Discriminant Analysis

Classification with Statistical Foundations

MA2003B - Multivariate Methods in Data Science

Dr. Juliho Castillo

Tecnologico de Monterrey

2025-10-17

The Classification Problem

Scenario: E-commerce company with thousands of customers

Question: Which segment does each customer belong to?

High-Value	Loyal	Occasional
Premium customers	Regular customers	Infrequent buyers
High spending	Moderate spending	Low engagement
Max engagement	Consistent activity	Need re-engagement

Real-World Applications

Business & Marketing

- Customer segmentation
- Credit risk assessment
- Churn prediction

Healthcare

- Disease diagnosis
- Treatment prediction
- Medical imaging

Manufacturing

- Quality control
- Defect classification
- Fault detection

Sports Analytics

- Athlete classification
- Talent identification
- Performance assessment

The Core Idea

Discriminant Analysis finds discriminant functions

Linear or quadratic combinations of predictors that **best separate groups**

Think of it as finding the **“best viewing angle”**

to distinguish groups in multidimensional space

Mathematical Framework

Setup:

- g distinct groups or populations
- p predictor variables per observation
- Training data with known group memberships

Key Notation:

- $\boldsymbol{x} = \left(x_1, ..., x_p\right)^\top$ predictor vector
- π_k prior probability of group k
- $\boldsymbol{\mu}_k$ mean vector for group k
- $\boldsymbol{\Sigma}_k$ covariance matrix for group k
- $f_k(\boldsymbol{x})$ probability density for group k

Bayes Theorem Foundation

Posterior probability of group k:

$$P(G = k \mid \boldsymbol{x}) = \frac{f_k(\boldsymbol{x})\pi_k}{\sum_{j=1}^{g} f_j(\boldsymbol{x})\pi_j}$$

Bayes Classification Rule:

Assign to group k^* that maximizes posterior probability:

$$k^* = \arg\max_k f_k(\boldsymbol{x})\pi_k$$

This is **optimal** under correct distributional assumptions

Example: Credit Risk - Setup

Business Context:

Bank evaluating loan application. Two possible outcomes:

- Group 0: Customer will **not default** (repay loan)
- Group 1: Customer will **default** (fail to repay)

Applicant Profile:

- Annual income: 50,000 USD
- Debt-to-income ratio: 0.4 (40%)
- Credit score: 650

Historical Data (Prior Probabilities):

- $\pi_0 = 0.95$ (95% of past customers did not default)

- $\pi_1 = 0.05$ (5% of past customers defaulted)

Example: Credit Risk - Likelihood

Probability Densities:

How likely is this profile in each group?

No Default Group ($k = 0$):

$f_0(\boldsymbol{x}) = 0.0008$

This profile is **uncommon** among non-defaulters (lower income, higher debt)

Default Group ($k = 1$):

$f_1(\boldsymbol{x}) = 0.0030$

This profile is **more typical** among defaulters (3.75 times more likely)

Example: Credit Risk - Calculation

Step 1: Calculate numerators (prior times likelihood)

- No default: $f_0(\boldsymbol{x}) \times \pi_0 = 0.0008 \times 0.95 = 0.00076$
- Default: $f_1(\boldsymbol{x}) \times \pi_1 = 0.0030 \times 0.05 = 0.00015$

Step 2: Calculate denominator (sum of numerators)

$$\text{Total} = 0.00076 + 0.00015 = 0.00091$$

Step 3: Calculate posterior probabilities

- $P(\text{no default}|\boldsymbol{x}) = \frac{0.00076}{0.00091} = 0.835\ (83.5\%)$
- $P(\text{default}|\boldsymbol{x}) = \frac{0.00015}{0.00091} = 0.165\ (16.5\%)$

Example: Credit Risk - Interpretation

Key Insight:

Even though this profile is **3.75x more common** among defaulters…

The **prior probability** (95% vs 5%) is so strong that we still classify as **no default**

Decision Rule:

Classify as **no default** (83.5% > 16.5%)

Business Implications:

- Approve loan, but consider higher interest rate
- Monitor account more closely
- May require additional collateral

- 16.5% risk is still significant for portfolio management

Linear Discriminant Analysis (LDA)

Two Critical Assumptions:

1. Multivariate Normality

Each group follows multivariate normal distribution

2. Equal Covariances

$$\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \ldots = \mathbf{\Sigma}_g = \mathbf{\Sigma}$$

Result: **Linear** decision boundaries

LDA: Discriminant Scores

Score for group k:

$$\delta_k(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log(\pi_k)$$

Classification Rule:

Assign $\boldsymbol{x}$ to group with largest $\delta_k(\boldsymbol{x})$

Geometric Interpretation:

- Decision boundaries are hyperplanes
- Perpendicular bisectors when priors are equal

Fisher's Approach

Alternative (equivalent) formulation:

Maximize ratio of between-group to within-group variance

For two groups:

$$\text{maximize} \quad \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_1^2 + s_2^2}$$

where $y = \boldsymbol{a}^\top \boldsymbol{x}$

Solution: $\boldsymbol{a} \propto \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

