

Unsupervised Learning: Validation

Exploratory vs. Confirmatory Analysis

Confirmatory Analysis:

- Seeks to test an a priori hypothesis.
- Examples:
 - ▶ Classical inferential statistics.
 - ▶ Prediction in supervised learning.

Exploratory Analysis:

- Seeks to explore and understand data.
- Hypothesis generating.

Which is unsupervised analysis?

Unsupervised Learning Objectives

① Data exploration.

- ▶ Explore data to find patterns and generate hypotheses.
- ▶ Exploratory analysis - Generated hypotheses will be later tested and confirmed.

② Data-driven discoveries.

- ▶ Make discoveries by finding important patterns and structures in the data.
- ▶ **Challenge:** Both exploratory & confirmatory analysis!

Data Exploration

- Understand patterns, trends, and anomalies in your data.
- Prepare your data for the modeling stage and confirmatory analysis.
- Give a visual summary of the data.
- Generate hypotheses to be confirmed later.

We need to use multiple techniques to fully explore and visualize data!

From Data Exploration to Data-Driven Discoveries

Challenge:

- Any potential discoveries found during data exploration (hypothesis generation) cannot be confirmed on the same data set (hypothesis testing).

Why not?

From Data Exploration to Data-Driven Discoveries

Challenge:

- Any potential discoveries found during data exploration (hypothesis generation) cannot be confirmed on the same data set (hypothesis testing).

Why not?

Overfitting!

From Data Exploration to Data-Driven Discoveries

Challenge:

- Any potential discoveries found during data exploration (hypothesis generation) cannot be confirmed on the same data set (hypothesis testing).

Why not?

Overfitting!

What can we do to validate data-driven discoveries?

Validating Data-Driven Discoveries

- ① Corroborate via existing literature.
 - ▶ Suggestive, but further confirmation needed.

Validating Data-Driven Discoveries

- ① Corroborate via existing literature.
 - ▶ Suggestive, but further confirmation needed.
- ② Show data-driven discovery is **stable**.
 - ▶ Small changes to the data, the algorithm, the method, the parameters, etc. yield the same or similar result.
 - ▶ Communicates uncertainty in discovery and avoids overfitting.

Validating Data-Driven Discoveries

- ➊ Corroborate via existing literature.
 - ▶ Suggestive, but further confirmation needed.
- ➋ Show data-driven discovery is **stable**.
 - ▶ Small changes to the data, the algorithm, the method, the parameters, etc. yield the same or similar result.
 - ▶ Communicates uncertainty in discovery and avoids overfitting.
- ➌ Confirm via a completely separate **test set**.
 - ▶ Strong validation and standard in machine learning.
 - ▶ Sometimes challenging for unsupervised learning.

Validating Data-Driven Discoveries

- ➊ Corroborate via existing literature.
 - ▶ Suggestive, but further confirmation needed.
- ➋ Show data-driven discovery is **stable**.
 - ▶ Small changes to the data, the algorithm, the method, the parameters, etc. yield the same or similar result.
 - ▶ Communicates uncertainty in discovery and avoids overfitting.
- ➌ Confirm via a completely separate **test set**.
 - ▶ Strong validation and standard in machine learning.
 - ▶ Sometimes challenging for unsupervised learning.
- ➍ Validate via biological experiments.
 - ▶ Gold Standard! True confirmation.
 - ▶ Expensive & sometimes not possible.

Confirming Discoveries on a Test Set

Idea:

- Use a “training set” for data exploration to make discoveries.
- Use a separate “test set” for confirming the discovery.

Approach:

- Use a separate study for a test set.
- Randomly split samples into a separate training and test set before any analysis.

Important:

- Keep test set hidden until confirmatory stage!

But, this is often challenging for unsupervised learning . . .

Discussion

How would you use a separate test set to confirm the following types of data-driven discoveries'?

- Pattern.
- Clusters.
- Selected features.

Example: Utilizing a Test Set for Clustering

- Cluster the training data with K clusters.
- Separately, cluster the test data with K clusters.
- On training data, build a predictive model, $\hat{f}()$, to predict the K cluster labels.
 - ▶ E.g. KNN or RF.
- Use $\hat{f}()$ to predict labels on test set.
- Use a metric to compare test set clusters to test set predicted labels.
 - ▶ E.g. Rand, Jaccard, or Adjust Rand Index.

Stability Principle

Idea:

- Reproducible discoveries are likely to be true discoveries.

