

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

Premium customers

High spending

Max engagement

Loyal

Regular customers

Moderate spending

Consistent activity

Occasional

Infrequent buyers

Low engagement

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:

- $\mathbf{x} = (x_1, \dots, x_p)^\top$ 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
- $f_k(\mathbf{x})$ probability density for group k

Bayes Theorem Foundation

Goal: Classify observation with features \mathbf{x} into one of g groups

Bayes Theorem gives posterior probability:

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

where:

- π_k = prior probability of group k
- $f_k(\mathbf{x})$ = probability density of \mathbf{x} in group k

Bayes Optimal Classification:

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

Denominator is same for all groups, so we can ignore it

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 ($k = 0$):

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

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

Default Group ($k = 1$):

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

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: $f_0(x) \times \pi_0 = 0.0008 \times 0.95 = 0.00076$
- Default: $f_1(x) \times \pi_1 = 0.0030 \times 0.05 = 0.00015$

Step 2: Calculate denominator (sum of numerators)

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

Step 3: Calculate posterior probabilities

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

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

From Bayes to Discriminant Analysis

The Challenge:

We need to specify $f_k(\mathbf{x})$ for each group

The Assumption:

Assume each group follows **multivariate normal distribution**:

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right)$$

Key Parameters:

- $\boldsymbol{\mu}_k$ = mean vector for group k
- Σ_k = covariance matrix for group k

Simplifying the Math

Recall: We want to maximize $f_k(\mathbf{x})\pi_k$

Trick: Maximize $\log(f_k(\mathbf{x})\pi_k)$ instead (same result, easier math)

Taking the logarithm:

$$\begin{aligned}\log(f_k(\mathbf{x})\pi_k) &= -\frac{p}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma_k| \\ &\quad - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) + \log(\pi_k)\end{aligned}$$

Drop constant terms (same for all groups), define discriminant score $\delta_k(\mathbf{x})$

Two Scenarios: LDA vs QDA

Scenario 1: Equal Covariances (LDA assumption)

If $\Sigma_1 = \Sigma_2 = \dots = \Sigma_g = \Sigma$

Then $\log|\Sigma_k|$ is constant across groups

The quadratic term $(\mathbf{x} - \boldsymbol{\mu}_k)^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)$ expands to terms linear in \mathbf{x}

Result: **Linear discriminant function**

Scenario 2: Different Covariances (QDA assumption)

Each group has Σ_k

Keep all terms including $\log|\Sigma_k|$

Result: **Quadratic discriminant function**

Summary: Bayes to LDA/QDA

The Complete Connection

Step 1: Bayes optimal rule requires maximizing $f_k(\mathbf{x})\pi_k$

Step 2: Assume multivariate normal: $f_k(\mathbf{x}) \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

Step 3: Take logarithm for computational convenience

Step 4: Simplify based on covariance assumption:

- **Equal covariances** \rightarrow LDA (linear boundaries)
- **Different covariances** \rightarrow QDA (quadratic boundaries)

Both methods are **Bayesian classifiers** under normality assumption

Linear Discriminant Analysis (LDA)

Two Critical Assumptions:

1. Multivariate Normality

Each group follows multivariate normal distribution

2. Equal Covariances

$$\Sigma_1 = \Sigma_2 = \dots = \Sigma_g = \Sigma$$

Result: **Linear** decision boundaries

LDA: Deriving the Discriminant Scores

Start with log-likelihood, assume $\Sigma_k = \Sigma$ for all k :

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

Expand the quadratic term:

$$(\mathbf{x} - \boldsymbol{\mu}_k)^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) = \mathbf{x}^\top \Sigma^{-1} \mathbf{x} - 2\mathbf{x}^\top \Sigma^{-1} \boldsymbol{\mu}_k + \boldsymbol{\mu}_k^\top \Sigma^{-1} \boldsymbol{\mu}_k$$

Drop terms constant across groups:

Drop $-\frac{1}{2}\log|\Sigma|$ and $\mathbf{x}^\top \Sigma^{-1} \mathbf{x}$

Define LDA discriminant score:

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

This is **linear** in \boldsymbol{x}

Fisher's Approach

Alternative (equivalent) formulation:

Maximize ratio of between-group to within-group variance

For two groups:

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

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

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

Quadratic Discriminant Analysis (QDA)

Relaxes equal covariance assumption

Each group k has own covariance Σ_k

When to use QDA:

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

Result: Quadratic (curved) decision boundaries

QDA: Deriving the Discriminant Scores

Now allow different Σ_k for each group:

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

Key difference from LDA:

Cannot drop $\log|\Sigma_k|$ (varies by group)

Cannot drop $\mathbf{x}^\top \Sigma_k^{-1} \mathbf{x}$ (different Σ_k for each group)

Define QDA discriminant score:

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

This is **quadratic** in x , producing curved decision 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
```

```
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