Quadratic Discriminant Analysis (QDA)

Relaxes equal covariance assumption

Each group k has own covariance $\boldsymbol{\Sigma}_k$

When to use QDA:

- Groups have different variability patterns
- Sufficient sample size
- Linear boundaries inadequate

Result: Quadratic (curved) decision boundaries

QDA: Discriminant Scores

Score for group k:

$$\delta_k(\boldsymbol{x}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) + \log(\pi_k)$$

Quadratic in $\boldsymbol{x}$ leads to curved boundaries

LDA vs QDA Trade-offs

Criterion	**LDA**	**QDA**
Parameters	Fewer	More
Sample size need	Smaller	Larger
Decision boundaries	Linear	Curved
Interpretability	Simpler	Complex
Overfitting risk	Lower	Higher

Rule of thumb: Start with LDA, move to QDA if needed

Analysis Workflow

Step 1: Data Preparation

- Feature selection (avoid multicollinearity)
- Standardization (equal scales)
- Stratified train-test split

Step 2: Assumption Checking

- Multivariate normality (Q-Q plots, tests)
- Equal covariances (Box's M test)
- Multicollinearity (VIF)

Step 3: Model Fitting

- Fit LDA and/or QDA
- Extract discriminant functions

Analysis Workflow (cont.)

Step 4: Interpretation

- Examine discriminant coefficients
- Identify key separating variables
- Calculate group means on functions

Step 5: Validation

- Test set accuracy
- Confusion matrix
- Cross-validation
- ROC curves and AUC
- Visualize decision boundaries

Python Implementation

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis

# Prepare data
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, stratify=y, random_state=42
)

# Standardize
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Fit LDA
```

```python
lda = LinearDiscriminantAnalysis()
lda.fit(X_train_scaled, y_train)

# Predict
y_pred = lda.predict(X_test_scaled)
accuracy = lda.score(X_test_scaled, y_test)
```

Marketing Example: Setup

Business Problem:

E-commerce with 1,200 customers, 3 segments for targeting

Three Segments:

- High-Value (30%): Premium customers
- Loyal (40%): Regular customers
- Occasional (30%): Infrequent buyers

Eight Behavioral Metrics:

Purchase frequency, order value, browsing time, cart abandonment, email open rate, loyalty points, support tickets, social engagement

Marketing Example: Results

Discriminant Functions:

- **LD1 (95.8%):** Overall customer value
  - Drivers: frequency, loyalty points, order value
  - Separates High-Value from Occasional
- **LD2 (4.2%):** Order size patterns
  - Drivers: order value, browsing time

Performance:

- LDA: 99.9% accuracy
- QDA: 100.0% accuracy

Recommendation: Use LDA (simpler, equally effective)

Business Insights

High-Value: High frequency, strong engagement, premium retention strategy

Loyal: Moderate metrics, upselling and cross-selling focus

Occasional: Low frequency, high abandonment, re-engagement campaigns

Applications:

- Auto-classify new customers (2-3 months)
- Monitor segment migration
- Optimize marketing ROI
- Personalize campaigns

Advanced Topics

Variable Selection:

- Stepwise methods (forward/backward)
- Regularized DA (RDA)
- Penalized LDA

Imbalanced Classes:

- Adjust prior probabilities
- Oversampling (SMOTE)
- Undersampling

Diagnostics:

- Wilks' Lambda
- Canonical correlation

Comparison with Other Methods

Method	Best For
Logistic Regression	Binary outcomes, no normality assumption
SVM	Non-linear boundaries, no assumptions
Random Forest	Non-linear, robust to outliers
Discriminant Analysis	Interpretability, understanding differences

Common Pitfalls

Mistakes to Avoid:

- Ignoring assumptions (normality, equal covariances)
- Not checking for outliers
- Overfitting (too many predictors)
- Evaluating only on training data
- Ignoring class imbalance
- Using correlated predictors

Best Practices

Data Quality:

- Handle missing data
- Screen for outliers
- Verify data integrity

Model Selection:

- Start with LDA baseline
- Use cross-validation
- Report multiple metrics

Validation:

- Independent test data
- Monitor over time
- Update as needed

Key Takeaways

Core Value:

- Not just prediction, but **understanding** group differences
- Interpretable discriminant functions
- Probabilistic classification confidence

When to Use:

- Labeled training data
- Need interpretability
- Moderate dimensionality
- Approximate multivariate normality

Decision: LDA vs QDA

Start simple (LDA), add complexity (QDA) only if justified

Hands-On Learning

Interactive Notebook:

`ch5_guiding_example/marketing_discriminant_analysis.ipynb`

Complete workflow:

1. Data generation (reproducible)
2. Exploratory analysis
3. LDA implementation
4. QDA comparison
5. Decision boundaries
6. Performance evaluation

Experiment with different splits, features, priors!

Questions?

“The goal is to turn data into information,

and information into insight.“

- *Carly Fiorina*

MA2003B - Multivariate Methods

Dr. Juliho Castillo

Tecnologico de Monterrey