Logistic Regression

For a better understanding of logistic regression and probably helping interviews.

- Definition
  - It is similar to the binary classification we are familiar with.
  - It predicts the **categorical** dependent variables, usually binary. The dependent variables can also be multinomial or ordinal. Sklearn describes multinomial logistic regression with details.
    - https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

- Assumptions
  - The Observations are Independent (check with residual plot)
  - There is No Multicollinearity Among Explanatory (check with Variables variance inflation factor VIF)
  - There are No Extreme Outliers
  - p(x) = exp(b0+b1*x)/(1+exp(b0+b1*x)). Logit = log(p/(1-p))=b0+b1*x

- Two phases
  - Use the logistic function(sigmoid) g(z) = 1/(1+exp(-z)) to calculate the probability(0~1), z=b0+b1*x.
  - Predict the class of a test sample based on the threshold.

- Maximum likelihood estimation
  - Why log likelihood?
    - If we use MSE like Linear Regression, the loss function may not be convex. So here we use log likelihood as the loss function as it is convex.
  - Use MLE to fit the model:

    - Betas, x, y —> likelihood function
    - Estimate betas to maximize likelihood
    - Can use gradient descent to update betas and use mini-batch to improve the efficiency. Mini-batch may bring bad gradients if the data is noisy but we can still optimize them

- Why logistic instead of linear regression?
  - Logistic regression can predict categorical Ys.
  - Linear regression:
    - Hard to fit the data especially with outliers.
    - Linear regression predictions has the meaning of order and implies that the differences between different categories are the same. But in fact the difference should not be the same.
    - Linear regression can generate probability>1 or <0.

- Other tips:
  - Logistic regression is also linear because it has a linear decision boundary.

    p(x) = exp(b0+b1*x)/(1+exp(b0+b1*x)). Logit = log(p/(1-p))=b0+b1*x

    if we have p(x)=exp(1+x), and threshold=0.5, then exp(1+x)=0.5. The boundary is linear.
  - For linear separable data, logistic regression may generate probability very close to 1/0, and it won't stop unless you set the iterations when using sklearn packages.

See logistic regression implementations:

https://github.com/Alleria1809/LogisticRegressionImplementation

References:

https://www.statology.org/assumptions-of-logistic-regression/