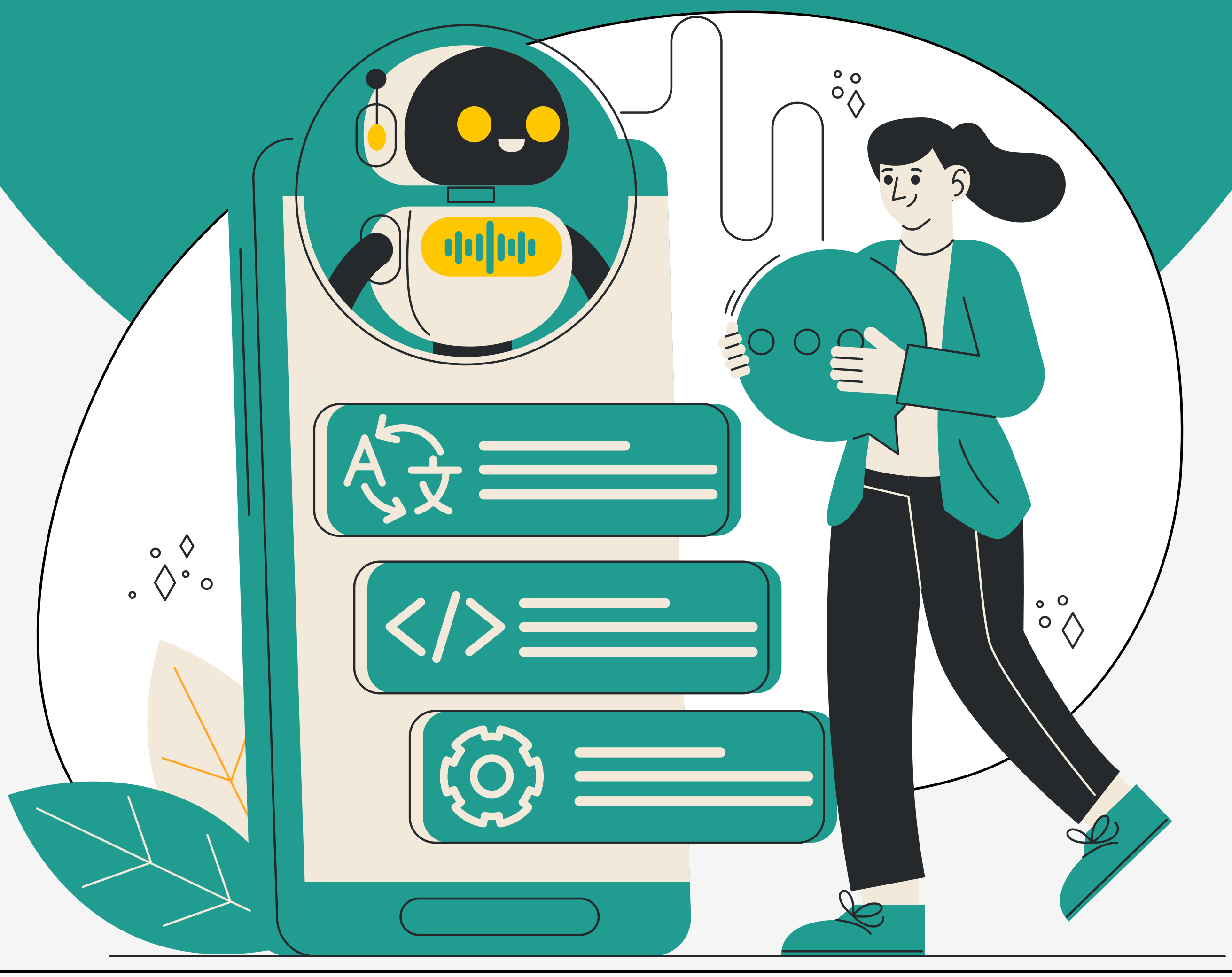
Logistic Regression

Interview Questions -1

(Practice Project)



PW SKILLS

Easy:

1. Q: How does logistic regression handle non-linear relationships between features and the target variable?

Answer : Logistic regression doesn't directly handle non-linear relationships. However, you can incorporate non-linearity by:

- Adding polynomial features
- Using interaction terms
- Applying non-linear transformations to features (e.g., log, square root)
- Binning continuous variables

These techniques allow logistic regression to capture more complex relationships within its linear framework.

2. Q: Explain why logistic regression is called "regression" when it's used for classification tasks.

Answer : Despite being primarily used for classification, logistic regression is called "regression" because:

- It predicts the probability of class membership (a continuous value between 0 and 1)
- The underlying model is similar to linear regression, using a linear combination of features
- The decision boundary is linear in the feature space
- The "logistic" part refers to the logistic function used to transform the linear combination of features into a probability.

