Question 8

• [4 points] John tells his professor that he forgot to submit his homework assignment. From experience, the professor knows that students who finish their homework on time forget to turn it in with probability 0.87. She also knows that 0.4 of the students who have not finished their homework will tell her they forgot to turn it in. She thinks that 0.96 of the students in this class completed their homework on time. What is the probability that John is telling the truth (i.e. he finished it given that he forgot to submit it)?

• Answer: Calculate

$$Pr\{Forget \mid Finish\} = 0.87$$

$$Pr\{Forget \mid \neg Finish\} = 0.4$$

$$Pr\{Finish\} = 0.96$$

$$Pr\{Finish \mid Forget\} = \frac{Pr\{Forget \mid Finish\} \cdot Pr\{Finish\}}{Pr\{Forget \mid Finish\} \cdot Pr\{Finish\} + Pr\{Forget \mid \neg Finish\} \cdot Pr\{\neg Finish\}}$$

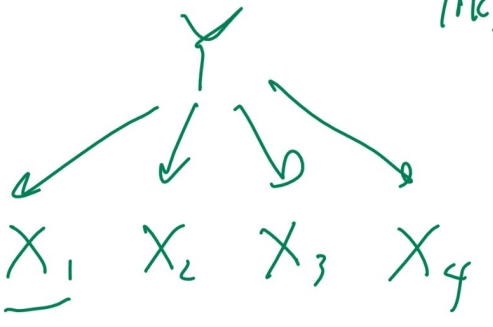
(Bayes Rule; 0.87, 0.96, 0.4, 0.04)

Question 9

• [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 total 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 numbers?

• Answer: Calculate

$5 \times 6 \times 4 + 2$



likelihood

$Pr\{X_i \mid Y\} = ?$ number

$\times 3$

$\times 4$

$Pr\{Y\}$ → 2 number

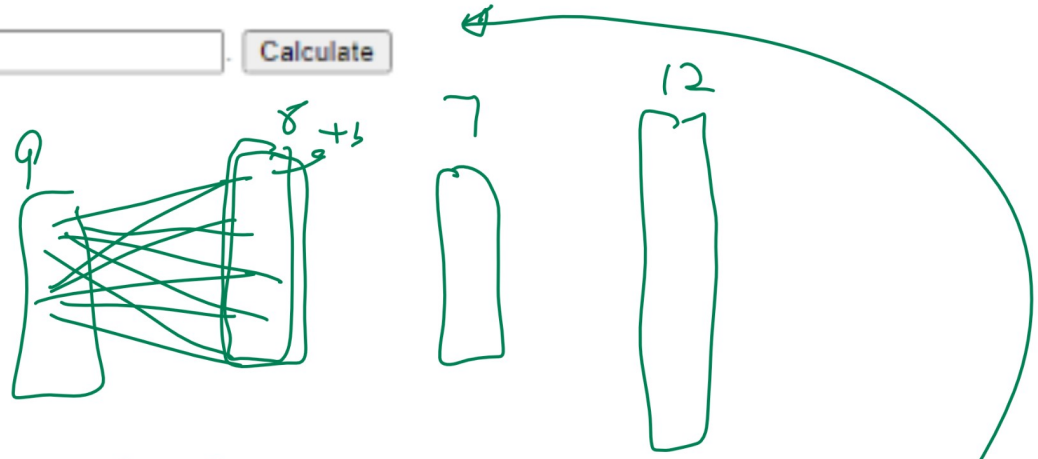
prior

Question 2

• [2 points] In a three-layer (fully connected) neural network, the first layer contains 8 sigmoid units, the second layer contains 7 units, and the output layer contains 12 units. The input is 9 dimensional. How many weights plus biases does this neural network have? Enter one number.

▶ Hint

• Answer: Calculate



weights $9\times 8 + 8\times 7 + 7\times 12$]

biases $8 + 7 + 12$] add.

Question 11

• [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)?

