**RAS 585: Assignment 1 Report**

# Part 1 (70 pts). Discriminant Analysis

## 1. Classes in dataset

Raisin dataset includes two classes: Kecimen and Besni.

## 2. Log Odds and Discriminant Function

Discriminant function between two Gaussian classes:

Since priors are equal,

## 3. Mean Vectors and Covariance Matrices

Mean and covariance matrices were computed for each class.  
Example values are shown in the console output (see Appendix if required).  
  
For Kecimen:  
μK=[63413.47,352.86,229.35,0.74,65696.36,0.71,983.69]\mu\_{K} = [63413.47, 352.86, 229.35, 0.74, 65696.36, 0.71, 983.69]μK​=[63413.47,352.86,229.35,0.74,65696.36,0.71,983.69]  
  
Covariance matrix (7×7):  
[[3.14e+08, 9.48e+05, 5.26e+05, 1.22e+02, 3.34e+08, -2.08, 2.55e+06],

[9.48e+05, 3.55e+03, 1.18e+03, 2.52, 1.04e+06, -0.52, 8.55e+03],

… [2.55e+06, 8.55e+03, 3.87e+03, 2.94, 2.79e+06, -1.03, 2.26e+04]]

For Besni:  
μB=[112194.79, 509.00, 279.62, 0.82, 116675.82, 0.69, 1348.12]\mu\_{B} = [112194.79, 509.00, 279.62, 0.82, 116675.82, 0.69, 1348.12] μB ​= [112194.79, 509.00, 279.62, 0.82, 116675.82, 0.69, 1348.12]  
  
Covariance matrix (7×7):  
[[1.54e+09, 3.69e+06, 1.78e+06, 3.13e+02, 1.59e+09, 3.50e+02, 9.10e+06],

[3.69e+06, 1.12e+04, 3.34e+03, 3.53, 3.92e+06, -0.72, 2.51e+04],

... [9.10e+06, 2.51e+04, 9.63e+03, 4.77, 9.72e+06, -1.03, 6.09e+04]]

## 4. Synthetic Samples

20 samples were generated from each multivariate Gaussian distribution.  
First 5 samples for each class are reported.  
  
- Kecimen  
[80225, 433, 245, 0.81, 88735, 0.63, 1212],

[58151, 356, 207, 0.83, 57637, 0.71, 916],   
…  
  
- Besni  
[167731, 625, 354, 0.82, 175978, 0.68, 1699],

[113945, 520, 277, 0.81, 116946, 0.57, 1399],   
...

## 5. Joint Distribution Visualization

3D surface plots of MinorAxisLength vs Perimeter were created.  
Finding: distributions are approximately Gaussian jointly (elliptical contours), but individually the features do not perfectly follow univariate Gaussians.

## 6. Likelihood Ratio

Where:

## 7. Discriminant Functions (separate covariance)

Discriminant functions were calculated using class-specific covariance matrices.  
Sample labels assigned correctly.

## 8. Discriminant Functions (pooled covariance)

Discriminant functions were computed with pooled covariance matrix.  
Labels differ slightly from Q7.

## 9. Confusion Matrices and Accuracy

Separate covariance:  
CM = [[430, 20], [106, 344]], Accuracy = 0.86  
  
Pooled covariance:  
CM = [[392, 58], [68, 382]], Accuracy = 0.86  
  
Conclusion: accuracy is the same, but error patterns differ.

# Part 2 (60 pts). 5-Fold Cross Validation

## 1. Mean Vectors and Covariance Matrices

Computed for each fold’s training data.  
For each fold, the training set was split into two classes (Kecimen and Besni).

- Cxomputed the class-specific mean vectors and covariance matrice**s** on the training data.

- This repeats the calculation of Part 1 (3), but only on the fold’s training portion.

## 2. Discriminant Functions (separate covariance)

Applied Gaussian discriminant with class-specific covariances.  
  
