Discriminant Analysis

Companion Notes for Chapter 5

MA2003B - Multivariate Methods in Data Science

Dr. Juliho Castillo · Tecnologico de Monterrey

Contents

Hands-On Learning Resource

This document is accompanied by an interactive Jupyter notebook that demonstrates all concepts with executable Python code and visualizations:

`ch5_guiding_example/marketing_discriminant_analysis.ipynb`

The notebook analyzes a customer segmentation case study with 1,200 e-commerce customers, comparing Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) for marketing campaign optimization.

Each major concept in these notes references specific modules in the notebook where you can see the implementation and results.

1. Introduction: The Classification Problem

1.1. Why Discriminant Analysis?

Imagine you are a marketing analyst at an e-commerce company with thousands of customers. You want to answer a fundamental question: **Given a customer's behavior, which segment do they belong to?** Are they a high-value customer worth premium retention efforts? A loyal regular who might respond to upselling? Or an occasional buyer who needs re-engagement?

This is a **classification problem** - we have several known groups and want to assign new observations to the correct group based on measured characteristics. Discriminant Analysis provides a principled statistical approach to this problem.

1.2. Real-World Applications

Discriminant analysis finds extensive application across diverse domains where classification decisions require both accuracy and interpretability. In business and marketing contexts, the technique enables customer segmentation for targeted campaigns by identifying behavioral patterns that distinguish high-value customers from occasional buyers. Financial institutions employ discriminant analysis for credit risk assessment, where the method helps determine whether to approve or reject loan applications based on applicant characteristics and historical default patterns. Additionally, churn prediction models leverage discriminant functions to identify customers at risk of discontinuing service, allowing proactive retention efforts.

The healthcare and medical sectors utilize discriminant analysis for diagnostic purposes, where the method combines patient symptoms, laboratory test results, and medical history to classify individuals into disease categories. Treatment response prediction represents another critical application, as clinicians can use discriminant models to identify which patients are likely to benefit from specific therapeutic interventions. Medical imaging classification similarly benefits from discriminant techniques, particularly when distinguishing between normal and pathological tissue patterns in radiological studies.

Manufacturing industries apply discriminant analysis extensively in quality control processes, where products must be classified as acceptable, borderline, or defective based on multiple measurement criteria. Defect type classification helps identify the root causes of manufacturing problems by associating specific defects with process parameters. Process monitoring and fault detection systems employ discriminant functions to distinguish normal operating conditions from various failure modes, enabling timely intervention before quality degradation occurs.

In sports and performance contexts, discriminant analysis assists in athlete classification for training programs by identifying physiological and performance characteristics that predict success in different athletic domains. Talent identification systems use discriminant models to evaluate prospective athletes based on multiple attributes, helping coaches allocate resources effectively. Performance level assessment employs these techniques to classify athletes into skill categories, informing coaching strategies and competitive placement decisions.

1.3. The Core Idea

Discriminant Analysis finds **discriminant functions** — linear (or quadratic) combinations of your predictor variables that best separate your groups. Think of it as finding the "best viewing angle" to distinguish between groups in multidimensional space.

Key Question: Which combination of behavioral metrics (purchase frequency, order value, engagement, etc.) best distinguishes high-value customers from loyal customers from occasional customers?

2. Mathematical Foundations

2.1. The Setup

The discriminant analysis framework requires formalization of the classification problem structure. The analysis involves g distinct groups or populations into which observations must be classified, with p predictor variables measured on each observation. A training dataset provides examples of observations with known group memberships, enabling the estimation of discriminant functions. The ultimate goal consists of developing a classification rule that assigns new observations to one of the g groups based on their predictor values.

The mathematical notation employed throughout discriminant analysis includes several key quantities. The vector $\boldsymbol{x} = \left(x_1, x_2, \ldots, x_p\right)^\top$ represents the predictor variables for a given observation, where the superscript indicates vector transposition. The prior probability π_k denotes the probability that a randomly selected observation belongs to group k, which may reflect either the proportion of group k in the population or be specified based on domain considerations. The mean vector $\boldsymbol{\mu}_k$ characterizes the central tendency of group k in the p-dimensional predictor space, while the covariance matrix $\boldsymbol{\Sigma}_k$ describes the variability and correlation structure within group k. Finally, the probability density function $f_k(\boldsymbol{x})$ specifies the likelihood of observing predictor values $\boldsymbol{x}$ given membership in group k, with the multivariate normal distribution serving as the standard assumption in classical discriminant analysis.

2.2. Classification Rules: Bayes Theorem

The foundation of discriminant analysis is Bayes theorem, which provides the optimal classification framework when distributional assumptions hold. Understanding this theorem in depth clarifies how discriminant analysis achieves its classification decisions and why these decisions are optimal under certain conditions.

2.2.1. Understanding the Notation

Before examining the theorem itself, we must clarify the mathematical notation employed throughout discriminant analysis:

Group Membership Variable G: The random variable G represents the group or class to which an observation belongs. This categorical variable can take values $1, 2, \ldots, g$, where g denotes the total number of groups in the classification problem. For example, in a credit risk application, G might take values from the set {0, 1}, where 0 indicates "no default" and 1 indicates "default". In the marketing segmentation example, G takes values from {1, 2, 3}, corresponding to High-Value, Loyal, and Occasional customer segments respectively.

Predictor Vector x: The vector $\boldsymbol{x} = \left(x_1, x_2, ..., x_p\right)^\top$ contains the values of p predictor variables measured on a specific observation. In the marketing context, this might be $\boldsymbol{x} = (\text{purchase frequency}, \text{order value}, \text{browsing time}, ...)^\top$. Each component x_j represents a measured characteristic that potentially helps discriminate between groups.

Prior Probability pi_k: The quantity π_k denotes the prior probability that a randomly selected observation belongs to group k, calculated before observing the predictor values. This probability reflects either population frequencies or domain knowledge about group prevalence. The prior probabilities must satisfy $\pi_k > 0$ for all groups and $\sum_{k=1}^{g} \pi_k = 1$.

Class-Conditional Density f_k(x): The function $f_k(\boldsymbol{x})$ represents the probability density of observing predictor values $\boldsymbol{x}$ given that the observation belongs to group k. In discriminant analysis, we typically assume $f_k(\boldsymbol{x})$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$. This density quantifies how typical or likely the observed predictor values are for members of group k.

2.2.2. The Bayes Theorem Formula

Given observation $\boldsymbol{x}$, the posterior probability of belonging to group k is:

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

This equation decomposes into several interpretable components:

Left Side - Posterior Probability: The quantity $P(G = k \mid \boldsymbol{x})$ represents the probability that an observation with observed predictor values $\boldsymbol{x}$ belongs to group k. This is precisely what we want to calculate for classification purposes. The vertical bar notation | reads as "given" or "conditional on", so we interpret this as "the probability of group membership k given that we have observed predictor values $\boldsymbol{x}$".