Approach:

- Use repeated data perturbations of training data to mimic having new test data.
 - ▶ Subsampling, bootstrapping, random corruptions, random noise, random tuning parameters, random initializations, etc.
- Apply machine learning technique to each perturbed data set.
- If the same discovery appears repeatedly, then it's likely a true discovery.

Stability Principle

Idea:

- Reproducible discoveries are likely to be true discoveries.

Approach:

- Use repeated data perturbations of training data to mimic having new test data.
 - ▶ Subsampling, bootstrapping, random corruptions, random noise, random tuning parameters, random initializations, etc.
- Apply machine learning technique to each perturbed data set.
- If the same discovery appears repeatedly, then it's likely a true discovery.
- Useful when a separate test set is unavailable or difficult to confirm via a test set.
- Communicates level of uncertainty associated with the data-driven discovery.

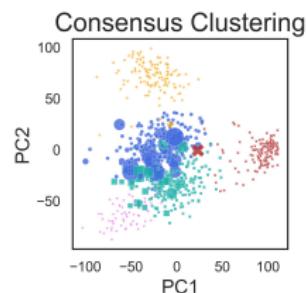
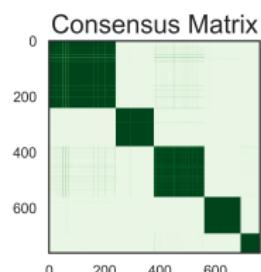
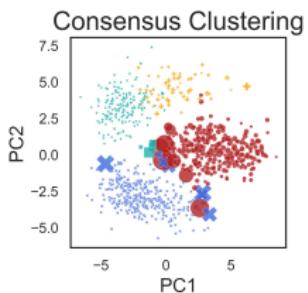
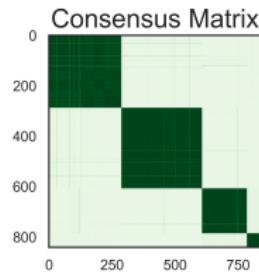
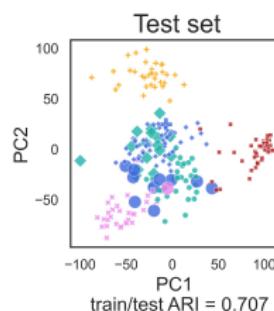
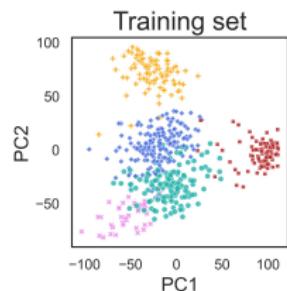
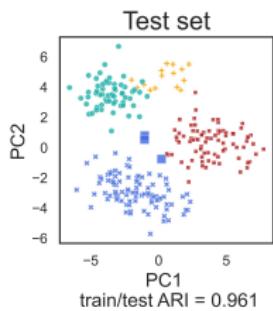
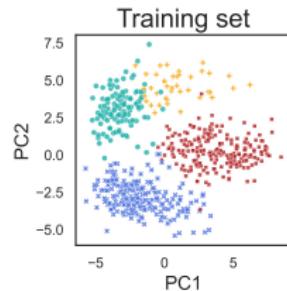
Discussion

How would you use the stability principle to confirm the following types of data-driven discoveries?

- Patterns.
- Clusters.

How would you randomly perturb your data to perform stability analysis?

Cluster Validation Example



(a) Author

(b) PANCAN

Allen, Gan, & Zheng, "Interpretable Machine Learning for Discovery: Statistical Challenges and Opportunities", ARSIA 2024

Unsupervised Learning: Best Practices

Summary

Topics & Techniques covered:

- Dimension Reduction.
 - ▶ PCA.
- Pattern Recognition & Data Visualization.
 - ▶ PCA, NMF, ICA, MDS, tSNE, UMAP.
- Clustering.
 - ▶ K-means, Hierarchical, Bioclustering, Convex clustering.
- Feature Filtering via Association Testing.
 - ▶ FWER, FDR, Permutation approaches.
- Graphical Models.
 - ▶ Graph types, Gaussian graphical models.
- Validation.

Some Good Rules for Unsupervised Analysis

- ① Always visualize.
- ② Use multiple techniques.
- ③ Validate discoveries.
- ④ Communicate uncertainty.
- ⑤ Make your analysis reproducible.