⍰ How does logistic regression differ from linear regression?

Linear regression is used for predicting continuous outcomes (e.g., house prices) by fitting a straight line to the data, minimizing the sum of squared errors. It outputs real numbers without bounds. Logistic regression, however, is for binary classification (e.g., spam vs. not spam), predicting probabilities between 0 and 1. It uses the sigmoid function to transform the linear output into a probability, and optimizes using maximum likelihood estimation rather than least squares. While both are linear models, logistic regression handles categorical outcomes and avoids predicting invalid values like negative probabilities.

⍰ What is the sigmoid function?

The sigmoid function, also called the logistic function, is a mathematical function that maps any real-valued number to a value between 0 and 1. It's defined as: σ(x)=11+e−x \sigma(x) = \frac{1}{1 + e^{-x}} σ(x)=1+e−x1 where e e e is the base of the natural logarithm. In logistic regression, it squashes the linear combination of inputs into a probability. It has an S-shaped curve, approaching 0 as x x x goes to negative infinity and 1 as x x x goes to positive infinity.

⍰ What is precision vs recall?

Precision and recall are evaluation metrics for classification models, especially in binary settings:

- **Precision** measures the accuracy of positive predictions: the proportion of true positives among all predicted positives (TP / (TP + FP)). It's useful when false positives are costly (e.g., spam detection).
- **Recall** (also called sensitivity) measures how well the model captures actual positives: the proportion of true positives among all actual positives (TP / (TP + FN)). It's key when false negatives are costly (e.g., disease detection). There's often a trade-off between them; the F1-score (harmonic mean of precision and recall) balances both.

⍰ What is the ROC-AUC curve?

The ROC (Receiver Operating Characteristic) curve plots the true positive rate (recall) against the false positive rate at various threshold settings for a binary classifier. It shows the model's ability to distinguish between classes across thresholds. AUC (Area Under the Curve) is the area beneath the ROC curve, ranging from 0 to 1: 0.5 indicates random guessing, while 1.0 means perfect classification. Higher AUC means better performance, especially for imbalanced datasets, as it's threshold-independent.

⍰ What is the confusion matrix?

A confusion matrix is a table that summarizes the performance of a classification model by comparing predicted vs. actual labels. For binary classification, it's a 2x2 grid:

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

	Predicted Positive	Predicted Negative

From this, metrics like accuracy ((TP + TN) / total), precision, recall, and F1-score are derived. It's extendable to multi-class problems with larger matrices.

□ What happens if classes are imbalanced?

In imbalanced datasets (e.g., 95% negative, 5% positive), models like logistic regression may bias toward the majority class, leading to high accuracy but poor minority class performance (e.g., low recall for positives). This can result in misleading evaluations, as the model might simply predict the majority class most of the time. Solutions include resampling (oversampling minority or undersampling majority), using class weights in the model, focusing on metrics like precision-recall or ROC-AUC instead of accuracy, or employing techniques like SMOTE for synthetic data generation.

□ How do you choose the threshold?

In logistic regression, the default threshold is 0.5: predict positive if probability ≥ 0.5. However, choose based on the problem's needs—e.g., lower it to increase recall (catch more positives, but more false positives) or raise it for higher precision. Use tools like the ROC curve to find the threshold maximizing the Youden's J statistic (sensitivity + specificity - 1), precision-recall curve for imbalanced data, or cost-based analysis if false positives/negatives have different costs. Cross-validation helps validate the choice.

□ Can logistic regression be used for multi-class problems?

Yes, but it requires extensions since basic logistic regression is binary. Common approaches:

- **One-vs-Rest (OvR)**: Train one binary classifier per class (class vs. all others), then pick the class with the highest probability.
- **Multinomial Logistic Regression (Softmax)**: Generalizes the sigmoid to multiple classes using the softmax function to output probabilities summing to 1. Libraries like scikit-learn handle this automatically with the 'multinomial' solver. It's effective for mutually exclusive classes but assumes linear decision boundaries.