Q1. 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**.

P(Reply gets more likes than Tweet | Tweet is Bad)

$P(B) = 0.73$

$P(G) = 0.27$

Tweet

- R1 — 0.96 | 0.04
- R2 — 0.96 | 0.04
- R3 — 0.96 | 0.04

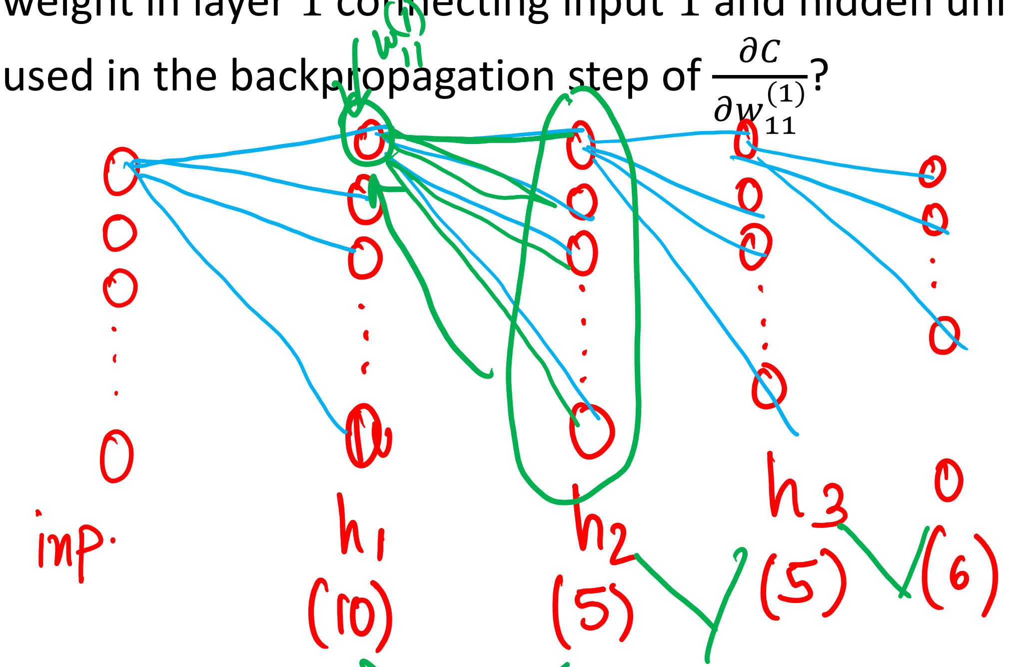
$$P(R|B) = 1 - [P(\text{Reply has less likes} \mid \text{Bad})]^3 = 1 - 0.04^3 = 0.999936$$

$$P(R|G) = 1 - [P(\text{'' '' '' ''} \mid \text{Good})]^3 = 1 - 0.89^3 = \ldots$$

$$P(B|R) = \frac{P(R|B) \cdot P(B)}{P(R)} = \ldots$$

$$P(R) = P(R|B) \cdot P(B) + P(R|G)P(G)$$

Q2. Suppose you are given a neural network with 3 hidden layers, 9 input units, 6 output units, and $[10 \;\; 5 \;\; 5]$ hidden units. In one backpropagation 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 backpropagation step of $\frac{\partial C}{\partial w_{11}^{(1)}}$?



$5 + 25 + 30 = 60$

Q3. 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

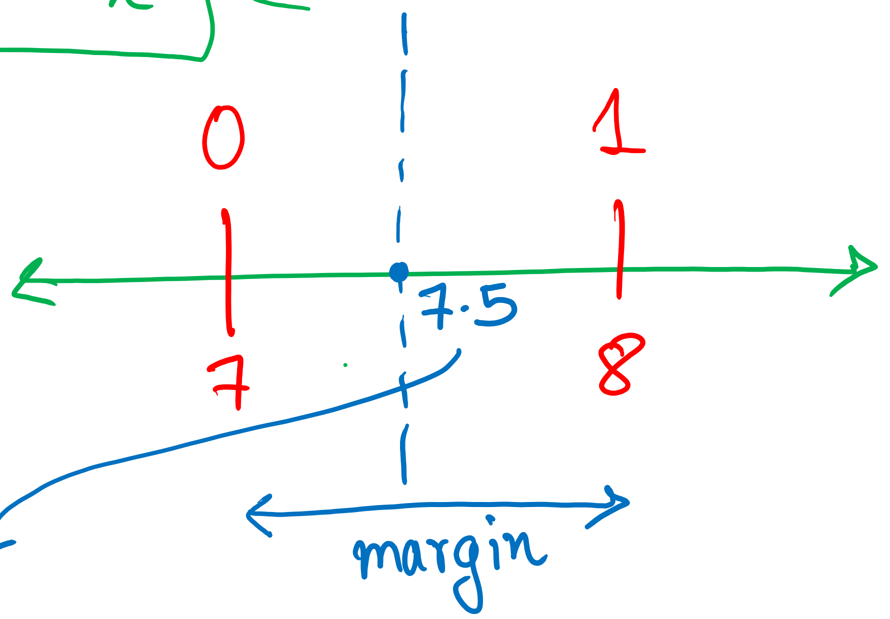
$$wx + b \geq 0 \Rightarrow wx - 1 \geq 0$$

