4.5 A Comparison of Classification Methods

Dr. Lauren Perry

Classification Methods

We discussed

- Logistic regression
- LDA
- QDA
- Naive Bayes
- KNN

Here, we dig into the performance of these five methods.

Why discriminant analysis?

- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. Linear discriminant analysis does not suffer from this problem.
- If n is small and the distribution of the predictors X is approximately normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model.
- LDA is popular when we have more than two response classes, because it also provides low-dimensional views of the data.

Which method is best?

- No one method uniformly dominates the others.
- In any setting, choice of method will depend on
  - the distribution of the predictors in each of the K classes
  - the values of n and p

Logistic Regression

Assumptions:

- Linear decision boundary
- No assumption about feature distribution
- Models $P(Y|X)$ directly

Logistic Regression

When to use:

- Classes are approximately linearly separable
- You want interpretable coefficients
- You care about well-calibrated probabilities
- Moderate sample size
- Features may not follow a Gaussian distribution

Logistic Regression

Strengths

- Interpretability
- Simple and robust
- Works well in high dimensions
- Handles correlated features better than Naive Bayes

Weaknesses

- Cannot model nonlinear boundaries unless you engineer features

LDA

Assumptions

- Gaussian class-conditional distributions
- Equal covariance matrices across classes
- Linear decision boundary
- Models $P(X|Y)$ (generative model)

LDA

When to use:

- Data is approximately normal
- Classes have similar covariance structure
- Small-to-moderate sample size
- You want dimension reduction and classification

LDA

Strengths

- Lower variance than logistic regression when assumptions hold
- Works well with small samples
- Computationally efficient

Weaknesses

- Sensitive to violation of equal covariance assumption
- Linear boundary only

QDA

Assumptions:

- Normality of class distributions
- Different covariance matrices per class
- Quadratic decision boundary

QDA

When to use:

- Normal data
- Class covariances differ substantially
- Enough data to estimate covariance matrices reliably

QDA

Strengths

- More flexible than LDA
- Captures nonlinear boundaries

Weaknesses

- High variance
- Requires a lot of data
- Prone to overfitting

Naive Bayes

Assumptions:

- Conditional independence of features given the class
- Often Gaussian (for continuous features)
  - . . . or some other distribution

Naive Bayes

When to use:

- High-dimensional data (e.g., text classification)
- Small datasets
- Need fast training
- Independence assumption roughly holds

Naive Bayes

Strengths

- Very fast
- Works surprisingly well even when independence assumption is violated
- Very good for text and sparse data

Weaknesses

- Can have poor probability calibration
- Can perform badly if strong feature correlations exist

KNN

Assumptions:

- No distributional assumptions
- Local similarity structure matters

KNN

When to use:

- Nonlinear decision boundary
- Low-dimensional data
- Large dataset ($n >> p$)
- You don't want parametric assumptions

KNN

Strengths

- Very flexible
- Captures complex patterns
- Simple conceptually

Weaknesses

- Computationally expensive at prediction time
- Performs poorly in high dimensions
- Sensitive to feature scaling

Overall Summary of Methods

- Logistic regression is very popular, especially when $K = 2$.
- LDA is useful when n is small, or the classes are well-separated, and normality is a reasonable assumption.
- QDA is best for nonlinear boundaries where normality is a reasonable assumption.
- Naive Bayes is useful when p is very large.
- KNN will dominate when the decision boundary is highly non-linear and $n >> p$.