Numerator Components:

- $f_k(\boldsymbol{x})$: The likelihood of observing these predictor values if the observation belongs to group k
- π_k: The prior probability of group k before observing any predictor values
- Product $f_k(\boldsymbol{x})\pi_k$: The joint probability of both belonging to group k and having predictor values $\boldsymbol{x}$

Denominator: The sum $\sum_{j=1}^{g} f_j(\boldsymbol{x})\pi_j$ aggregates the joint probabilities across all g groups. This normalization constant ensures that the posterior probabilities sum to unity across all possible group assignments: $\sum_{k=1}^{g} P(G = k \mid \boldsymbol{x}) = 1$.

2.2.3. Intuitive Interpretation

Bayes theorem combines two distinct sources of information to produce classification decisions:

Prior Knowledge: The prior probabilities π_k encode what we know about group frequencies before examining any specific observation. If historical data shows that 5 percent of loan applicants default, we set $\pi_1 = 0.05$ for the default group. This prior information prevents the classifier from ignoring base rates and making unrealistic predictions.

Evidence from Data: The class-conditional density $f_k(\boldsymbol{x})$ quantifies how typical the observed predictor values are for group k. If a loan applicant has income equal to 50,000 dollars, debt ratio equal to 0.4, and credit score equal to 650, the density $f_1(\boldsymbol{x})$ tells us how likely these values are among customers who ultimately default, while $f_0(\boldsymbol{x})$ tells us how likely they are among customers who do not default.

The posterior probability is proportional to the product of these two components: "how likely this data is for group k" multiplied by "how common group k is". Groups that are both common (high π_k) and compatible with the observed data (high $f_k(\boldsymbol{x})$) receive high posterior probabilities.

2.2.4. Detailed Example: Credit Risk Assessment

Consider a credit risk classification problem with the following characteristics:

Groups: $k = 0$ (no default), $k = 1$ (default)

Prior Probabilities: Based on historical data, $\pi_0 = 0.95$ (95 percent of customers do not default) and $\pi_1 = 0.05$ (5 percent default)

Observation: A loan applicant with $\boldsymbol{x} = [\text{income} = 50000, \text{debt ratio} = 0.4, \text{credit score} = 650]^\top$

Class-Conditional Densities: Suppose our fitted discriminant analysis model produces:

- $f_0(\boldsymbol{x}) = 0.0008$ (this financial profile is somewhat typical for non-defaulters)
- $f_1(\boldsymbol{x}) = 0.0030$ (this financial profile is more typical for defaulters)

The posterior probabilities calculate as:

Numerator for group 0: $f_0(\boldsymbol{x})\pi_0 = 0.0008 \times 0.95 = 0.00076$

Numerator for group 1: $f_1(\boldsymbol{x})\pi_1 = 0.0030 \times 0.05 = 0.00015$

Denominator: $\sum_{j=0}^{1} f_j(\boldsymbol{x})\pi_j = 0.00076 + 0.00015 = 0.00091$

Posterior probabilities:

- $P(G = 0 \mid \boldsymbol{x}) = \frac{0.00076}{0.00091} \approx 0.835$ (83.5 percent probability of no default)
- $P(G = 1 \mid \boldsymbol{x}) = \frac{0.00015}{0.00091} \approx 0.165$ (16.5 percent probability of default)

Despite the financial profile being more typical of defaulters (density ratio $\frac{f_1}{f_0} = 3.75$), the strong prior probability favoring non-default prevents the classifier from predicting default. The model balances the evidence from the data against the base rate information.

2.2.5. The Bayes Classification Rule

Bayes Classification Rule: Assign observation $\boldsymbol{x}$ to the group k^* that maximizes the posterior probability:

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

The second equality follows because the denominator $\sum_{j=1}^{g} f_j(\boldsymbol{x})\pi_j$ is identical across all groups and therefore does not affect which group achieves the maximum. Consequently, we need only compare the numerators $f_k(\boldsymbol{x})\pi_k$ across groups.

In the credit risk example above, we would classify the applicant as non-default (group 0) because $P(G = 0 \mid \boldsymbol{x}) = 0.835 > P(G = 1 \mid \boldsymbol{x}) = 0.165$.

2.2.6. Optimality of the Bayes Rule

The Bayes classification rule is optimal in a precise mathematical sense: it minimizes the total probability of misclassification. This result, known as the Bayes risk minimization theorem, states that among all possible classification rules, assigning observations to the group with the highest posterior probability achieves the lowest possible error rate.

This optimality holds under two critical conditions:

Condition 1 - Correct Prior Probabilities: The specified prior probabilities π_k accurately reflect the true prevalence of groups in the population or appropriately weight the relative importance of different groups.

Condition 2 - Correct Density Specification: The assumed class-conditional densities $f_k(\boldsymbol{x})$ correctly specify the distribution of predictor values within each group. In discriminant analysis, we assume multivariate normal distributions, so optimality requires that this assumption holds or approximately holds.

When these conditions are violated, the Bayes rule using the misspecified priors or densities may perform suboptimally. Nevertheless, the framework remains valuable because:

Robustness: Discriminant analysis often performs well even when the normality assumption is violated moderately, particularly with large sample sizes.

Flexibility in Priors: We can adjust prior probabilities to reflect business costs or domain knowledge, even if these differ from sample proportions.

Interpretability: The probabilistic framework provides posterior probabilities that quantify classification uncertainty, enabling risk-based decision making beyond simple classification.

2.2.7. Connection to Discriminant Functions

The discriminant analysis methods (LDA and QDA) implement the Bayes classification rule by making specific assumptions about the class-conditional densities $f_k(\boldsymbol{x})$. Both methods assume multivariate normal distributions:

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

Taking logarithms and simplifying yields discriminant functions, which are equivalent to comparing posterior probabilities:

LDA: Assuming $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ for all groups yields linear discriminant functions

QDA: Allowing group-specific $\boldsymbol{\Sigma}_k$ yields quadratic discriminant functions

Thus, discriminant analysis provides a computationally efficient implementation of the optimal Bayes rule under normality assumptions, transforming the classification problem into comparison of discriminant scores rather than explicit probability calculations.

2.3. Linear Discriminant Analysis (LDA)

2.3.1. Assumptions

LDA makes two critical assumptions:

1. **Multivariate Normality**: Each group follows a multivariate normal distribution
2. **Equal Covariances**: All groups share the same covariance matrix: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$

Under these assumptions, the discriminant functions become **linear** in $\boldsymbol{x}$.

2.3.2. Fisher Linear Discriminant

An alternative (but equivalent) approach by R.A. Fisher: Find linear combinations of variables that maximize the ratio of between-group variance to within-group variance.

For two groups, find weights $\boldsymbol{a}$ that maximize:

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

where $y = \boldsymbol{a}^\top \boldsymbol{x}$ is the discriminant score.

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

This gives us the **linear discriminant function**.

2.3.3. Discriminant Scores

For observation $\boldsymbol{x}$, the discriminant score for group k is:

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

Classification Rule: Assign $\boldsymbol{x}$ to the group with the largest discriminant score $\delta_k(\boldsymbol{x})$.

