

Ques 1:

- a). Model was trained on data from stable economic conditions and tested on data from economic downturn. Training data distribution no longer matches testing distribution.
- b). High quality, diverse and representative data enables the model to capture general patterns. A simple model can generalize well with good data, while a complex model will fail with biased data.
- c). Large volume of data is normal and unlabeled. Very few labeled fraud cases exist. Clustering based anomaly detection algorithm can be used.
- d). Outcome is binary with a probability ρ of success. Bernoulli distribution can be used.

Ques 2:

a). Initial Parameters: $w=0.1$, $b=0$, $\alpha=0.05$

Applicant Data:

Applicant	Income(x)	Default(y)
A	$2 \rightarrow x^1$	$0 \rightarrow y^1$
B	$5 \rightarrow x^2$	$1 \rightarrow y^2$

Predicted Probability:

Applicant A, $z^1 = w x^1 + b = 0.1 \times 2 + 0 = 0.2$

$$\sigma(z^1) = \frac{1}{1+e^{-z^1}} = \frac{1}{1+e^{-0.2}} = \frac{1}{1+0.8187} = 0.5498$$

Applicant B, $z^2 = w x^2 + b = 0.1 \times 5 + 0 = 0.5$

$$\sigma(z^2) = \frac{1}{1+e^{-z^2}} = \frac{1}{1+e^{-0.5}} = \frac{1}{1+0.6065} = 0.6225$$

$$\hat{y}^1 = 0.55 \quad \hat{y}^2 = 0.62$$

Batch Gradient:

$$\frac{\partial L}{\partial w} = \frac{1}{m} \left(\sum_{i=1}^m (\hat{y}^{(i)} - \hat{f}^{(i)}) \right) x^{(i)}$$

$$= \frac{1}{m} \left(\sum_{i=1}^m (\hat{y}^{(i)} - y^{(i)}) \right) x^{(i)}$$

Applicant A, $(\hat{y}^1 - y^1) x^1 = (0.5498 - 0)(2) = 1.0996$

Applicant B, $(\hat{y}^2 - y^2) x^2 = (0.6225 - 1)(5) = -1.8875$

$$\frac{\partial L}{\partial w} = \frac{1}{2} (1.0996 - 1.8875) = \frac{-0.7879}{2} = -0.394$$

$$\frac{\partial L}{\partial w} = -0.394$$

b). SVM decision boundary $f(x) = 0.8x_1 - 1.5x_2 - 50 = 0$

Substitute new stock values:

$$\begin{aligned} f(x) &= 0.8 * 72 - 1.5 * 15 - 50 \\ &= -14.9 \end{aligned}$$

Buy class support vector $f(x) = 0.8 * 75 - 1.5 * 10 - 50$
 $= -5$

Sell class support vector $f(x) = 0.8 * 60 - 1.5 * 20 - 50$
 $= -32$

Importance of support vectors:

SV are data points closest to decision boundary. They are the critical points that constrain the optimal separating hyperplane.

Why removing non-support vectors doesn't affect the boundary:

Non-support vectors lie outside the margin. They don't contribute to the optimization constraints.

Ques 3:

a).

Avg Intra Distance measures how compact the cluster is.
(lower is better)

Closest Inter Distance measures how well the cluster is separated from other clusters (higher is better).

Among C₁, C₂ and C₃, C₂ has highest intra distance and lowest inter distance. C₂ shows the strongest indication of cluster assignment ambiguity.

b). Cumulative Variance:

$$PC_1 + PC_2 : 58 + 30 = 88\%$$

$$PC_1 + PC_2 + PC_3 : 58 + 30 + 9 = 97\%$$

$$PC_1 + PC_2 + PC_3 + PC_4 : 58 + 30 + 9 + 3 = 100\%$$

c).

Precision: among all predicted as +ve, how many are actually +ve.

Recall: among those which are actually +ve, how many were predicted +ve.

Security team goal is to increase Recall
and

decrease unnecessary alerts

↳ decrease False Positives → Increase Precision

Threshold	Precision	Recall
0.40	0.72	0.88
0.75	0.93	0.46

→ 88% Recall and 28% FP } Threshold
 → 46% Recall and 7% FP } 0.8
 } preferable

Ques 4:

a). Validation accuracy represents generalization to unseen data.

Method	Batch Size	Training Accuracy	Validation Accuracy
full Batches GD	120,000	72%	70%
Mini Batch GD	256	89%	87%
Stochastic GD	1	97%	75%

Mini Batch GD performs best.

	Validation Accuracy
No Regularization	87%
L2 Regularization	90%
Dropout ($p=0.4$)	92%

} Mini Batch with dropout is best choice.

b).

$$\text{Leaky ReLU } (a_i) = \begin{cases} z_i & z_i > 0 \\ 0.1z_i & z_i \leq 0 \end{cases}$$

$$z = \begin{bmatrix} 3 \\ -4 \\ 2 \end{bmatrix} \quad \frac{\partial L}{\partial a} = \begin{bmatrix} 5 \\ -1 \\ 4 \end{bmatrix}$$

Local gradient :

$$\frac{\partial a_i}{\partial z_i} = \begin{cases} 1 & z_i > 0 \\ 0.1 & z_i \leq 0 \end{cases}$$

$$z_1 = 3 > 0 \Rightarrow \frac{\partial a_1}{\partial z_1} = 1$$

$$z_3 = 2 > 0 \Rightarrow \frac{\partial a_3}{\partial z_3} = 1$$

$$z_2 = -4 < 0 \Rightarrow \frac{\partial a_2}{\partial z_2} = 0.1$$

feature matrix = $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

Shapley value (z)