  
Example Fold 1 confusion matrix:  
[[66, 20], [7, 87]], Accuracy = 0.85

## 3. Discriminant Functions (pooled covariance)

Used pooled covariance across classes.  
  
  
Example Fold 1 confusion matrix:  
[[76, 10], [14, 80]], Accuracy = 0.867

## 4. Average Rates across 5 folds

Separate covariance (averaged across 5 folds):   
- Accuracy ≈ 0.850,   
- TPR ≈ 0.947,   
- TNR ≈ 0.756,   
- FPR ≈ 0.244,   
- FNR ≈ 0.053  
  
Pooled covariance (averaged across 5 fold):   
- Accuracy ≈ 0.858,   
- TPR ≈ 0.870,   
- TNR ≈ 0.848,   
- FPR ≈ 0.152,   
- FNR ≈ 0.130

## 5. Performance Comparison

Separate covariance:   
higher sensitivity (TPR ~95%), but at the cost of more folse positives (FPR ~24%).  
  
Pooled covariance:   
slightly higher overall accuracy (85.8% vs 85.0%) and much better specificity (TNR ~85% vs 76%), but sensitivity is lower (TPR ~87%).

Conclusion:   
- If the goal is to catch as many positives as possible (minimize false negatives), separate covariance is better.

- If the goal is to balance errors and reduce false positives, pooled covariance is preferable.

- On this dataset, pooled covariance performed slightly better overall.

# Part 3 (70 pts). Regression and PCA

## 1. Visualization of Perimeter vs Independent Variables

Scatter plots show trends:  
- Strong positive: Area, MajorAxisLength, ConvexArea  
- Moderate positive: MinorAxisLength  
- Weaker: Eccentricity  
- Negative: Extent  
  
Findings**:**

- Strong positive trends: Area, MajorAxisLength, ConvexArea (all nearly linear).

- Moderate positive trend: MinorAxisLength.

- Weaker relationship: Eccentricity.

- Negative correlation: Extent decreases as Perimeter increases.

## 2. Correlation Matrix (training data - first 700 samples)

Top correlations with Perimeter:  
- ConvexArea: 0.979  
- MajorAxisLength: 0.979  
- Area: 0.963  
- MinorAxisLength: 0.820  
- Eccentricity: 0.442  
- Extent: -0.220

Observation (Conclusion):  
- Variables like ConvexArea, MajorAxisLength, and Area have the largest impact on Perimeter.

- Correlation does not imply causation — the relationships may be due to geometric dependencies, not direct cause-effect.

## 3. Linear Regression Model (training data - first 700 samples)

Coefficients:  
- Eccentricity: +54.11  
- MinorAxisLength: +1.72  
- MajorAxisLength: +1.70  
- ConvexArea: +0.00725  
- Area: -0.00755  
- Extent: -88.81  
  
Interpretation:

- Eccentricity has a surprisingly high weight (likely due to scaling).

- MajorAxisLength, MinorAxisLength, ConvexArea positively affect Perimeter.

- Extent has a strong negative coefficient.

- Area has a small negative coefficient due to collinearity with ConvexArea.

## 4. Prediction Performance (training data - first 700 samples)

Test MSE (linear regression) ≈ 666.0

## 5. Linear Dependence Discussion

Condition number ≈ 1.3× → strong multicollinearity (very high).  
Impact: unstable coefficients, inflated variance.

## 6. Principal Component Analysis (PCA)

Applied PCA to the independent variables (training data).

- The explained variance ratio decreases across components.  
- Most variance is captured by the first few PCs.

## 7. Pareto Chart

Pareto bar chart shows variance first few PCs and line explain majority of cumulative variance.

## 8. Regression with First 4 PCs

- Linear regression on original features (MSE ~666) performed better than PCA regression (MSE ~749).

- This suggests PCA is not helpful here since the dataset is not very high-dimensional and original features already capture strong linear relationships

## 9. Practical Scenarios for PCA

- Use PCA: many correlated predictors, dimensionality reduction needed.  
- Do NOT use PCA: when interpretability is important or predictors are already few.