2.3.4. Geometric Interpretation

The geometric perspective on linear discriminant analysis provides valuable intuition about the classification mechanism. Each discriminant function defines a hyperplane in the p-dimensional predictor space, with these hyperplanes serving as decision boundaries that partition the space into regions corresponding to different group assignments. The linearity of these decision boundaries, which gives the method its name, implies that the boundaries consist of straight lines in two dimensions, flat planes in three dimensions, and hyperplanes in higher dimensions. When prior probabilities are equal across groups, the decision boundaries take a particularly elegant form as perpendicular bisectors of the lines connecting group centroids, reflecting the intuitive principle that observations should be assigned to the nearest group center when groups are equally likely a priori.

2.4. Quadratic Discriminant Analysis (QDA)

2.4.1. When LDA Is Not Enough

QDA relaxes the equal covariance assumption. Each group k has its own covariance matrix $\boldsymbol{\Sigma}_k$.

When to use QDA:

- Groups have genuinely different variability patterns
- You have sufficient sample size (more parameters to estimate)
- Linear boundaries do not fit the data well

2.4.2. Quadratic Discriminant Functions

The discriminant score becomes:

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

This is **quadratic** in $\boldsymbol{x}$, leading to curved (quadratic) decision boundaries.

2.4.3. Trade-offs: LDA vs QDA

The choice between Linear Discriminant Analysis and Quadratic Discriminant Analysis involves balancing model complexity against estimation precision. LDA offers several compelling advantages that make it the preferred starting point for most applications. The method requires estimation of only $\frac{p(p+1)}{2}$ covariance parameters for the pooled covariance matrix, compared to $g \cdot \frac{p(p+1)}{2}$ parameters required by QDA when each of the g groups maintains its own covariance structure. This parameter efficiency translates to greater stability when working with smaller sample sizes, as fewer quantities require estimation from limited data. The reduced complexity also makes LDA less prone to overfitting, particularly when the number of predictors approaches the sample size. Additionally, the linear structure of LDA discriminant functions facilitates interpretation, as the coefficients directly indicate how variables combine to separate groups.

Conversely, QDA provides advantages when the assumption of equal covariances proves untenable. The method's flexibility allows it to fit complex, curved decision boundaries that better capture the structure of data when groups genuinely exhibit different variability patterns. This flexibility translates to improved classification accuracy when the equal covariance assumption is violated substantially. Furthermore, QDA does not constrain groups to share variance structures, making it more appropriate when domain knowledge suggests heterogeneous within-group variability.

A practical rule of thumb guides the selection between these methods. The recommended approach begins with LDA as the baseline model. Three conditions suggest moving to QDA warrants consideration. First, the sample size should be substantially larger than the number of predictors, as QDA's additional parameters require more data for reliable estimation. Second, exploratory analysis should reveal clearly different spread patterns across groups, suggesting that the equal covariance assumption may not hold. Third, LDA should demonstrate unsatisfactory performance, with classification accuracy falling short of acceptable thresholds or exhibiting systematic patterns of misclassification that curved boundaries might address.

3. Practical Implementation

Interactive Tutorial: The notebook `ch5_guiding_example/marketing_discriminant_analysis.ipynb` demonstrates each step of this workflow with executable Python code, detailed explanations, and visualizations.

3.1. The Analysis Workflow

3.1.1. Step 1: Data Preparation

The data preparation phase establishes the foundation for reliable discriminant analysis, with Module 2 of the accompanying notebook providing complete implementation details. Feature selection represents the initial critical decision, requiring analysts to identify predictors that effectively discriminate between groups while avoiding redundancy. The selection process should eliminate highly correlated predictors to mitigate multicollinearity issues that can destabilize coefficient estimates and inflate standard errors. Domain knowledge plays an essential role in this process, as subject matter expertise often reveals which variables possess theoretical relevance for group separation beyond what correlation analysis alone might suggest.

Standardization becomes necessary when variables exist on disparate scales, as the raw magnitude differences can cause variables with larger numerical ranges to dominate the discriminant functions inappropriately. For instance, mixing monetary variables measured in dollars (ranging from 0 to 200) with rate variables expressed as proportions (ranging from 0 to 1) without standardization would give undue weight to the monetary variables. The notebook implementation employs the `StandardScaler` transformation to convert all features to have mean zero and standard deviation one, ensuring that each variable contributes to the analysis based on its discriminatory power rather than its measurement scale.

The train-test split procedure requires careful attention to maintain the integrity of model evaluation. Stratified sampling preserves the proportional representation of each group in both the training and testing subsets, preventing bias that could arise if certain groups were over-represented in one subset. The typical split allocates seventy percent of observations to the training set for parameter estimation and thirty percent to the testing set for unbiased performance evaluation. The implementation utilizes the train_test_split function with the stratify=y parameter, which ensures all groups appear in both subsets according to their original proportions.

3.1.2. Step 2: Assumption Checking

Verification of underlying assumptions ensures the validity of discriminant analysis results and guides methodological choices. The multivariate normality assumption, which posits that observations within each group follow a multivariate normal distribution, can be assessed through several diagnostic approaches. Q-Q plots constructed for each variable within each group provide visual evidence of univariate normality, while multivariate tests such as the Mardia test and Henze-Zirkler test offer formal statistical evaluation of multivariate normality. Fortunately, discriminant analysis demonstrates robustness to moderate violations of normality when sample sizes are large, as the Central Limit Theorem ensures that parameter estimates remain approximately normally distributed even when the underlying data deviate somewhat from normality.

The equal covariance assumption, which is fundamental to LDA, merits particular scrutiny. Box's M test provides a formal statistical test of covariance homogeneity across groups, though practitioners should note that this test exhibits considerable sensitivity and frequently rejects the null hypothesis even when differences are practically negligible. Visual inspection of covariance matrices computed separately for each group often provides more practical insight, allowing analysts to judge whether observed differences justify the additional complexity of QDA. When the equal covariance assumption is violated substantially, the analyst faces a choice between transitioning to QDA, which accommodates group-specific covariances, or applying variance-stabilizing transformations to the data.

Multicollinearity assessment protects against numerical instability and interpretation difficulties arising from highly correlated predictors. Examination of the correlation matrix among predictors reveals pairwise relationships, while Variance Inflation Factors quantify the degree to which each predictor's variance is inflated due to linear relationships with other predictors. When multicollinearity proves problematic, the analyst should remove redundant predictors, retaining those with greater theoretical relevance or stronger discriminatory power.

3.1.3. Step 3: Model Fitting

Notebook Reference: Module 3 (LDA) and Module 6 (QDA) demonstrate model fitting with scikit-learn.

Linear Discriminant Analysis estimation in Python employs the scikit-learn library's LinearDiscriminantAnalysis class. The estimation procedure fits the model to training data, producing several key outputs that inform interpretation and prediction. The lda.scalings_ attribute contains the discriminant function coefficients, which indicate how the original variables combine to form discriminant scores. The lda.means_ attribute stores the group centroids, representing the mean position of each group in the predictor space. While the pooled covariance matrix remains internal to the implementation, the lda.priors_ attribute reveals the prior probabilities assigned to each group, and the lda.explained_variance_ratio_ attribute quantifies how much between-group variance each discriminant function captures.

```python
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis

lda = LinearDiscriminantAnalysis()
lda.fit(X_train, y_train)
```

Quadratic Discriminant Analysis follows a parallel implementation pattern through scikit-learn's QuadraticDiscriminantAnalysis class. The model fitting process estimates group-specific parameters, with the qda.means_ attribute providing group centroids analogous to those in LDA. The group-specific covariance matrices, which distinguish QDA from LDA, remain internal to the implementation but govern the quadratic decision boundaries. The `qda.priors_` attribute contains prior probabilities, while the `predict_proba()` method generates posterior probabilities that quantify classification confidence for each observation and group combination.

```python
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis

qda = QuadraticDiscriminantAnalysis()
qda.fit(X_train, y_train)
```

3.1.4. Step 4: Interpretation

Notebook Reference: Module 4 provides detailed coefficient interpretation with group means analysis.

