

ECON484 Spring 2025

HW 3

Goal: To gain experience in the analysis, design and implementation of dimension reduction.

Explanation:

Consider the in-class example for Linear Discriminant Analysis (LDA). In that example we had

- Labeled data from week 34 of 2022-2023 and 2023-2024 seasons of TFF Super League. The labels were somewhat arbitrarily made.
- Unlabeled data from week 34 of 2024-2025 season of the same league. The labels for these were assigned by LDA.

Now you will be required to make some changes, observe the results and make comments.

Follow the steps below:

1. Create a similar dataset with the following properties
 1. Take the data from week 17 of the same league for 5 seasons, skipping the COVID break (2022-2023, 2021-2022, 2018-2019, 2017-2018, 2016-2017).
 2. Label the teams using the last week of each season, so that the labels are all end results (ie. last standings).
 1. Champion (Single Team, 1st place)
 2. RunnerUp (Single Team, 2nd place)
 3. European (Multiple teams, those that qualify for European cups)
 4. Mid (All other teams)
 5. Down (Bottom 3 or 4 teams, subject to relegation)
 3. Add the last year's season (2023-2024) week 17 data without labels. Note that for that season we know who was champion, etc.
2. Run the same algorithm (not much change needed for the code, but make as you wish).
3. Discuss the results, with a particular focus on misclassified (i.e., incorrect) predictions.
 1. Does the classification scheme correctly classify the Champion and RunnerUp? Discuss the results with your research on the concept of a “imbalanced or sparse class distributions” and how it effects classification algorithms, in particular LDA.
 2. For all classes, calculate the True Positive, True Negative, False Positive, False Negative values by comparing the algorithm result and manual classification results.
 3. How reliable is this application of LDA as an estimation method?

4. State the null and alternative hypotheses implied in using LDA to classify team outcomes based on mid-season statistics. For example: 'H0: Mid-season team performance does not significantly predict end-of-season classification.'
5. Would you consider the labels Champion, RunnerUp, Mid, etc. part of a taxonomy or an ontology? How would an ontological structure differ in representing this knowledge?
6. Try running a clustering algorithm (e.g., k-means) on the same data and compare how the groupings differ from the supervised labels. What does this tell you about ontology vs. data-driven group emergence?