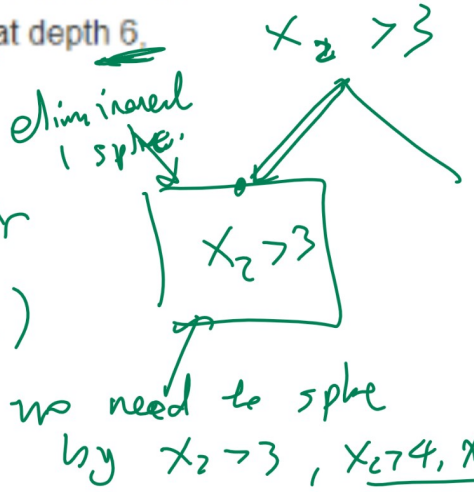
• Answer: Calculate

9

$3 \cdot 10 - 6 = 24$

info gain

possible feature-threshold pair (candidate split)



Question 1

• [4 points] Given two instances $x_1 = 8$ and $x_2 = -2$, suppose the feature map for a kernel SVM (Support Vector Machine) is $\varphi(x) = \begin{bmatrix} \exp(x) \\ x \\ x \end{bmatrix}$, what is the kernel (Gram) matrix?

2x2

dim $\varphi = \infty$

• Answer (matrix with multiple lines, each line is a comma separated vector): Calculate

$K_{ij} = \phi^T(x_i)\,\phi(x_j)$

$\phi_1 = \begin{pmatrix} e^8 \\ 8 \\ 8 \end{pmatrix}$ $\phi_2 = \begin{pmatrix} e^{-2} \\ -2 \\ -2 \end{pmatrix}$

$e^{16} + 64 \cdot 2$

$\begin{bmatrix} \phi_1^T\phi_1 & \phi_1^T\phi_2 \\ \phi_2^T\phi_1 & \phi_2^T\phi_2 \end{bmatrix}$

$\phi_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$

$\begin{bmatrix} \phi_1^T\phi_1 & \phi_1^T\phi_2 & \phi_1^T\phi_3 \\ \phi_2^T\phi_1 & \cdots & \\ & & \phi_3^T\phi_3 \end{bmatrix}$ 3x3