Discriminant Functions

For LDA with g groups, we get up to $\min(g - 1, p)$ discriminant functions. The maximum number of discriminant functions is determined by the formula:

$$m = \min(g - 1, p)$$

where g is the number of groups and p is the number of predictors. This limitation arises because:

- With g groups, there are only $g - 1$ independent contrasts between group means
- With p predictors, the discriminant space has at most p dimensions

For example, with 3 customer segments and 8 predictors, we obtain $\min(3 - 1, 8) = 2$ discriminant functions.

First discriminant function (LD1): Explains most between-group variation **Second discriminant function** (LD2): Second-most variation (orthogonal to LD1)

Marketing Example Interpretation (from the notebook):

- **LD1** (95.8% of variance) separates High-Value from Occasional customers
  - Strong negative coefficients: purchase frequency, loyalty points
  - Represents overall customer engagement and activity level
- **LD2** (4.2% of variance) captures remaining separation
  - Strong positive coefficient: average order value
  - Strong negative coefficient: browsing time
  - Represents order size patterns versus browsing efficiency

The notebook's Module 4 displays both the discriminant coefficients (scalings) and the group means on standardized features, allowing interpretation of which behavioral patterns define each segment.

Understanding Discriminant Coefficients

Discriminant analysis produces two types of coefficients, each serving distinct interpretive purposes:

Unstandardized Coefficients: These represent the raw weights applied to predictor variables in their original measurement units. An unstandardized coefficient of 0.5 for income (measured in thousands of dollars) means that a 1,000 dollar increase in income increases the discriminant score by 0.5 units. These coefficients depend on the original measurement scales and cannot be directly compared across variables measured in different units.

Standardized Coefficients: These coefficients indicate the relative importance of each predictor independent of its original scale. Standardization converts all predictors to have mean zero and standard deviation one before computing discriminant functions. A standardized coefficient of 0.8 for income and 0.3 for age suggests that income contributes more strongly to group separation than age, after accounting for differences in their variability. Standardized coefficients facilitate comparison across predictors and are analogous to beta coefficients in standardized regression.

The standardized form proves particularly valuable when comparing the relative importance of predictors, while the unstandardized form is necessary when applying the discriminant function to new observations in their original measurement units.

Discriminant Loadings

Similar to factor analysis loadings - correlation between original variable and discriminant function. Discriminant loadings represent the simple correlation between each original predictor and the discriminant scores, providing an alternative measure of variable importance.

High absolute loading = variable is important for that discriminant function. Loadings above 0.30 in absolute value typically indicate meaningful contributions to the discriminant function.

The relationship between coefficients and loadings mirrors factor analysis: coefficients represent regression-like weights accounting for correlations among predictors, while loadings represent simple correlations ignoring other predictors. When predictors are highly correlated, coefficients and loadings can differ substantially, with loadings often providing more intuitive interpretation.

The notebook creates a DataFrame of coefficients to easily identify the most influential features for each discriminant function.

3.1.5. Step 5: Validation

Comprehensive model validation ensures that discriminant analysis results generalize beyond the training data and meet performance requirements. Modules 3, 6, 9, and 10 of the accompanying notebook provide detailed validation analyses that illustrate these principles. Classification accuracy assessment begins with the confusion matrix constructed from test set predictions, as presented in Module 9. This matrix reveals not only overall accuracy but also class-specific performance through precision, recall, and F1-scores compiled in the classification report. Practitioners must resist the temptation to focus solely on overall accuracy, as this aggregate metric can obscure poor performance on specific groups, particularly when class sizes differ substantially.

Cross-validation provides a more robust estimate of model performance by evaluating the model across multiple train-test partitions of the data. K-fold cross-validation, typically employing five or ten folds, assesses performance stability across different data splits and guards against the possibility that a single train-test split might yield unrepresentative results. The notebook implementation utilizes the `cross_val_score` function with five folds, generating a distribution of accuracy estimates that characterizes model performance more comprehensively than a single test set evaluation.

Visual validation techniques complement numerical performance metrics by revealing the geometric structure of group separation. Module 5 presents scatter plots of discriminant scores in the two-dimensional space defined by LD1 and LD2, allowing visual assessment of how well the discriminant functions separate groups. Module 8 provides decision boundary plots that project the classification regions onto two-dimensional feature subspaces, illustrating where the model assigns observations to each group. Module 10 generates ROC curves for each segment using the one-versus-rest approach, with the area under these curves quantifying the model's ability to distinguish each segment from all others. These visualizations verify that groups are well-separated and reveal any regions of potential classification ambiguity.

Advanced validation metrics provide additional insights into model performance characteristics. Module 10′s ROC curve analysis with AUC scores quantifies discrimination ability for each group independently, revealing whether certain groups prove easier to classify than others. Module 7 analyzes posterior probabilities to assess prediction confidence, identifying observations where the model assigns high probability to the predicted class versus those where probabilities spread across multiple classes. Module 9′s side-by-side confusion matrix comparison for LDA versus QDA facilitates direct assessment of whether QDA's additional complexity translates to meaningful performance improvements.

4. Applied Example: Marketing Segmentation

Complete
Implementation: See `ch5_guiding_example/marketing_discriminant_analysis.ipynb` for the full interactive analysis with visualizations and detailed interpretations.

4.1. Business Problem

An e-commerce company has 1,200 customers and wants to classify them into three segments for targeted marketing:

1. **High-Value** (30%): Premium customers, high spending and engagement
2. **Loyal** (40%): Regular customers, moderate spending, consistent
3. **Occasional** (30%): Infrequent buyers, need re-engagement

The dataset is synthetically generated using `fetch_marketing.py` to ensure reproducibility and known statistical properties for educational purposes.

4.2. Variables (p = 8)

The analysis employs eight behavioral metrics that characterize customer engagement patterns across multiple dimensions. Purchase frequency measures the average number of transactions a customer completes per month, providing insight into shopping regularity. Average order value, denominated in US dollars, captures the typical monetary amount spent per transaction, distinguishing high-spending customers from those making smaller purchases. Browsing time quantifies the minutes spent per website session, reflecting engagement level and product consideration depth. Cart abandonment rate, expressed as a proportion between zero and one, indicates the frequency with which customers initiate but fail to complete purchases. Email open rate, similarly scaled from zero to one, measures responsiveness to marketing communications. Loyalty points represent the accumulated rewards a customer has earned through the company's retention program. Support tickets count the average monthly customer service interactions, potentially indicating product satisfaction or purchase complexity. Social engagement tracks monthly interactions with the company's social media presence, including likes, shares, and comments.