Downstream Gradient

$$\frac{\partial L}{\partial z} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} = \begin{bmatrix} 5 \times 1 \\ -1 \times 0.1 \\ 4 \times 1 \end{bmatrix} = \begin{bmatrix} 5 \\ -0.1 \\ 4 \end{bmatrix}$$

c)

$$v(\emptyset) = 50 \quad v(\{c\}) = 60$$

$$v(\{e\}) = 70 \quad v(\{e, c\}) = 90$$

Case	Subsets without c	Contribution of c
1.	\emptyset	$v(\{c\}) - v(\emptyset) = 60 - 50 = 10$
2.	$\{e\}$	$v(\{e, c\}) - v(\{e\}) = 90 - 70 = 20$

Shapley value for feature c = $\sum_{z'} \frac{|z'|! (M-|z'|-1)!}{M!} [v(z' \cup c) - v(z')]$

features in subset
↓
Subset without c ↓ total # features

for case 1, $|z'|=0, M=2$

$$\frac{|z'|! (M-|z'|-1)!}{M!} = \frac{0! (2-0-1)!}{2!} = \frac{1}{2}$$

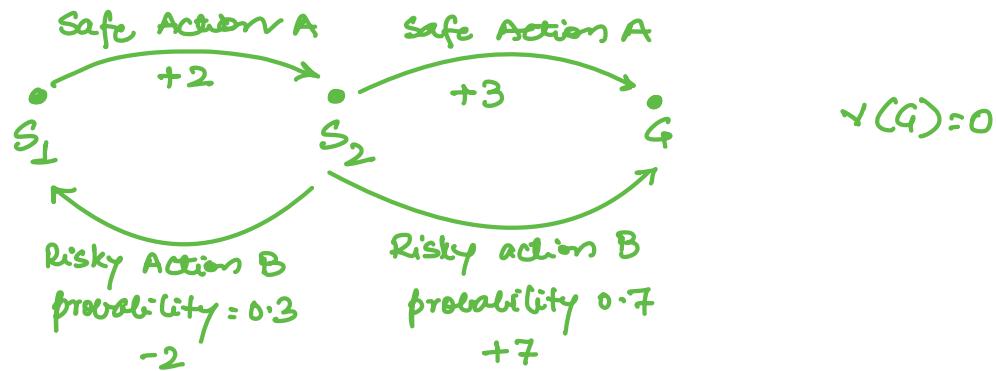
for case 2, $|z'|=1, M=2$

$$\frac{|z'|! (M-|z'|-1)!}{M!} = \frac{1! (2-1-1)!}{2!} = \frac{1}{2}$$

$$\phi_c = \frac{1}{2} \times 10 + \frac{1}{2} \times 20 = 5+10 = 15$$

Ques 5:

a).



$$\gamma = 0.9$$

Policy π always chooses safe (A) in both S_1 and S_2 .

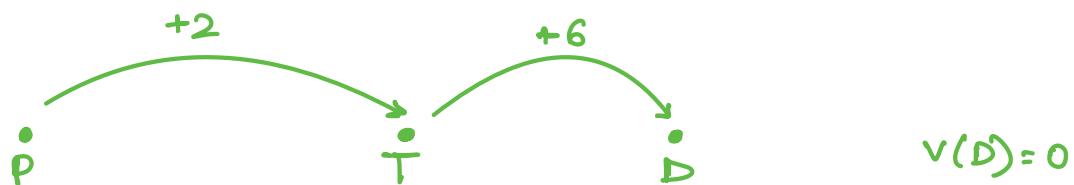
$$v^\pi(S_2) = 3 + \gamma v^\pi(G) = 3 + 0.9^* 0 = 3$$

$$v^\pi(S_1) = 2 + \gamma v^\pi(S_2) = 2 + 0.9^* 3 = 4.7$$

Risky (B) in S_2

$$\begin{aligned}
 Q^\pi(S_2, B) &= \sum_{S'} P(S'|S_2, B) [r(S_2, B, S') + \gamma v^\pi(S')] \\
 &= 0.7 [7 + 0.9 \times 0] + 0.3 [-2 + 0.9^* 4.7] \\
 &= 0.7^* 7 + 0.3^* 2.23 = 4.9 + 0.669 \\
 &= 5.569
 \end{aligned}$$

b).



Discount factor $\gamma = 0.5$
learning rate $\alpha = 0.5$

After first episode, $v_1(P) = 1, v_1(T) = 3$

$$\begin{aligned}
 v_2(p) &= v_1(p) + \alpha [r + \gamma v_1(\tau) - v_1(p)] \\
 &= 1 + 0.5 [2 + 0.5 * 3 - 1] \\
 &= 1 + 0.5 [2.5] = 1 + 1.25 = 2.25
 \end{aligned}$$

$$\begin{aligned}
 v_2(\tau) &= v_1(\tau) + \alpha [r + \gamma v_1(D) - v_1(\tau)] \\
 &= 3 + 0.5 [6 + 0.5 * 0 - 3] \\
 &= 3 + 0.5 [3] = 3 + 1.5 = 4.5
 \end{aligned}$$
