# Question bank on Supervised and Deep Learning

1. A real estate company is predicting house prices using features like area, number of rooms, and location. The model performs very well on training data but poorly on test data.

- Why might overfitting be happening here?
- How can **L2 regularization (Ridge Regression)** help in this scenario?

2. A telecom company uses logistic regression to predict whether customers will leave (churn). The dataset contains many features, including some irrelevant ones.

- Which type of regularization (L1 or L2) would be more suitable here and why?
- What advantage does **L1 (Lasso)** provide in this case?

3. A financial firm uses a machine learning model with hundreds of indicators (features) to predict stock price movement. The model is overfitting due to noise in data.

- How can **Elastic Net regularization** (combination of L1 and L2) be useful here?
- What is the advantage of combining both penalties instead of choosing just one?

4. **Perform Forward Propagation and Backward Propagation based on the below information.**

Input layer: 2 neurons (x1, x2)

Hidden layer: 1 neuron

Output layer: 1 neuron

Activation function: **Sigmoid**

**Given Values**

Inputs: x1 = 0.6, x2 = 0.1

Weights: w1 = 0.5, w2 = -0.4 (input → hidden)

Hidden bias: b1 = 0.1

Weight hidden → output: w3 = 0.3

Output bias: b2 = -0.2

Target (expected output): $t = 0.75$

Learning rate: $\eta = 0.5$

**Answer- w3new=0.3+0.01883=0.3188**

**b2new=−0.2+0.03198=−0.1680**

**w1new=0.5+0.00139=0.5014**

**w2new=−0.4+0.00023=−0.3998**

**b1new=0.1+0.00232=0.1023**

5. Compare Gradient Descent, Stochastic Gradient Descent (SGD), and Mini-batch Gradient Descent. When would you prefer each?
6. Why is the learning rate a critical hyperparameter in gradient-based optimization? What happens if it is too high or too low?
7. Why does SGD with momentum sometimes converge to a better minimum than Adam, even though Adam converges faster?
8. Explain the concept of a flat minimum vs. sharp minimum in loss landscapes. How do optimization algorithms influence which minimum is chosen?
9. Why does adaptive learning rate optimizers (Adam, RMSProp) sometimes generalize poorly compared to SGD?
10. What are the key differences between Batch Gradient Descent, Stochastic Gradient Descent (SGD), and Mini-batch Gradient Descent?
11. A sparse text classification model is underperforming. Which optimizer (SGD, AdaGrad, RMSProp, Adam) is best suited and why?
12. Explain the vanishing gradient problem in sigmoid and tanh activations.
13. Compare Sigmoid, ReLu, Tanh, softmax, softplus activation functions.
14. What is the significance of Learning Rate?
15. What is Cost function. Derive the cost function for Linear regression algorithm?
16. What is logistic regression. Derive its equation.
17. A binary classifier gives the following confusion matrix:

| | Predicted Positive | Predicted Negative |
|---|---|---|
| Actual Positive | 50 | 10 |
| Actual Negative | 5 | 35 |

**Compute Accuracy**
**Compute Precision**
**Compute Recall (Sensitivity)**
**Compute Specificity**
**Compute F1-score**

18. Out of 1000 patients tested for a rare disease:
    10 actually have the disease (positive)

990donot(negative)

The model predicts:

9 patients positive (8 true positives, 1 false positive)

991 patients negative (2 false negatives, 989 true negatives)

**Calculate Accuracy.**

**Calculate Precision, Recall, and F1-score.**

**Why might accuracy be misleading here?**

19. Why is a loss function important in training a neural network?

20. Differentiate between loss function and cost function.
21. Compare Mean Squared Error (MSE) and Mean Absolute Error (MAE). When would you prefer one over the other?
22. Explain the intuition behind Cross-Entropy Loss. Why is it preferred for classification tasks?
23. What is Hinge Loss? For which type of model is it typically used?
24. Explain the difference between binary cross-entropy and categorical cross-entropy.
25. What is Kullback-Leibler (KL) Divergence loss? How is it used in deep learning?
26. Why are smooth approximations like Huber Loss preferred in regression with outliers?
27. **A dataset of 4 points: -**

| x | y |
|---|---|
| 1 | 2 |
| 2 | 3 |
| 3 | 5 |
| 4 | 4 |

- Fit a linear regression model:

  $y=w \cdot x+b$

  Compute slope w and intercept b using least squares formula.

- Predict y for x=5.

28. **A dataset is given: -**

| x | y |
|---|---|
| 0 | 0 |
| 1 | 0 |
| 2 | 1 |
| 3 | 1 |

**Fit a logistic regression model y^=σ(wx+b), with sigmoid activation σ(z)=1/ 1+e-z**

**Suppose after training: w=1, b=−2.**

**Compute predicted probabilities for x=0,1,2,3.**

**Classify using threshold = 0.5.**