The `MARKETING_DATA_DICTIONARY.md` document provides comprehensive variable descriptions, including detailed range specifications, measurement precision, and the multivariate normal data generation methodology employed to create the synthetic dataset.

4.3. Analysis Strategy

4.3.1. Why Both LDA and QDA?

LDA: Assumes all customer segments have similar variability patterns **QDA**: Allows different patterns (e.g., Occasional customers might be more variable)

We'll fit both and compare performance.

4.3.2. Feature Standardization

Since variables are on different scales (dollars, rates, counts), we standardize:

$$z_j = \frac{x_j - \overline{x}_j}{s_j}$$

This ensures no variable dominates due to scale.

4.4. Results

Note: The following results are from the Jupyter notebook analysis. Run `marketing_discriminant_analysis.ipynb` to reproduce these findings and generate visualizations.

4.4.1. Discriminant Functions

The notebook's Module 4 (Interpreting Discriminant Functions) reveals:

LD1 (95.8% of between-group variance):

- Separates High-Value from Occasional customers along the primary axis
- Key drivers: Purchase frequency (strong negative), loyalty points (strong negative), average order value (negative)
- Interpretation: Overall customer value and activity level

LD2 (4.2% of between-group variance):

- Distinguishes remaining group differences
- Key drivers: Average order value (strong positive), browsing time (negative)
- Interpretation: Order size versus browsing efficiency

Insight: Two independent dimensions describe customer segments:

1. Overall engagement and frequency (LD1 - dominant factor)
2. Purchase value patterns (LD2 - secondary factor)

4.4.2. Classification Performance

LDA Results (Module 3):

- Test accuracy: Nearly perfect classification
- Cross-validation: 99.9% (± 0.3%)
- Strong performance across all segments

QDA Results (Module 6):

- Test accuracy: Perfect classification
- Cross-validation: 100.0% (± 0.0%)
- Slightly outperforms LDA with more flexible boundaries

Interpretation: Both models achieve excellent performance. The synthetic data's well-separated structure allows for near-perfect classification, demonstrating the power of discriminant analysis when groups have distinct multivariate profiles.

4.4.3. Model Comparison and Visualization

The notebook includes comprehensive visualizations:

Module 5: Discriminant space scatter plot showing customer distribution in LD1-LD2 space with group centroids

Module 8: QDA decision boundaries visualization using purchase frequency and average order value

Module 9: Side-by-side confusion matrices comparing LDA and QDA performance

Module 10: ROC curves for each segment showing excellent discrimination (AUC near 1.0)

4.4.4. Model Selection Recommendation

Module 11 provides a comprehensive comparison and recommends **LDA** despite QDA's marginally better performance because:

- Similar accuracy (difference is negligible)
- Simpler model with fewer parameters
- More interpretable discriminant functions
- Lower overfitting risk
- Easier to explain to stakeholders

4.5. Business Insights

4.5.1. Segment Characteristics

Module 4′s group means analysis reveals distinct behavioral profiles that characterize each customer segment. High-Value customers exhibit positive standardized values across purchase frequency, order value, and browsing time, indicating sustained engagement with the company's offerings. This segment demonstrates exceptionally high email open rates and social engagement, suggesting strong brand affinity and responsiveness to marketing communications. Notably, these customers show low cart abandonment rates and minimal support ticket generation, reflecting purchase certainty and product satisfaction. The highest loyalty points accumulation among this group confirms their long-term value to the organization. The recommended strategy for High-Value customers emphasizes retention through premium services, personalized product recommendations, and exclusive benefits that reinforce their valued status.

Loyal customers present moderate purchase frequency and order values, positioning them between the extremes of occasional and high-value segments. This group maintains good email engagement and accumulates loyalty points steadily, though not at the accelerated pace observed among High-Value customers. Their behavioral metrics demonstrate balance across most dimensions, suggesting consistent but not exceptional engagement. The strategic approach for Loyal customers focuses on upselling opportunities that encourage larger purchases and cross-selling initiatives that broaden their product adoption. Enhancements to the loyalty program can incentivize increased spending and engagement, potentially facilitating migration toward High-Value status.

Occasional customers display negative standardized values across most engagement metrics, indicating sporadic and limited interaction with the company. This segment exhibits elevated cart abandonment rates and generates more support tickets per capita, suggesting hesitation in purchase decisions or difficulties in the buying process. Minimal loyalty points accumulation and social engagement further characterize this group's tenuous connection to the brand. The strategic imperative for Occasional customers involves re-engagement campaigns designed to increase purchase frequency, cart recovery mechanisms that address abandonment triggers, and educational content that builds product knowledge and purchase confidence.

4.5.2. Actionable Applications

The discriminant analysis framework enables several practical applications that translate statistical insights into business value. New customer classification becomes feasible once a customer accumulates sufficient behavioral data, typically requiring two to three months of transaction history. The model then automatically assigns the customer to a segment, enabling tailored marketing strategies from an early stage in the customer relationship. Module 7 demonstrates how posterior probabilities provide confidence scores for each classification, allowing the marketing team to gauge classification certainty and adjust campaign intensity accordingly.

Monitoring segment migration over time reveals customer lifecycle dynamics that inform retention strategies. Tracking discriminant scores longitudinally identifies Occasional customers who transition toward Loyal status, validating re-engagement efforts, as well as Loyal customers whose scores drift toward Occasional levels, signaling attrition risk. The discriminant scores serve as early warning indicators of segment transition, enabling proactive interventions before customers fully migrate to less valuable segments.

Marketing return on investment optimization leverages the classification framework to allocate resources efficiently across customer segments. Expensive retention campaigns targeting High-Value customers, where classification confidence is typically high, maximize the impact of premium marketing investments. Cost-effective email campaigns suit Loyal customers, whose moderate value justifies ongoing engagement without intensive personalization. Automated cart recovery mechanisms address Occasional customers efficiently, deploying technology-enabled interventions that require minimal manual effort.

Campaign personalization benefits from Module 7′s posterior probability analysis, which identifies customers with ambiguous segment membership reflected in probability distributions spread across multiple classes. These customers, exhibiting characteristics of multiple segments, may respond favorably to hybrid marketing strategies that combine elements designed for different segments, such as re-engagement messaging paired with loyalty program incentives.

4.6. Credit Risk Classification Example

To illustrate discriminant analysis in a financial context, consider a credit risk assessment problem where a bank classifies loan applicants as either “No Default” (low risk) or “Default” (high risk) based on financial characteristics.

4.6.1. Problem Setup

Groups: $g = 2$
- Group 0: No Default (historically 95% of applicants)
- Group 1: Default (historically 5% of applicants)

Predictors: $p = 3$
- x_1: Annual income (thousands of dollars)
- x_2: Debt-to-income ratio (proportion, 0 to 1)
- x_3: Credit score (300 to 850)

4.6.2. Discriminant Score Interpretation

After fitting LDA, suppose the discriminant function is:

$$\text{Score} = 0.002 \cdot \text{income} - 8.5 \cdot \text{debt ratio} + 0.015 \cdot \text{credit score} - 5.2$$

The cutoff threshold with equal prior probabilities would be zero. However, recognizing the imbalanced class distribution and the higher cost of default misclassification, the bank adjusts prior probabilities to reflect business considerations rather than sample proportions.