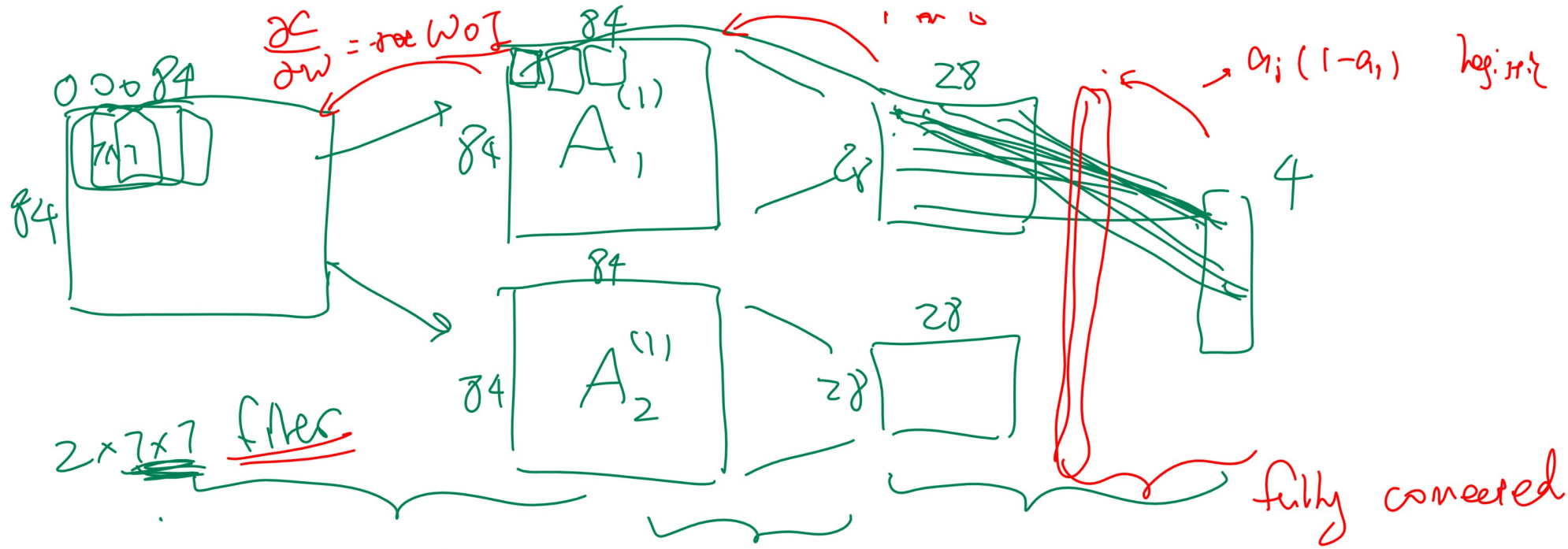
if add $x_3 = 0$

Question 7

• [4 points] A convolutional neural network has input image of size 84 x 84 that is connected to a convolutional layer that uses a 7 x 7 filter, zero padding of the image, and a stride of 1. There are 2 activation maps. The convolutional layer is then connected to a pooling layer that uses 3 x 3 max pooling, a stride of 3 (non-overlapping), and no padding of the convolutional layer. The pooling layer is then fully connected to an output layer that contains 4 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.

LeNet

• Answer: Calculate



weights $2 \cdot 7 \times 7$ | 0 | $28 \cdot 28 \cdot 2 \cdot 4$

(hidden) units | output units

biases 2 or 0 | 0 | 4

Question 11

• [4 points] What is the conditional entropy $H(B|A)$ for the following set of training examples.

item	A	B
1	T	T
2	F	F
3	T	F
4	F	T
5	F	T
6	F	F
7	F	T
8	F	T

$\log 2$

$$\Pr[A=T] \cdot H(B|A=T) + \Pr[A=F] \cdot H(B|A=F)$$

$$\Rightarrow \frac{4}{8} \cdot \left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{3}{4} \cdot \left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right)$$

Question 5

• [4 points] Consider a kernel $K(x,y) = 9 \cdot 2^{x+y} + 5 \cdot x^2 \cdot y^2 + 10 \cdot \exp(x+y)$, where both x and y are positive real numbers. What is the feature vector $\varphi(x)$ induced by this kernel evaluated at $x = 8$?

• Answer (comma separated vector): Calculate

$$K(x,y) = \phi^T(x)\phi(y) \qquad \text{guess } \phi(x)\phi(y) = \begin{pmatrix} 3 \cdot 2^x \\ \sqrt{5}\,x^2 \\ \sqrt{10}\,e^x \end{pmatrix}^T \begin{pmatrix} 3 \cdot 2^y \\ \sqrt{5}\,y^2 \\ \sqrt{10}\,e^y \end{pmatrix}$$

Question 14

• [4 points] What is the gradient magnitude of the center element (pixel) of the image $\begin{bmatrix} 8 & -2 & 10 \\ -4 & -3 & -5 \\ -1 & -8 & 5 \end{bmatrix}$. Use the x gradient filter: $\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$, and the y gradient filter: $\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$. Remember to flip the filters.

$\begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$

∇_x center $= -6$

∇_y center $= 26$

$\sqrt{6^2 + 26^2}$

Question 2

• [4 points] You have a data set with 44 positive items and 22 negative items. You perform a "leave-one-out" procedure: for each item i, learn a separate kNN (k Nearest Neighbor) classifier on all items except item i, and compute that kNN's accuracy in predicting item i. The leave-one-out accuracy is defined to be the average of the accuracy for each item. What is the leave-one-out accuracy when k = 65?

• Answer: [ ] Calculate

classifier always returns majority = positive here

65 NN

1 NN

100%

66 instances

65 training

1 test

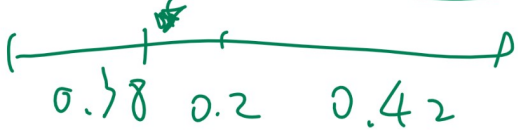
neighbors are all 65

$\frac{44}{66}$

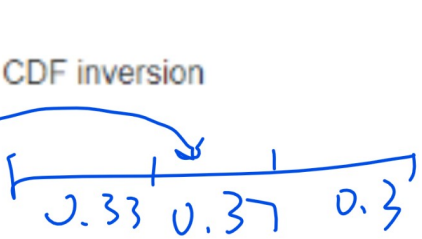
Question 13

• [4 points] Given the following transition matrix for a bigram model with words "I" (label 0), "am" (label 1) and "Groot" (label 2) $\begin{bmatrix} 0.38 & 0.2 & 0.42 \\ 0.33 & 0.37 & 0.3 \\ 0.35 & 0.13 & 0.52 \end{bmatrix}$. Row i column j is $\mathbb{P}\{w_t = j | w_{t-1} = i\}$. Two uniform random numbers between 0 and 1 are generated to simulate the words after "I", say $u_1 = 0.42$ and $u_2 = 0.5$. Using the CDF inversion method, which two words are generated? Enter two integer labels (0, 1, or 2), not strings.

