Here is a set of questions and answers about loss functions, structured for easy review and including common interview questions.

---

1. Fundamental Concepts

Q: What is a loss function?

A: A loss function is a mathematical function that measures the error—or "loss"—between a model's prediction and the actual, true value.

Q: What is the main purpose of a loss function in training?

A: Its purpose is to provide a signal for optimization. The optimizer (like Gradient Descent) uses the gradient (derivative) of the loss function to update the model's parameters in a way that minimizes the error.

Q: What is the difference between a "loss function" and a "cost function"?

A: They are often used interchangeably. Technically, a loss function applies to a single training example, while a cost function is the average loss over the entire training batch or dataset (e.g., MSE is a cost function).

---

2. Regression Losses

Q: What is Mean Squared Error (MSE)?

A: It is the average of the squared differences between predictions and true values.

- **Formula:** $L_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Q: What is Mean Absolute Error (MAE)?

A: It is the average of the absolute (non-negative) differences between predictions and true values.

- **Formula:** $L_{\text{MAE}} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$

Q: When would you use MAE instead of MSE?

A: Use MAE when your dataset has significant outliers. MSE squares the error, so large errors from outliers can dominate the loss and skew the model. MAE is more robust to these outliers.

---

3. Classification Losses

Q: What is Binary Cross-Entropy (BCE)?

A: It is the standard loss function for binary classification (two classes). It measures the difference between the true label (0 or 1) and the model's predicted probability (a value between 0 and 1).

Q: What activation function is paired with Binary Cross-Entropy?

A: The Sigmoid activation function, because it outputs a single value between 0 and 1, which is required for BCE.

Q: What is Categorical Cross-Entropy (CCE)?

A: It is the standard loss function for multi-class classification (three or more classes). It measures how well the model's predicted probability distribution matches the true class.

Q: What activation function is paired with Categorical Cross-Entropy?

A: The Softmax activation function. It converts the model's raw outputs (logits) into a probability distribution where all class probabilities sum to 1.

Q: What is Hinge Loss?

A: A loss function used for "max-margin" classifiers, most famously Support Vector Machines (SVMs). It aims to find the best decision boundary by penalizing any predictions that are incorrect or "not confident enough" (i.e., too close to the boundary).

---

4. Interview-Style Questions (Conceptual)

Q: (⭐ Important) What's the difference between a loss function and an evaluation metric?

A:

- **Loss Function:** Used for *training* the model. It must be differentiable so the optimizer can use its gradient. Example: MSE.
- **Evaluation Metric:** Used for *evaluating* the model's performance in a human-understandable way. It does *not* need to be differentiable. Example: Accuracy.

You train a model by minimizing MSE (loss), but you *report* its performance using R-squared or Accuracy (metric).

Q: Your model has 99% accuracy, but it's performing terribly in production. What is the most likely reason?

A: The data is severely imbalanced.

For example, in a fraud detection (1% fraud) or cancer diagnosis (1% positive) task, a model that always predicts "not fraud" or "no cancer" will be 99% accurate but is completely useless.

In this case, accuracy is the wrong metric. You should use a metric like F1-Score or Precision-Recall and a loss function like Weighted Cross-Entropy.

Q: You are training a regression model, and you see the loss is being dominated by a few data points. What is happening, and what can you do?

A:

- **What is happening:** You are likely using **Mean Squared Error (MSE)**, and your dataset has significant outliers. MSE squares large errors, so these outliers are creating an exploding loss.
- **What to do:** Switch to a loss function that is more robust to outliers, such as **Mean Absolute Error (MAE)** or **Huber Loss**.

Q: What does it mean for a loss function to be "convex"? Why is this important?

A: A convex function has a "bowl" shape, meaning it has only one global minimum. This is important because it guarantees that an optimizer (like Gradient Descent) can find the single best possible solution.

Simple models like linear regression have convex loss functions. Deep neural networks, however, are highly **non-convex**, meaning they have many local minima, and optimization is about finding a "good enough" solution, not necessarily the single best one.