Example Classification:

Consider applicant A with income equal to 50,000 dollars, debt ratio equal to 0.40, and credit score equal to 650:

$$\text{Score}_A = 0.002(50) - 8.5(0.40) + 0.015(650) - 5.2$$

$$\text{Score}_A = 0.1 - 3.4 + 9.75 - 5.2 = 1.25$$

A positive discriminant score (1.25 greater than 0) indicates classification into Group 0 (No Default), suggesting this applicant presents low credit risk. The magnitude of the score reflects the strength of classification confidence.

Consider applicant B with income equal to 35,000 dollars, debt ratio equal to 0.65, and credit score equal to 580:

$$\text{Score}_B = 0.002(35) - 8.5(0.65) + 0.015(580) - 5.2$$

$$\text{Score}_B = 0.07 - 5.525 + 8.7 - 5.2 = -1.955$$

A negative discriminant score (-1.955 less than 0) indicates classification into Group 1 (Default), flagging this applicant as high credit risk. The bank would likely deny this loan application or offer modified terms with higher interest rates to compensate for elevated risk.

Coefficient Interpretation:

- **Income** (+0.002): Higher income reduces default risk, though the small coefficient reflects income's measurement in thousands
- **Debt Ratio** (-8.5): Strong negative coefficient indicates high debt loads substantially increase default risk
- **Credit Score** (+0.015): Higher credit scores reduce default risk, reflecting proven creditworthiness

The debt-to-income ratio emerges as the strongest predictor when coefficients are standardized, consistent with credit risk theory that emphasizes debt burden as a primary default driver.

Business Application:

The discriminant score threshold can be adjusted to reflect business priorities. Setting a higher threshold (e.g., requiring scores above 0.5 instead of 0) makes classification more conservative, approving fewer borderline applicants but reducing default rates. Conversely, lowering the threshold (e.g., −0.5) approves more applicants, increasing loan volume but accepting higher default risk. Banks optimize this threshold by analyzing the trade-off between foregone interest revenue from rejected non-defaulters versus losses from accepted defaulters.

5. Advanced Topics

5.1. Variable Selection

The inclusion of all available predictors may not optimize discriminant analysis performance, motivating variable selection procedures that identify parsimonious models. Stepwise discriminant analysis employs iterative algorithms to build or refine predictor sets. Forward selection begins with an empty model and sequentially adds variables that contribute most substantially to group separation, typically measured by F-to-enter statistics or Wilks' Lambda changes. Conversely, backward elimination starts with all candidate predictors and iteratively removes those contributing least to discrimination, guided by F-to-remove statistics. Both approaches utilize Wilks' Lambda or related F-statistics to quantify how much each variable enhances group separation. While computationally convenient, stepwise methods can be unstable when predictors are highly correlated and may capitalize on chance associations in finite samples.

Shrinkage methods offer modern alternatives that stabilize coefficient estimates through regularization. Regularized Discriminant Analysis (RDA) introduces a tuning parameter that shrinks the group-specific covariance matrices toward a common structure, effectively interpolating between LDA and QDA while reducing parameter variance. Penalized LDA extends this concept by applying penalties such as the L1 norm to discriminant coefficients, encouraging sparse solutions where many coefficients equal zero. This sparsity proves valuable in high-dimensional settings where interpretability benefits from identifying a small subset of influential predictors.

5.2. Handling Imbalanced Classes

Substantial imbalance in group sizes, such as ninety-five percent acceptable items versus five percent defective items, poses challenges for discriminant analysis. When sample proportions serve as prior probabilities, the classifier tends to favor the majority class, achieving high overall accuracy by predominantly predicting the larger group while performing poorly on the minority class. Two complementary strategies address this issue.

Adjusting prior probabilities away from sample proportions recalibrates the classification rule to account for the practical importance of different groups. Setting equal priors regardless of sample sizes gives each group equal a priori weight, preventing the majority class from dominating predictions solely due to its prevalence. Alternatively, priors can reflect business costs, assigning higher prior probability to groups where misclassification carries greater consequences. For instance, if failing to detect a defective item costs substantially more than falsely flagging an acceptable item, elevated prior probability for the defective class appropriately adjusts the decision threshold.

Sampling techniques modify the training data composition to mitigate imbalance effects. Oversampling the minority class through replication or synthetic data generation increases its representation in the training set, allowing the model to learn its characteristics more effectively. Undersampling the majority class reduces its dominance by randomly selecting a subset of observations for model training. The Synthetic Minority Over-sampling Technique (SMOTE) generates artificial minority class observations by interpolating between existing minority class instances in feature space, effectively expanding the minority class representation while introducing variation beyond simple replication.

5.3. Model Diagnostics

Several diagnostic statistics quantify the strength and significance of group separation in discriminant analysis. Understanding these diagnostics helps assess model quality and interpret the degree to which groups differ in the multivariate space.

5.3.1. Wilks' Lambda

Wilks' Lambda tests whether group means differ significantly across the predictor space, calculated as the ratio of the determinant of the within-group sum of squares matrix $\boldsymbol{W}$ to the determinant of the total sum of squares matrix $\boldsymbol{T}$:

$$\Lambda = \frac{|\boldsymbol{W}|}{|\boldsymbol{T}|}$$

Small values approaching zero indicate strong group separation, as the within-group variation becomes negligible relative to total variation. The test statistic ranges from 0 to 1, where:

- $\Lambda \approx 0$: Perfect group separation (all variation is between groups)
- $\Lambda \approx 1$: No group separation (all variation is within groups)

The associated F-statistic provides formal hypothesis testing of whether group means differ significantly. Wilks' Lambda is also used in stepwise discriminant analysis as a criterion for variable selection, where variables that minimize Lambda (maximize group separation) are preferred for inclusion in the model.

5.3.2. Hotelling's T-squared Test

For testing whether two groups have different centroids (mean vectors), Hotelling's T-squared test provides the multivariate generalization of the univariate t-test. The test statistic is:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{\text{pooled}}^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)$$

where n_1 and n_2 are the sample sizes for the two groups, $\overline{\boldsymbol{x}}_1$ and $\overline{\boldsymbol{x}}_2$ are the group mean vectors, and $\boldsymbol{S}_{\text{pooled}}$ is the pooled covariance matrix. This test determines whether the multivariate means differ significantly, providing a foundation for discriminant analysis by confirming that groups occupy distinct regions of the predictor space.

For more than two groups, multivariate analysis of variance (MANOVA) extends this concept, testing the null hypothesis that all group centroids are equal. MANOVA and discriminant analysis are closely related: MANOVA tests whether groups differ, while discriminant analysis describes how they differ and builds classification rules based on those differences.

5.3.3. Eigenvalues and Discriminant Functions

Each discriminant function has an associated eigenvalue that quantifies its discriminatory power. The eigenvalue represents the ratio of between-group variance to within-group variance along that discriminant function:

$$\lambda_i = \frac{\text{between-group SS for LD}_i}{\text{within-group SS for LD}_i}$$

Larger eigenvalues indicate stronger discrimination. A high eigenvalue means that particular discriminant function effectively separates groups, as the between-group variation dominates the within-group variation along that dimension.