• Answer (comma separated vector): Calculate



I am Groot



I am am

Question 13

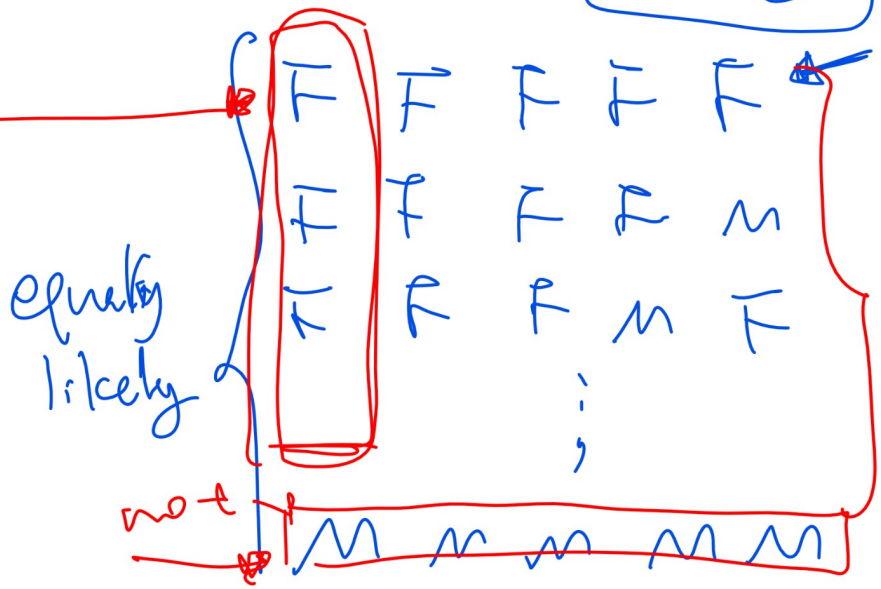
• [3 points] Assume the prior probability of having a female child is the same as having a male child, both are 0.5. The Smith family has 5 kids. One day you saw one of the Smith children, and she is a girl. The Wood family has 5 kids, too, and you heard that at least one of them is a girl. What is the chance that the Smith family has a boy? What is the chance that the Wood family has a boy?