3. Q: How does logistic regression perform multi-class classification?

Answer : Logistic regression can be extended to multi-class problems using two main approaches:

1. **One-vs-Rest (OvR):** Train binary classifiers for each class against all others
2. **Multinomial Logistic Regression:** Use softmax function instead of sigmoid, predicting probabilities for all classes simultaneously

The choice depends on the problem and the number of classes, with multinomial generally preferred for problems with mutually exclusive classes.

4. Q: What is the difference between L1 and L2 regularization in logistic regression, and when would you choose one over the other?

Answer : L1 (Lasso) and L2 (Ridge) regularization differ in their effects:

- L1 adds the absolute value of coefficients to the loss function, promoting sparsity (some coefficients become zero)
- L2 adds the squared value of coefficients, shrinking all coefficients but rarely to zero

Choose L1 when:

- Feature selection is important
- You suspect many features are irrelevant

Choose L2 when:

- You want to prevent overfitting but keep all features
- Multicollinearity is present in your data

5. Q: How do you interpret the coefficients in logistic regression?

Answer : In logistic regression, coefficients represent:

- The change in log-odds of the positive class for a one-unit increase in the corresponding feature, holding other features constant
- Positive coefficients increase the probability of the positive class, negative coefficients decrease it
- The magnitude indicates the feature's importance
- To interpret in terms of odds, exponentiate the coefficient. For probability, the relationship is non-linear due to the sigmoid function.

6. Q: Explain the concept of maximum likelihood estimation in logistic regression.

Answer : Maximum Likelihood Estimation (MLE) in logistic regression:

- Aims to find the model parameters that maximize the likelihood of observing the given data
- Uses the likelihood function, which is the product of probabilities for each observation
- Often works with log-likelihood for computational efficiency
- Is solved iteratively, typically using optimization algorithms like gradient descent
- MLE provides a principled way to estimate parameters that best fit the observed data.

7. Q: How does logistic regression handle outliers, and what are some strategies to mitigate their impact?

Answer : Logistic regression can be sensitive to outliers. Strategies to handle them include:
- Removing clear outliers if they're errors
- Winsorization: Capping extreme values
- Using robust logistic regression methods (e.g., Huber loss)
- Transforming features (e.g., log transformation)
- Using regularization to reduce the model's sensitivity to extreme values

The choice depends on the nature of outliers and the specific problem context.

Medium:

8. Q: What is the difference between a generative and discriminative model, and where does logistic regression fit?

Answer : Generative vs. Discriminative models:
- Generative models learn the joint probability P(X,Y), allowing generation of new samples
- Discriminative models learn the conditional probability P(Y|X), focusing on the decision boundary

Logistic regression is a discriminative model. It directly models the probability of the output class given the input features, without attempting to model how the data was generated.

9. Q: How does the choice of threshold affect the performance metrics in logistic regression, and how would you choose an optimal threshold?

Answer : Threshold choice impacts precision, recall, and F1-score:

- **Lower threshold:** Higher recall, lower precision
- **Higher threshold:** Lower recall, higher precision
- To choose an optimal threshold:
- Use ROC curve and choose based on the desired trade-off between TPR and FPR
- Use precision-recall curve for imbalanced datasets
- Consider business requirements and costs of false positives vs. false negatives
- Use techniques like Youden's J statistic or F1-score maximization

10. Q: Explain the concept of cross-entropy loss in logistic regression and why it's preferred over squared error loss.

Answer : Cross-entropy loss in logistic regression:

- Measures the difference between predicted probability distribution and true distribution
- **Is defined as:** -[y*log(p) + (1-y)*log(1-p)], where y is the true label and p is the predicted probability

Is preferred because:

1. It's derived from the maximum likelihood principle for Bernoulli distribution
2. It provides stronger gradients for incorrect predictions, leading to faster learning
3. Squared error can lead to slower convergence and suboptimal solutions for classification tasks

11. Q: How would you handle multicollinearity in logistic regression?

Answer : Strategies to handle multicollinearity:

1. Identify multicollinearity using correlation matrix or VIF (Variance Inflation Factor)
2. Remove one of the highly correlated features
3. Combine correlated features (e.g., using PCA)
4. Use L1 regularization (Lasso) to perform feature selection
5. Use L2 regularization (Ridge) to reduce the impact of multicollinearity
6. Collect more data if possible

The choice depends on the specific problem and the nature of the multicollinearity.

12. Q: Explain the difference between odds and probability in the context of logistic regression.

Answer : Odds vs. Probability:

- **Probability:** The chance of an event occurring, ranging from 0 to 1
- **Odds:** The ratio of the probability of success to the probability of failure
- **Relationship:** odds = p / (1-p), where p is the probability
- **Log-odds (logit):** log(odds) = log(p / (1-p))

Logistic regression models the log-odds as a linear function of features. The sigmoid function converts log-odds back to probabilities.

