Data Mining - Lecture S6: Data Preprocessing

NOVA IMS

1 Introduction to Data Preprocessing

Data preprocessing transforms original data into a more compact and informative representation through:

- Missing values treatment
- Outlier detection
- Discretization and encoding
- Handling imbalanced datasets
- Feature selection and engineering
- Feature scaling and normalization

2 Reasons for Data Preprocessing

- Noise reduction and signal amplification
- Input space size reduction
- Domain-specific knowledge application
- Feature selection and engineering
- Scaling (normalization)

3 Curse of Dimensionality

- Input space grows exponentially with number of variables
- More data and computing power required in high dimensions
- Distance metrics lose meaning
- Cluster identification becomes more difficult

4 Feature Selection Principles

Two major principles for input space reduction:

4.1 Relevancy

- Irrelevant variables don't affect the target
- Feature selection measures which variables best separate classes
- Use information gain measures (e.g., entropy)
- Features are interesting if they reduce uncertainty

4.2 Redundancy

- Redundant variables are highly correlated with others
- Remove correlated variables to reduce dimensionality

5 Feature Engineering

Creating input combinations:

- Height^2/weight (BMI)
- Population/area
- Euros spent/number of purchases
- Euros spent/time as customer
- Debt/income

6 Principal Component Analysis (PCA)

PCA uses orthogonal transformation to convert correlated variables into linearly uncorrelated principal components.

6.1 Key Properties

- Number of PCs equals number of original variables
- PC1 captures maximum variance
- PC2 captures next largest variance (orthogonal to PC1)
- Subsequent PCs explain decreasing variance

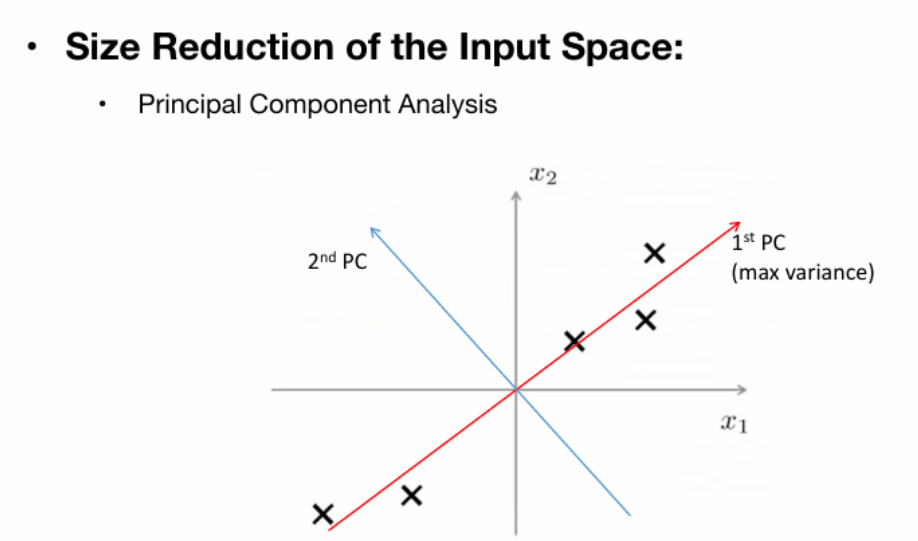

Figure 1: Principal Components showing maximum variance directions

6.2 Variance Explanation Example

- PC1: 60% of variance
- PC2: 25% of variance
- PC3: 10% of variance
- First three components: 95% total variance

Important Note: PCA preserves variance, not necessarily class separability. Dimensionality reduction can hurt supervised learning.

7 Data Standardization & Normalization

Many algorithms assume features are on comparable scales.

7.1 Normalization Methods

- Min-Max: $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$
- Z-score: $z = \frac{x - \mu}{\sigma}$

8 Useful Resources

- Principal Component Analysis (PCA) Documentation

- PCA Video Explanation
- Data Normalization & Standardization
- How to Calculate Standard Deviation