• Answer (comma separated vector): Calculate

no info $1 - \frac{1}{2^5}$

Smith $1 - \frac{1}{2^4}$

Wood $1 - \frac{1}{2^5 - 1}$

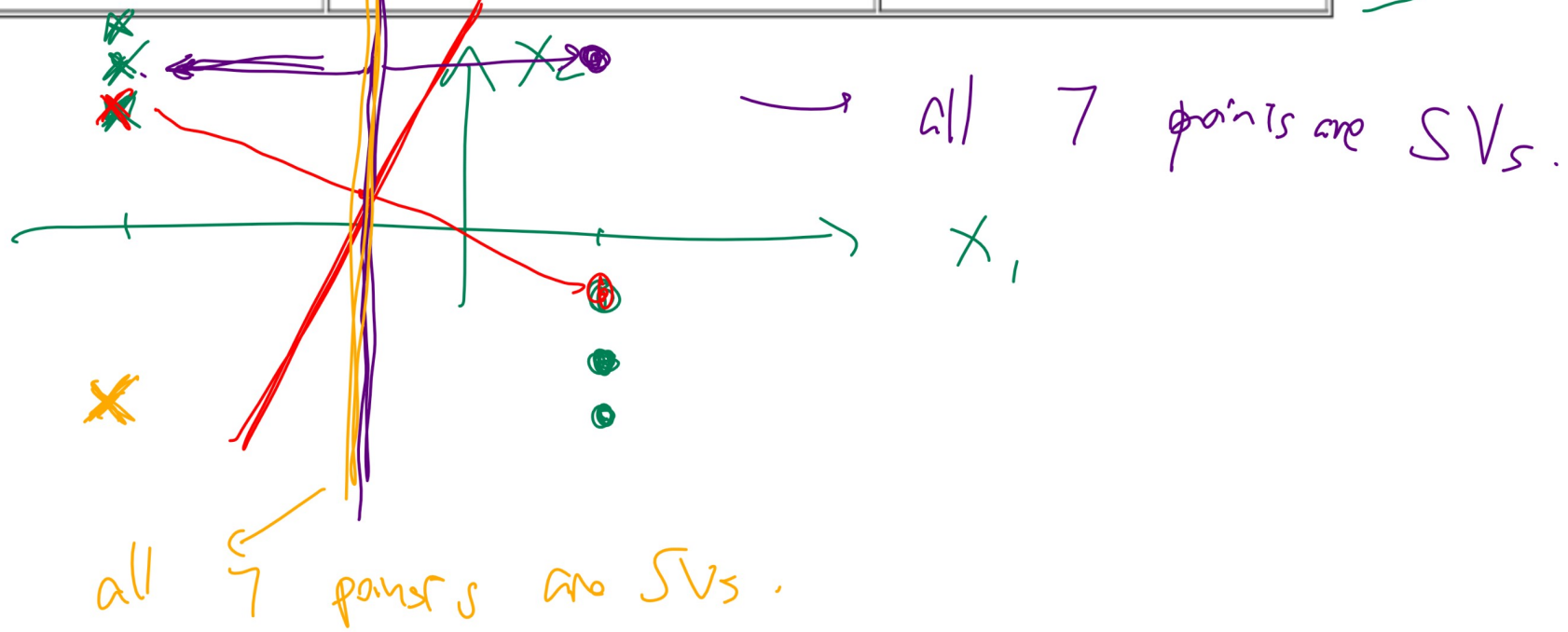


Question 8

M4 Q9 M3 Q9

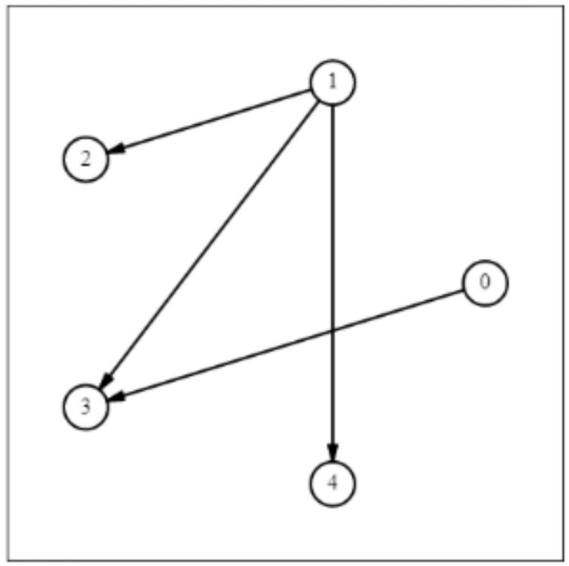
• [2 points] Given the following training set, add one instance $\begin{bmatrix} x_1, \\ x_2 \end{bmatrix}$ with $y = 1$ so that all instances are support vectors for the Hard Margin SVM (Support Vector Machine) trained on the new training set.

x_1	x_2	y
2	-5	0
2	-4	0
2	-2	0
-4	4	1
-4	9	1
-4	8	1



Question 5

• [4 points] Consider the following Bayesian Network containing 5 Boolean random variables. How many numbers must be stored in total in all CPTs (Conditional Probability Table) associated with this network (excluding the numbers that can be calculated from other numbers)?



▶ Hint

• Answer: Calculate

Question 6

• [3 points] You roll a 6-sided die 28 times and observe the following counts: 2, 4, 9, 2, 6, 5. Use Laplace smoothing (i.e. add-1 smoothing), estimate the probability of each side.

▶ Hint

• Answer (comma separated vector, 6 numbers): Calculate

Question 7

• [2 points] An n-gram language model computes the probability $\mathbb{P}\{w_n|w_1, w_2, \ldots, w_{n-1}\}$. How many parameters need to be estimated for a 4-gram language model given a vocabulary size of 63?

▶ Hint

• Answer: Calculate