1\}$

$X_1 \ldots X_6 = 1, 2, 3, 4$

$Pr\{X_1 = 4\} = Pr\{X_1=4, X_2 \in \{1,2,3,4\}, X_3 \in \{1,2,3,4\} \ldots X_6 \in \{1,2,3,4\}\}$

$4 \cdot 4 \cdot 4 \ldots$

4^5 entries accessed.

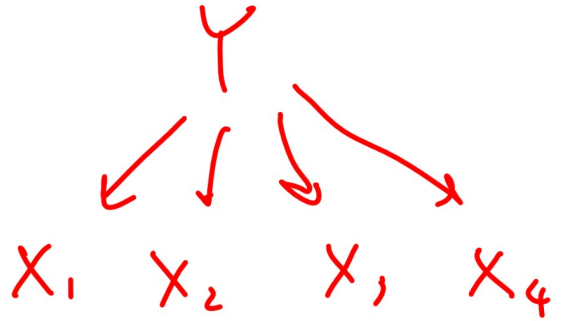
Question 7

X307

• [4 points] Say we use Naive Bayes in an application where there are 4 features represented by 4 variables, each having 6 possible values, and there are 3 classes. How many probabilities must be stored in the CPTs (Conditional Probability Table) in the Bayesian network for this problem? Do not include probabilities that can be computed from other probabilities.

• Answer: Calculate

CPT table of $X \mid P(X)$



$Pr\{X_1 = 6 \mid P(X)\} = 1 - Pr[X_1=1 \mid P(X)] - \ldots X_{1=2} \ldots - X_{1=5}$

Store 5 for each $P(X)$

$Pr\{Y=1\} = 1 - Pr\{Y=2\} - Pr\{Y=3\}$

2

Y	$X_1 \mid Y$	$X_2 \mid Y$	...	$X_4 \mid Y$
2	5+5+5	15		15

$4 \times 15 + 2 \longrightarrow$ CPT #.

Question 5

• [2 points] Consider the following directed graphical model over binary variables: $A \rightarrow B \leftarrow C$ with the following training set.

A	B	C
0	0	0
0	0	0
0	0	1
0	0	1
1	0	0
1	1	0
1	0	1
1	0	1

What is the MLE (Maximum Likelihood Estimate) with Laplace smoothing of the conditional probability that $\mathbb{P}\{B = 0 \mid A = 1, C = 1\}$?

• Answer: Calculate

Question 2

X6 Q2

• [3 points] A tweet is ratioed if a reply gets more likes than the tweet. Suppose a tweet has 3 replies, and each one of these replies gets more likes than the tweet with probability 0.96 if the tweet is bad, and probability 0.11 if the tweet is good. Given a tweet is ratioed, what is the probability that it is a bad tweet? The prior probability of a bad tweet is 0.73.