13. Q: What is the Newton-Raphson method in logistic regression, and how does it differ from gradient descent?

Answer : Newton-Raphson method in logistic regression:

- Is an optimization algorithm used to find the maximum likelihood estimates
- Uses both first and second derivatives of the likelihood function
- Generally converges faster than gradient descent, especially near the optimum
- Each iteration is more computationally expensive than gradient descent
- Is less sensitive to learning rate choice

Gradient descent only uses the first derivative and may require more iterations but is computationally cheaper per iteration.

Hard:

14. Q: How would you handle class imbalance in logistic regression, and what are the pros and cons of different approaches?

Answer : Approaches to handle class imbalance:

1. **Resampling:**
- Oversampling minority class (pro: no data loss, con: risk of overfitting)
- Undersampling majority class (pro: faster training, con: potential loss of information)
- SMOTE (Synthetic Minority Over-sampling Technique)

2. **Class weighting:** Assign higher weights to minority class (pro: uses all data, con: may not work well for extreme imbalance)

3. Adjust classification threshold (pro: simple, con: doesn't address fundamental imbalance in training)

4. Use different evaluation metrics (e.g., F1-score, AUC-ROC)
The choice depends on the degree of imbalance and specific problem requirements.

15. Q: Explain the concept of deviance in logistic regression and how it's used for model evaluation.

Answer : Deviance in logistic regression:

- Measures the lack of fit between the model and the data
- Is based on the likelihood ratio test
- **Null deviance:** Deviance of a model with only the intercept
- **Residual deviance:** Deviance of the fitted model
- The difference (null deviance - residual deviance) follows a chi-square distribution
- Used to assess model fit and compare nested models

A significant reduction in deviance indicates that the model is explaining a substantial amount of variation in the data.

16. Q: How does logistic regression perform feature selection inherently, and what are the limitations of this approach?

Answer : Logistic regression can perform feature selection through:

1. **Coefficient magnitude:** Larger absolute values indicate more important features
2. **L1 regularization:** Drives some coefficients to exactly zero Limitations:

- Doesn't account for feature interactions
- Can be misleading in the presence of multicollinearity
- May not capture non-linear relationships
- Sensitive to the scale of features

More sophisticated feature selection techniques (e.g., recursive feature elimination) might be necessary for complex problems.

17. Q: What is the difference between discriminative power and calibration in logistic regression, and why are both important?

Answer : Discriminative power vs. Calibration:

- **Discriminative power:** Ability to distinguish between classes (measured by AUC-ROC, accuracy)
- **Calibration:** Accuracy of predicted probabilities (how well they match observed frequencies)

Both are important because:

- Good discrimination ensures effective classification
- Good calibration is crucial for risk assessment and decision-making based on probabilities

A model can have good discriminative power but poor calibration, or vice versa. Techniques like Platt scaling can improve calibration without affecting discrimination.

18. Q: Explain the concept of the bias-variance tradeoff in the context of logistic regression.

Answer : Bias-variance tradeoff in logistic regression:

- **Bias:** Error from incorrect assumptions in the learning algorithm
- **Variance:** Error from sensitivity to small fluctuations in the training set
- **High bias (underfitting):** Model is too simple, misses important patterns
- **High variance (overfitting):** Model is too complex, captures noise in training data

Balancing this tradeoff:

Regularization can reduce variance at the cost of slightly increased bias
Feature engineering can reduce bias by allowing the model to capture more complex patterns
Cross-validation helps in finding the right balance

19. Q: How does the choice of link function affect logistic regression, and what are some alternatives to the logit link?

Answer : Link function in logistic regression:

- Connects the linear predictor to the mean of the distribution function
- **Logit link:** $\log(p/(1-p)) = \beta^T x$, most common for binary classification

Alternatives:

1. **Probit link:** $\Phi^{-1}(p) = \beta^T x$, where Φ is the cumulative normal distribution
2. **Complementary log-log:** $\log(-\log(1-p)) = \beta^T x$

Choice affects:

- Interpretation of coefficients
- Behavior at extreme probabilities
- Robustness to outliers

Probit is similar to logit but can be preferred in some fields (e.g., econometrics). Complementary log-log is asymmetric and can be useful for certain types of data.

20. Q: Describe the concept of separation in logistic regression and how it can be addressed.

Answer : Separation in logistic regression:

Complete separation: A feature or combination of features perfectly predicts the outcome
Quasi-complete separation: A feature or combination nearly perfectly predicts the outcome

Problems:
Leads to unstable or infinite coefficient estimates
Standard errors become very large

Addressing separation:

1. **Use regularization** (L1 or L2) to constrain coefficient values
2. **Firth's method:** Applies a modified score function to reduce bias in estimates
3. Exact logistic regression for small samples
4. Collect more data if possible
5. Combine rare categories in categorical predictors

Recognizing and addressing separation is crucial for obtaining reliable estimates in logistic regression.