$$wx \geq 1$$

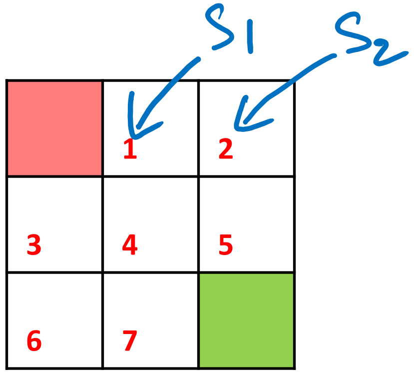
$$\boxed{w \geq \frac{1}{x}}$$

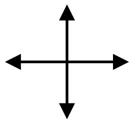
$$w < \frac{1}{7} \quad \text{— ①}$$

$$w \geq \frac{1}{8} \quad \text{— ②}$$



$$w = \frac{1}{7.5}$$





$R_t = -1$

$\gamma = 0.9$

π	ACTION			
States/Tiles	**UP**	**DOWN**	**LEFT**	**RIGHT**
S_1	0.25	0.5	0.1	0.15
S_2	0.1	0.3	0.3	0.3
S_3	0.2	0.25	0.25	0.3
S_4	$\boxed{0.4}$	$\boxed{0.2}$	0.15	0.25
S_5	0.22	0.18	0.5	0.1
S_6	0.25	0.25	0.25	0.25
S_7	0.2	0.2	0.4	0.2

$$V_{i+1} = \sum \pi(s|a) \cdot \sum P(s'|s,a) \times (r + \gamma V_i(s))$$

$$= 0.4 \times (-1 + 0.9 \cdot (-1))$$

$$0.2 \times (-1 + 0.9 \cdot (-1))$$

$$+$$

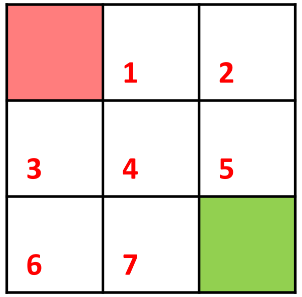
$$\vdots$$

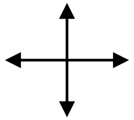
$$= -1.9$$

$V_{k=1} =$

0	-1	-1
-1	-1	-1
-1	-1	0

Find $V_{k=2}$ for states S_4





$R_t = -1$

$\gamma = 0.9$

π	ACTION			
States/Tiles	**UP**	**DOWN**	**LEFT**	**RIGHT**
S_1	0.25	0.5	0.1	0.15
S_2	0.1	0.3	0.3	0.3
S_3	0.2	0.25	0.25	0.3
S_4	$\boxed{0.4}$	0.2	0.15	0.25
S_5	0.22	0.18	0.5	0.1
S_6	0.25	0.25	0.25	0.25
S_7	0.2	0.2	0.4	0.2

-0.8

UP:
$\left(-1 + 0.9\left(-\cancel{2.9}\right)\right) \times 0.4$

DOWN:
$\left(-1 + 0.9\left(-3.45\right)\right) \times 0.2$

LEFT:
$\left(-1 + 0.9\left(-0.33\right)\right) \times 0.15$

DOWN: - - - -

+

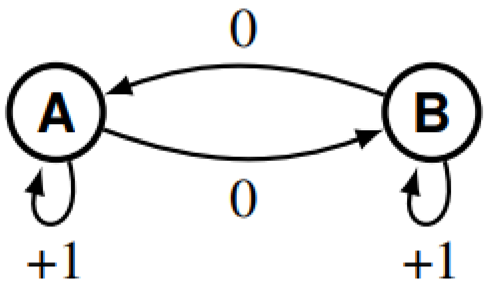
-2.09

$V_{k=i} =$

0	$\boxed{-0.8}$	-1.4
-0.33	-2.9	-0.64
-0.25	-3.45	0

Find $V_{k=i+1}$ for states S_4

$$V_{i+1} = \sum \pi(a|s) \cdot \sum P(s'|s,a) \times \left(r + \gamma V_i(s')\right)$$



