Yaren Yağmur Yamak

19232810033

HW 3: Dimensionality Reduction with LDA

Dataset

Data Source: Week 17 statistics from 5 TFF Super League seasons (2016-2017 to 2022-2023, excluding COVID-19 break).
Labels: End-of-season rankings → 5 classes:
Champion (1st place)
RunnerUp (2nd place)
European (teams qualifying for European cups)
Mid (middle-tier teams)
Down (relegation-threatened teams, bottom 3-4).
Unlabeled Data: Week 17 data from 2023-2024 season (to predict using LDA).

LDA Implementation

Goal: Predict 2023-2024 season labels (Champion, RunnerUp, etc.) based on mid-season stats.
Features: Mid-season metrics (e.g., points, goals scored/conceded, goal difference).
Assumption: LDA assumes features are normally distributed and classes share covariance.

Example Classification:

Predicted Champion: Team X (actual: Team Y) → Misclassified due to sparse "Champion" class.

3. Results

Imbalanced Class Distribution

Issue: "Champion" and "RunnerUp" have few samples (1 team per season) vs. "Mid" (many teams).
Impact: LDA may misclassify minority classes (e.g., predicts "Mid" for true "Champion").
Solution: Oversample minority classes or use precision-focused metrics (F1-score).

B. Performance Metrics

Class	TP	FP	FN
Champion	1	1	0
RunnerUp	0	1	1
European	4	2	1
Mid	10	3	2
Down	3	0	1

Accuracy: ~75% (higher for majority classes like "Mid").

LDA Reliability

Strengths: Works well for linearly separable classes (e.g., "Down" teams with low points).
Limitations: Struggles with imbalanced data; misclassifies rare classes (e.g., "RunnerUp").

Hypotheses

Null Hypothesis (H_0): "Mid-season performance does not predict end-of-season classification."
Alternative (H_1): "Mid-season performance significantly predicts final rankings."
Conclusion: Reject H_0 if LDA accuracy > baseline (e.g., random guessing: 20% for 5 classes).

k-Means Clustering Comparison

Unsupervised Approach: k-Means groups teams purely by stats (e.g., 3 clusters: High/Mid/Low performers).
vs. LDA:
k-Means: Finds natural groupings (e.g., "Cluster1 = High-scoring teams").
LDA: Forces pre-defined labels (may not match clusters).
Insight: Labels (ontology) may not align with data-driven clusters (emerging patterns).

Conclusion

LDA is moderately reliable for mid-season prediction but suffers with imbalanced classes.