Logistic Regression

Description: Models the probability of a binary outcome using a logistic function. Predictors can be continuous or categorical.

Pros:

- Interpretable coefficients (as odds ratios)
- Fast to train and easy to implement
- Works well when classes are linearly separable
- Handles both numerical and categorical variables

Cons / Cautions:

- Assumes a linear relationship in log-odds
- Limited flexibility for complex boundaries
- Sensitive to multicollinearity

Best For:

- When interpretability is important (e.g., medical risk modeling)
- Binary classification with moderate-sized datasets
- **Example**: Predicting whether a patient has diabetes based on BMI and blood pressure
- **Example**: Will a customer click on an ad? Yes/No

Linear Discriminant Analysis (LDA)

Description: Assumes normally distributed predictors with equal class covariances. Produces linear decision boundaries.

Pros:

- Low variance
- Stable with small sample sizes
- Can outperform logistic regression when assumptions hold

Cons / Cautions:

- Assumes equal variance-covariance matrices across classes
- Assumes normality of predictors
- Not suitable for non-linear boundaries

Best For:

- Small datasets with Gaussian features
- Structured settings with interpretable variables
- **Example**: Classifying handwritten digits when the features are pixel averages
- **Example**: Predicting whether a loan applicant is low or high risk based on age and income

Quadratic Discriminant Analysis (QDA)

Description: Extension of LDA that allows each class to have its own covariance matrix. Produces quadratic boundaries.

Pros:

- More flexible than LDA
- Can model complex class boundaries
- Lower bias in many cases

Cons / Cautions:

- High variance with small sample sizes
- Needs more parameters → more data
- Sensitive to outliers

Best For:

- Larger datasets where the decision boundary is curved
- **Example**: Classifying whether a tumor is malignant or benign based on complex cell shape features

- **Example**: Differentiating between types of credit card fraud based on transaction patterns

Naive Bayes

Description: Uses Bayes' theorem assuming predictors are conditionally independent given the class.

Pros:

- Extremely fast and simple
- Performs well in high dimensions
- Robust to irrelevant features
- Works well with small datasets

Cons / Cautions:

- Strong (and often unrealistic) independence assumption
- Probability estimates may be poor
- Less flexible than more complex models

Best For:

- Text classification and spam filtering
- High-dimensional or sparse data
- **Example**: Email spam detection based on words in an email
- **Example**: Sentiment analysis (positive/negative) in product reviews

K-Nearest Neighbors (KNN)

Description: Non-parametric method that assigns class based on the majority vote of the k closest points.

Pros:

- No assumptions about the data distribution
- Can capture complex, non-linear decision boundaries

- Easy to understand and implement

Cons / Cautions:

- Very sensitive to the choice of k and irrelevant features
- Poor performance in high dimensions (curse of dimensionality)
- Requires large datasets to be effective
- No model interpretability

Best For:

- Low-dimensional, well-behaved data
- Complex but smooth patterns
- **Example**: Classifying handwritten digits based on raw pixel values
- **Example**: Recommending movies based on user similarity

General Guidelines for Model Selection

Scenario	**Recommended Model(s)**
Need interpretability	Logistic Regression, LDA
Decision boundary is likely non-linear	QDA, KNN
Small sample size and many predictors	Naive Bayes
High-dimensional, sparse data (e.g., text)	Naive Bayes
Classes are approximately normal with similar variance	LDA
Quick prototype or baseline model	Logistic Regression, Naive Bayes
Avoid KNN when...	Dimensions are high or data is sparse

Key Takeaways

- **Logistic Regression**: Great baseline, interpretable, best when decision boundaries are simple and linear.
- **LDA**: Works well with small samples if normality holds, more stable than logistic regression.
- **QDA**: More flexible than LDA but can overfit unless you have lots of data.
- **Naive Bayes**: Surprisingly powerful for high-dimensional problems like text classification.
- **KNN**: Flexible and non-parametric, but impractical in high dimensions.