$\gamma = 0.8$

Consider the following Markov Decision Process. It has two states s. It has two actions a: move and stay. The state transition is deterministic: "move" moves to the other state, while "stay" stays at the current state. The reward r is 0 for move, 1 for stay. The agent starts at state A. In case of tie move. Use the following Bellman's Equation:

Bellman Equation: $\boxed{Q(s_t, a_t)_t + \gamma \max_{a'} Q(s_{t+1}, a')}$

$1 \quad + 0.8 \cdot 1$

$$Q_{i+1}(s_t, a_t) = Q_i(s_t, a_t) + r + \gamma \max_{a'} Q_i(s_{t+1}, a') - Q_i(s_t, a_t)$$

Find Q_1 , Q_∞ for this state transition table.

$1 + 0.8 + 0.8^2 + \dots = 5$

$Q_0 = \begin{bmatrix} 0, & 0 \\ 0, & 0 \end{bmatrix}, \quad Q_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$

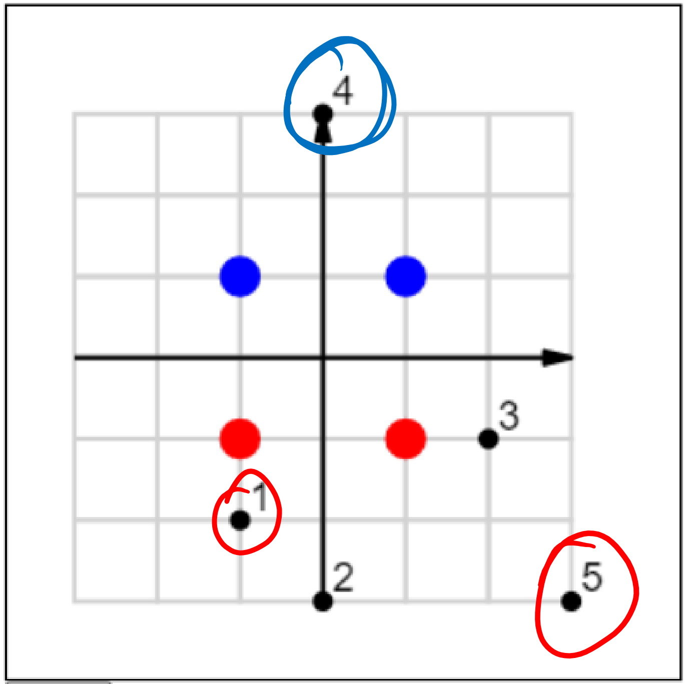
$Q_2 = \begin{bmatrix} 1+0.8 & 0 \\ 0 & 0 \end{bmatrix}, \quad Q_3 = \begin{bmatrix} 1+0.8(1+0.8) & 0 \\ 0 & 0 \end{bmatrix} \dots$

$Q_\infty = \begin{bmatrix} 5 & 0 \\ 0 & 0 \end{bmatrix}$

Question 8

• [3 points] Consider points in 2D and binary labels. Given the training data in the table, and use Manhattan distance with 1NN (Nearest Neighbor), which of the following points in 2D are classified as 1? Answer the question by first drawing the decision boundaries. The drawing is not graded.

index	x_1	x_2	label
1	-1	-1	1
2	-1	1	0
3	1	-1	1
4	1	1	0



Question 1

- [4 points] Consider a classification problem with $n = 36$ classes $y \in \{1, 2, \ldots, n\}$, and two binary features $x_1, x_2 \in \{0, 1\}$. Suppose $\mathbb{P}\{Y = y\} = \frac{1}{36}$, $\mathbb{P}\{X_1 = 1 | Y = y\} = \frac{y}{50}$, $\mathbb{P}\{X_2 = 1 | Y = y\} = \frac{y}{70}$. Which class will naive Bayes classifier produce on a test item with $X_1 = 1$ and $X_2 = 1$.