• Answer: Calculate

$\Pr[B \mid R]$

$\Pr[B] = 0.73$

$\Pr[\neg B] = 0.27$

$\Pr[\neg R \mid B] =$ all replies get fewer likes

$= 0.04 \cdot 0.04 \cdot 0.04$

$\Pr[\neg R \mid \neg B] = 0.89 \cdot 0.89 \cdot 0.89$

$1 - 0.04^3$

$\Pr[R \mid B] \cdot \Pr[B]$ 0.73

$\Pr[R \mid B] \cdot \Pr[B] + \Pr[R \mid \neg B] \cdot \Pr[\neg B]$ 0.27 $=$ ☐

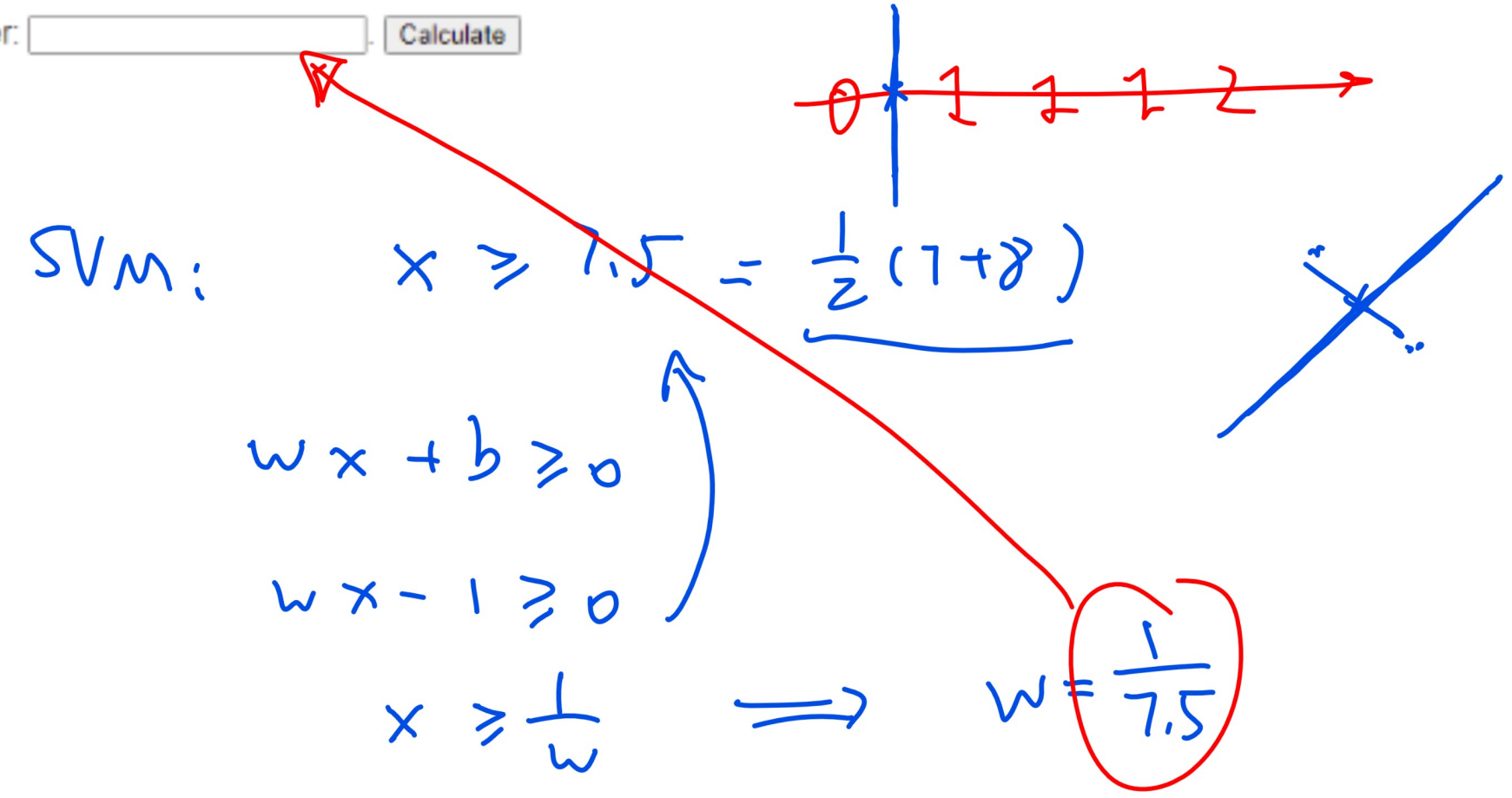
$1 - 0.89^3$

Question 3

• [3 points] A hard margin support vector machine (SVM) is trained on the following dataset. Suppose we restrict $b = -1$, what is the value of w? Enter a single number, i.e. do not include b. Assume the SVM classifier is $1_{\{wx+b\geq 0\}}$.

x_i	7	8	9	18	20
y_i	0	1	1	1	1

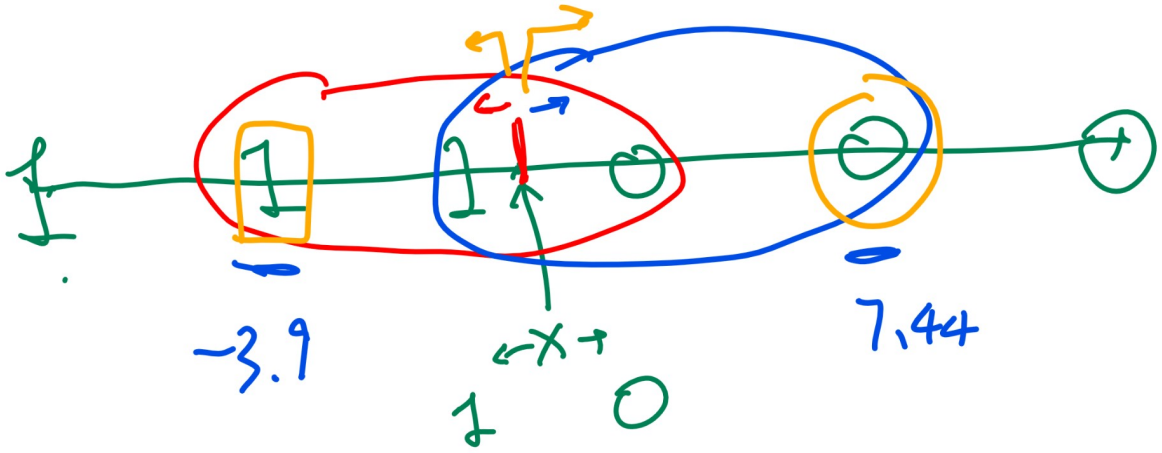
• Answer: Calculate



Question 6

• [4 points] You are given a training set of six points and their 2-class classifications (+ or -): $(-4.53, +)$, $(-3.9, +)$, $(-2.96, +)$, $(-1.95, -)$, $(7.44, -)$, $(9.29, -)$. What is the decision boundary associated with this training set using 3NN (3 Nearest Neighbor)? Note: there is one more point compared to the question from the homework.

• Answer: [ ] Calculate



decision boundary → $\frac{1}{2}(-3.9 + 7.44)$