The proportion of discriminatory power explained by each function is calculated as:

$$\text{Proportion explained by LD}_i = \frac{\lambda_i}{\sum_{j=1}^{m} \lambda_j}$$

where $m = \min(g - 1, p)$ is the total number of discriminant functions. Typically, the first discriminant function (LD1) accounts for the largest proportion of group separation, with subsequent functions capturing progressively less discrimination.

Example Interpretation: If LD1 has an eigenvalue of 12.5 and LD2 has an eigenvalue of 0.8, then LD1 explains $\frac{12.5}{12.5+0.8} = 94\%$ of the discriminatory power. This indicates that the first dimension captures nearly all meaningful group separation, and LD2 contributes minimally.

5.3.4. Canonical Correlation

Canonical correlation measures the strength of the relationship between discriminant functions and group membership. For each discriminant function, the canonical correlation represents the square root of the proportion of between-group variation to total variation:

$$R_{\text{can}} = \sqrt{\frac{\text{between-group SS}}{\text{total SS}}} = \sqrt{\frac{\lambda_i}{1 + \lambda_i}}$$

Values approaching one indicate excellent discrimination, signifying that the discriminant function effectively captures group differences. Canonical correlation ranges from 0 to 1, where:

- $R_{\text{can}} \approx 1$: Discriminant function perfectly separates groups
- $R_{\text{can}} \approx 0$: Discriminant function provides no group separation

Multiple discriminant functions yield multiple canonical correlations, with the first function typically exhibiting the highest value. Squaring the canonical correlation gives the proportion of variance in the discriminant scores explained by group membership, analogous to R-squared in regression.

5.3.5. Mahalanobis Distance

Mahalanobis distance quantifies the distance between an observation and a group centroid, accounting for the covariance structure of the data. For observation $\boldsymbol{x}$ and group k, the Mahalanobis distance is:

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

This metric plays multiple roles in discriminant analysis:

Classification: With equal prior probabilities, LDA assigns observations to the group with the smallest Mahalanobis distance to its centroid. This geometric interpretation clarifies why LDA works: it identifies the "nearest" group in multivariate space, where "nearest" accounts for variable correlations and scales.

Outlier Detection: Observations with large Mahalanobis distances from all group centroids may be multivariate outliers requiring investigation. Such observations lie far from typical group members in the multivariate space.

Variable Selection: Mahalanobis distance between group centroids can guide variable selection, as variables that increase the distance between groups enhance discrimination.

Unlike Euclidean distance, Mahalanobis distance accounts for variable correlations and unequal variances, making it more appropriate for multivariate data where predictors are not independent and identically scaled.

5.4. Comparison with Other Methods

Discriminant analysis occupies a particular niche among classification methods, with its suitability depending on data characteristics and analytical objectives. Logistic regression offers greater flexibility by avoiding the normality assumption, modeling the log-odds of class membership directly as a linear function of predictors. This approach works particularly well for binary outcomes and provides probability estimates through a natural probabilistic framework. However, extending logistic regression to multiple groups through multinomial models sacrifices some interpretability, as the method requires estimating separate coefficient sets for each group contrast rather than generating low-dimensional discriminant functions.

Support Vector Machines (SVM) provide powerful alternatives when decision boundaries exhibit complex, non-linear structure. The kernel trick enables SVMs to implicitly operate in high-dimensional feature spaces where linear separation becomes feasible, accommodating intricate decision boundaries without explicit feature engineering. SVMs impose no distributional assumptions, making them robust across diverse data types. The method's geometric focus on boundary optimization, however, yields less interpretable models than discriminant functions, and SVMs often excel in high-dimensional settings where discriminant analysis may struggle due to parameter estimation challenges.

Random Forests handle non-linear relationships naturally through recursive partitioning, building ensemble predictions from multiple decision trees. This approach demonstrates robustness to outliers and accommodates complex interactions without requiring explicit specification. Variable importance measures derived from random forests indicate which predictors contribute most to classification accuracy. The ensemble nature of random forests, however, creates black-box models that resist interpretation beyond variable importance rankings, limiting their utility when understanding group differences constitutes a primary analytical goal.

Discriminant analysis proves most appropriate when several conditions align. Moderate sample sizes and dimensionality suit the method's parameter estimation requirements, as extreme dimensionality or limited sample sizes introduce estimation instability. Interpretability requirements favor discriminant analysis, as the discriminant functions directly reveal which variable combinations separate groups.

Analytical objectives focused on understanding group differences rather than maximizing predictive accuracy align with the method's strengths. Finally, data conforming reasonably well to multivariate normality with similar within-group covariance structures satisfy the method's parametric assumptions, ensuring optimal performance.

6. Common Pitfalls and Best Practices

6.1. Common Mistakes

Several pitfalls frequently undermine discriminant analysis applications, warranting careful attention throughout the modeling process. Ignoring fundamental assumptions represents a primary source of error. Applying LDA when groups exhibit clearly different covariance structures violates the method's homogeneity assumption, potentially yielding biased decision boundaries that perform poorly on new data. Outliers exert disproportionate influence on discriminant functions due to their impact on estimated means and covariances, yet analysts sometimes proceed without outlier screening, particularly when sample sizes appear large. Mahalanobis distance calculations identify multivariate outliers that may not be apparent from univariate examinations.

The normality assumption requires that predictors follow multivariate normal distributions within each group. While discriminant analysis demonstrates robustness to moderate violations when sample sizes are large, severe departures from normality can compromise both classification accuracy and the optimality properties of the Bayes rule. Particularly problematic are categorical predictors, which inherently violate the continuity assumption underlying multivariate normality. Including categorical variables (such as gender coded as 0/1, or product category coded as 1/2/3) in discriminant analysis creates theoretical inconsistencies, as discrete distributions cannot be multivariate normal. When categorical predictors are essential for classification, analysts face several alternatives: convert categories to quantitative proxies when meaningful (e.g., replace product category with average price), use logistic regression which accommodates categorical predictors naturally, or employ non-parametric classification methods such as classification trees or k-nearest neighbors that impose no distributional assumptions.

Overfitting poses particular risks when the number of predictors approaches or exceeds the sample size relative to the number of groups. The discriminant analysis literature suggests requiring at least twenty observations per predictor per group to ensure stable parameter estimates. Violating this guideline risks producing discriminant functions that fit idiosyncrasies of the training sample rather than capturing genuine group differences, leading to poor generalization performance.

The practice of evaluating model performance on training data yields overly optimistic accuracy estimates that mislead analysts regarding true predictive performance. Training accuracy necessarily exceeds test accuracy because the model parameters are optimized to fit the training observations. Proper validation requires held-out test sets or cross-validation procedures that assess performance on observations not used for parameter estimation.

Class imbalance creates a subtle trap where models achieve high overall accuracy by predominantly predicting the majority class while failing to identify minority class members effectively. Overall accuracy masks this problem, as correct classification of the majority class dominates the aggregate metric. Examining per-class precision, recall, and F1-scores reveals whether the model performs adequately across all groups rather than succeeding primarily through majority class predictions.