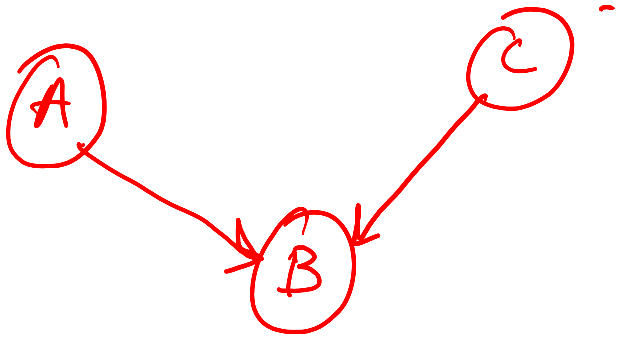
$$P(Y=y \mid X_1=1, X_2=1) = \frac{P(X_1=1, X_2=1 \mid Y=y) \cdot P(y)}{\sum_{y'} P(X_1=1, X_2=1 \mid Y=y') \cdot P(y')}$$

$$= \frac{P(X_1=1 \mid Y=y) \cdot P(X_2=1 \mid Y=y) \cdot P(y)}{\sum_{y'} P(X_1=1 \mid Y=y')\, P(X_2=1 \mid Y=y') \cdot P(y')} = \frac{\frac{y}{50} \times \frac{y}{70} \cdot \frac{1}{36}}{\sum_{y'=1}^{36} \frac{y}{50} \cdot \frac{y}{70} \cdot \frac{1}{36}}$$

$$\max_y \left(\frac{y^2}{c} \right) \Rightarrow y = 36$$

Question 3

• [3 points] Consider the following directed graphical model over binary variables: $A \to B \leftarrow C$. Given the CPTs (Conditional Probability Table):

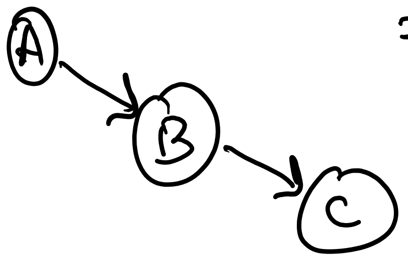


Variable	Probability	Variable	Probability
$\mathbb{P}\{A=1\}$	0.51		
$\mathbb{P}\{C=1\}$	0.84		
$\mathbb{P}\{B=1 \mid A=C=1\}$	0.64	$\mathbb{P}\{B=1 \mid A=0, C=1\}$	0.83
$\mathbb{P}\{B=1 \mid A=1, C=0\}$	0.27	$\mathbb{P}\{B=1 \mid A=C=0\}$	0.17

What is the probability that $\mathbb{P}\{A=0, B=0, C=1\}$?

$P(A=0) \cdot P(B=0 \mid A=0, C=1) \cdot P(C=1)$

$A \longrightarrow B \longrightarrow C$



$\Rightarrow P(A=0) \cdot P(B=0 \mid A=0) \cdot P(C=1 \mid B=0, A)$

$\downarrow$

$P(C=1 \mid B=0, A=0)$
$+$
$P(C=1 \mid B=0, A=1)$

Question 9

- [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?

K_{12}

$$K(x_1, x_2) = \begin{bmatrix} \boxed{\varphi(x_1)^T \varphi(x_1)} & \boxed{\varphi(x_1)^T \varphi(x_2)} \\ \varphi(x_2)^T \varphi(x_1) & \varphi(x_2)^T \varphi(x_2) \end{bmatrix} = \begin{bmatrix} \varphi(x_1) \\ \varphi(x_2) \end{bmatrix} \times \begin{bmatrix} \varphi(x_1) & \varphi(x_2) \end{bmatrix}$$

K_{11}

$$= \begin{bmatrix} e^{2x_1} + x_1^2 + x_1^2 & e^{x_1+x_2} + x_1x_2 + x_1x_2 \\ e^{x_1+x_2} + x_1x_2 + x_1x_2 & e^{2x_2} + x_2^2 + x_2^2 \end{bmatrix}$$

Question 15

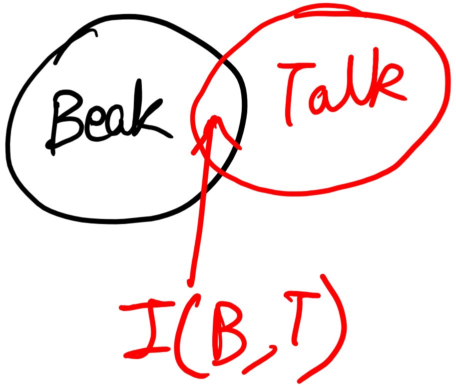
• [4 points] Given the following training data, what is the 6 fold cross validation accuracy if 1NN (Nearest Neighbor) classifier with Manhattan distance is used. The first fold is the first 1 instances, the second fold is the next 1 instances, etc. Break the tie (in distance) by using the instance with the smaller index. Enter a number between 0 and 1.

x_i	-5	-4	-3	-2	8	10
y_i	0	0	1	0	0	1

+1 +1 +0 +0 +0 +0

$\frac{2}{6} = \frac{1}{3} = 33.3\%$

Question 2



- [4 points] There are 94 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
$\boxed{22}$	$\boxed{\text{Red}}$	$\boxed{\text{Yes}}$
? x	Red	No
?? y	Black	Yes
25	Black	No

$$P(B = \text{Red}, \text{Talk} = \text{Yes}) = P(B=R) \cdot P(T=Y)$$

$$\Rightarrow \frac{22}{94} = \frac{22+x}{94} \cdot \frac{22+y}{94} \quad \text{— (1)}$$

$$x + y = 94 - 22 - 25 = 47 \quad \text{— (2)}$$

$$\left.\begin{matrix} \text{(1)} \\ \text{(2)} \end{matrix}\right\} \quad x = 25, \; y = 22$$

Question 6

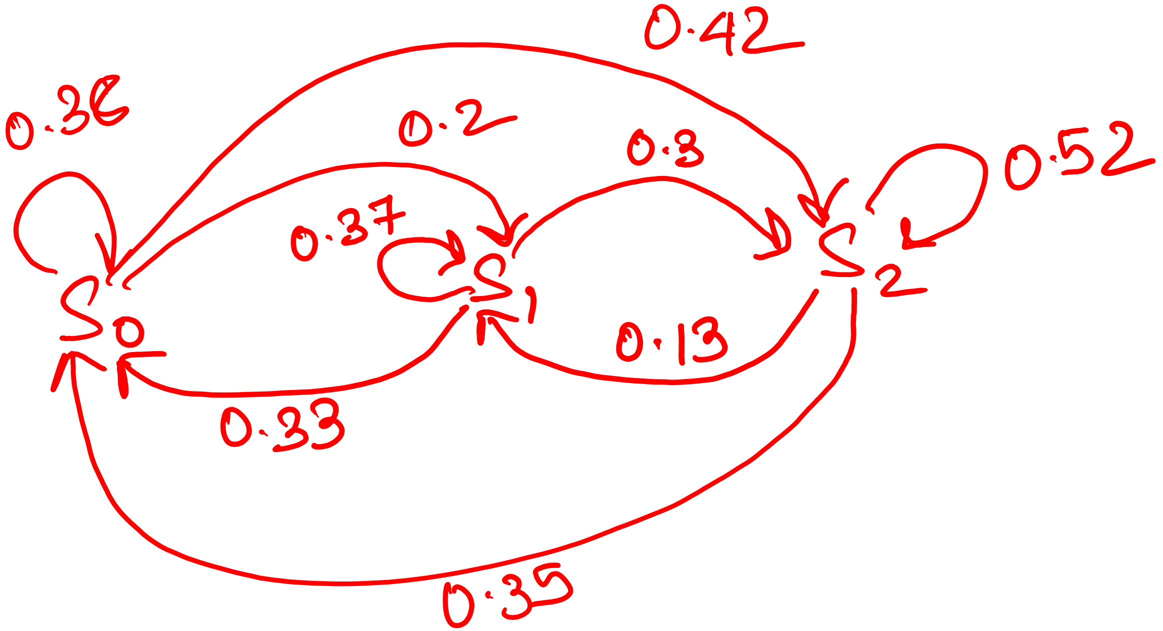
- [2 points] Given the following network $A \rightarrow B \rightarrow C$ where A can take on 4 values, B can take on 3 values, C can take on 4 values. Write down the minimum number of conditional probabilities that define the CPTs (Conditional Probability Table).

$3 + 2 \times 4 + 3 \times 3$

$(n_A - 1) + (n_B - 1) \times n_A + (n_C - 1 \times n_B)$

Question 10

• [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 0.38 chance that he will remain happy one hour later, a 0.2 chance to be hungry by then, and a 0.42 chance to have a wet diaper. Similarly, if he is hungry, one hour later he will be happy with 0.33 chance, hungry with 0.37 chance, and wet diaper with 0.3 chance. If he has a wet diaper, one hour later he will be happy with 0.35 chance, hungry with 0.13 chance, and wet diaper with 0.52 chance. He can smile (observation 0) or cry (observation 1). When he is happy, he smiles 0.47 of the time and cries 0.53 of the time; when he is hungry, he smiles 0.5 of the time and cries 0.12 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, smile (or Y_1, Y_2 = 1, 0) happens (in the first two periods)?



$$P(Y_1=1, Y_2=0)$$

$$= P(Y_1=1, Y_2=0, X_1, X_2)$$

$$= P(Y_1=1, Y_2=0, X_1=0, X_2)$$

$$= \sum_{x_2} P(Y_1=1 \mid X_1=0) \cdot P(Y_2=0 \mid x_2) \cdot P(x_2 \mid X_1=0) \underbrace{P(X_1=0)}_{1}$$