Correlated predictors introduce multicollinearity that inflates coefficient standard errors and produces unstable discriminant functions where small data perturbations yield substantially different coefficients. The presence of highly redundant variables provides no additional information for discrimination while complicating interpretation and increasing estimation variance. Variable selection or principal component analysis can address severe multicollinearity by reducing predictor dimensionality.

6.2. Best Practices

Successful discriminant analysis implementation rests on several foundational practices that enhance reliability and interpretability. Data quality requires meticulous attention before model fitting begins. Missing data must be handled appropriately, either through principled imputation methods or case deletion when missingness is minimal and random. Outlier screening using Mahalanobis distance identifies observations that deviate substantially from their group's multivariate center, allowing assessment of whether these points represent data entry errors, measurement anomalies, or genuine extreme cases warranting special handling. Verification of data entry errors prevents spurious results arising from typographical mistakes or unit conversion errors.

Model selection proceeds most effectively through a disciplined progression from simplicity to complexity. Beginning with LDA establishes a baseline that subsequent, more complex models must meaningfully improve upon to justify their adoption. Cross-validation provides honest performance estimates that account for sampling variability, generating multiple train-test partitions to characterize performance distributions rather than relying on a single partition that might be unrepresentative. Multiple performance metrics paint a comprehensive picture of model behavior, with accuracy, precision, recall, F1-scores, and AUC each illuminating different aspects of classification performance.

Interpretation extends beyond merely reporting classification accuracy to explaining the substantive meaning of discriminant functions. Coefficient examination reveals which predictors drive group separation and how they combine to form discriminant scores. Visualizing decision boundaries, when feasible through dimensionality reduction, provides geometric intuition about classification regions. Translating statistical results into domain-specific insights ensures that stakeholders understand not just which observations are classified into which groups, but why the classification occurs and what it means for practical decision-making.

Validation practices maintain model reliability over time and across contexts. Testing on truly independent data, distinct from any data used during model development, provides the most conservative performance assessment. Monitoring performance over time in production deployments detects concept drift, where the relationships between predictors and groups evolve, degrading model accuracy. Updating models as patterns change ensures continued relevance, with retraining schedules determined by the rate of observed performance degradation.

7. Summary and Key Takeaways

Discriminant analysis provides a classification framework for assigning observations to predefined groups based on measured predictor variables. The method identifies linear combinations of predictors in LDA or quadratic combinations in QDA that maximally separate groups in the feature space. Under the assumptions of multivariate normality and known covariance structures, discriminant analysis implements the optimal Bayes classification rule, minimizing the total probability of misclassification when these conditions hold.

The method proves most appropriate when analysts possess labeled training data with known group memberships, seek interpretable insights into group differences rather than black-box predictions, work with moderate sample sizes and dimensionality where parameter estimation remains stable, and observe data that conform reasonably well to multivariate normality assumptions. These conditions position discriminant analysis as a valuable tool when understanding group structure constitutes a primary analytical objective alongside prediction.

Several key decisions shape discriminant analysis applications. The choice between LDA and QDA balances the parsimony of equal covariance assumptions against the flexibility of group-specific covariances, with LDA serving as the recommended starting point. Predictor selection combines domain knowledge about theoretically relevant variables with statistical criteria that identify redundant or non-discriminating features. Prior probability specification chooses between equal priors that weight groups uniformly, proportional priors that reflect sample or population frequencies, or cost-based priors that account for asymmetric misclassification consequences.

Interpretation distinguishes discriminant analysis from purely predictive methods. The discriminant functions reveal how groups differ by showing which variable combinations drive separation. Discriminant loadings or coefficients indicate which variables contribute most substantially to each function. Classification accuracy quantifies predictive performance, but the interpretive insights into group structure often provide greater analytical value than prediction accuracy alone.

Validation requirements apply universally to discriminant analysis applications. Hold-out test sets or cross-validation procedures provide honest performance estimates uncontaminated by training data optimization. Per-group performance metrics ensure that classification succeeds across all groups rather than primarily through majority class prediction. Monitoring performance over time in operational deployments detects degradation and informs model updating schedules.

The power of discriminant analysis ultimately lies not merely in classification accuracy, but in understanding what makes groups different. This interpretive capability transforms multivariate data into actionable insights that inform both scientific understanding and practical decision-making, distinguishing the method from alternative classification approaches that prioritize prediction over explanation.

8. Putting It Into Practice

To solidify your understanding of discriminant analysis, work through the complete marketing segmentation example:

Step 1: Generate the customer dataset

```
cd ch5_guiding_example
python fetch_marketing.py
```

Step 2: Open and run the Jupyter notebook

```
jupyter notebook marketing_discriminant_analysis.ipynb
```

Step 3: Execute each module sequentially, reading the explanations and examining the outputs

Step 4: Experiment with modifications:
- Try different train/test splits
- Exclude certain features to see impact on performance
- Adjust prior probabilities in the discriminant models
- Create additional visualizations

Step 5: Review the generated PNG files:
- `marketing_lda_scores.png`: Discriminant space visualization
- `marketing_qda_boundaries.png`: Decision boundaries in 2D
- `marketing_confusion_comparison.png`: LDA vs QDA performance
- `marketing_roc_curves.png`: Segment-specific classification quality

By working through the complete pipeline from data generation to model comparison, you will develop practical skills in applying discriminant analysis to real-world classification problems.

9. AI-Assisted Content Development

These companion notes were developed with the assistance of artificial intelligence (Claude, Anthropic) to enhance their pedagogical value and comprehensiveness. The AI contribution involved:

Content Organization and Clarity: Restructuring technical concepts into a logical progression from foundational theory to advanced applications, ensuring accessibility for students with diverse mathematical backgrounds.

Mathematical Exposition: Translating formal statistical notation into clear explanations with concrete interpretations, balancing mathematical rigor with practical understanding.

Example Development: Creating worked examples (marketing segmentation, credit risk classification) that demonstrate discriminant analysis application to realistic business problems, complete with numerical calculations and business interpretations.

Coverage Alignment: Enhancing content to ensure comprehensive coverage of assessment topics, including detailed discussions of diagnostic statistics (Wilks' Lambda, Hotelling's T-squared, eigenvalues, canonical correlation, Mahalanobis distance), coefficient interpretation (standardized vs unstandardized), and methodological considerations (categorical predictors, assumption violations).

Technical Accuracy: Verifying formulas, statistical concepts, and Python implementation details against authoritative references to ensure correctness.

The pedagogical structure, learning objectives, and example selection reflect the course instructor's vision (Dr. Juliho Castillo), with AI serving as a tool for content development and refinement rather than as the primary author. All technical content has been reviewed for accuracy and alignment with course objectives.

AI Tool Used: Claude 3.5 Sonnet (Anthropic) **Development Dates**: October 2024 - October 2025 **Human Oversight**: Dr. Juliho Castillo, Tecnologico de Monterrey

This disclosure reflects the growing role of AI in educational content creation while maintaining transparency about the development process. Students are encouraged to engage critically with all course materials, verify understanding through practice problems, and seek clarification from the instructor when concepts require additional explanation.

"The goal is to turn data into information, and information into insight."
- Carly Fiorina