Data preprocessing and feature engineering

APSC 8280: Machine learning applied to plant science

04/02/2020

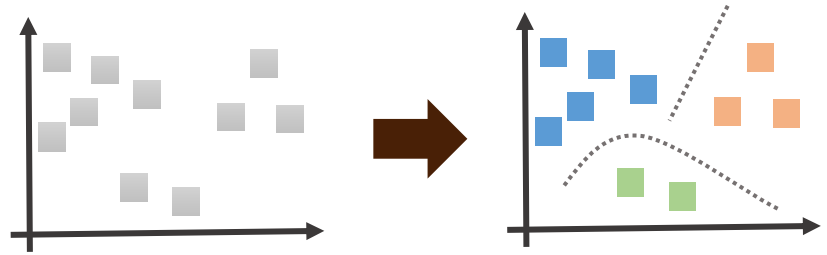
Outline

- Unsupervised learning
- Dealing with missing values
- Feature transformation
- Feature extraction
- Feature selection

Unsupervised learning

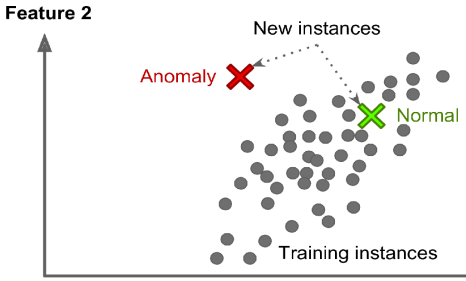
Clustering

- K-means clustering
- Hierarchical clustering
- Model-based clustering



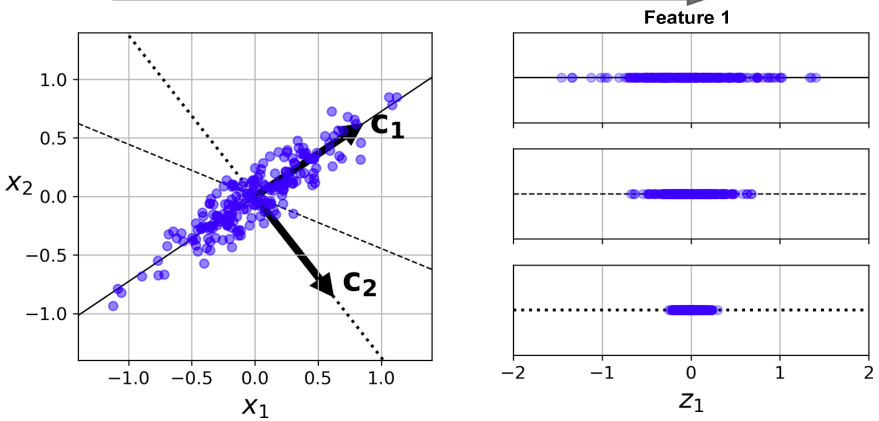
Anomaly and novelty detection

- One-class SVM
- Isolation forest



Visualization and dimensionality reduction

- (Kernel) Principal component analysis
- Locally Linear Embedding(LLE)
- t-Distributed Stochastic Neighbor Embedding (t-SNE)



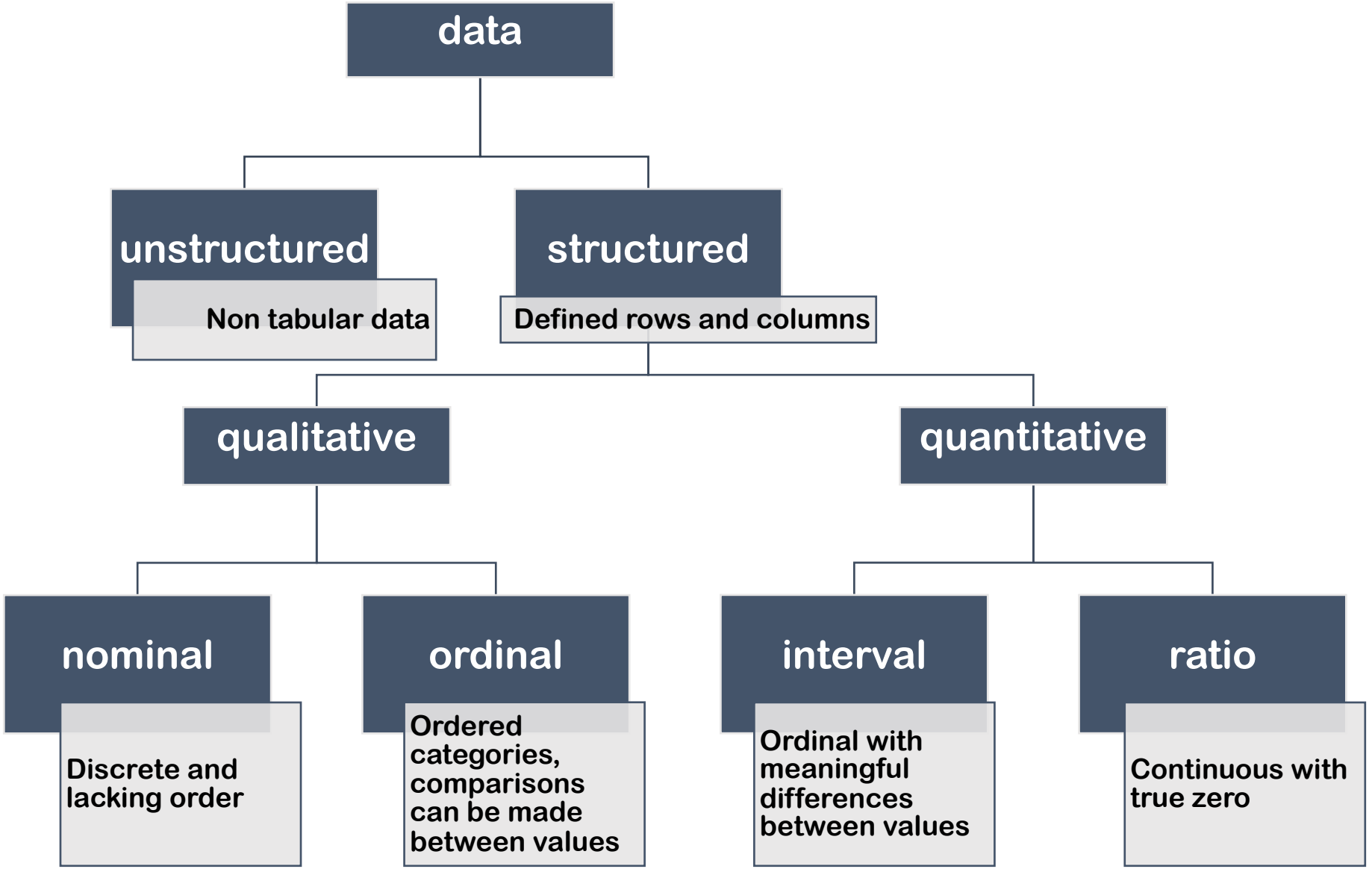
Association rule learning

- Apriori
- Eclat

https://www.mathworks.com/help/stats/dendrogram.html

Hands-on machine learning with Scikit-Learn, Keras and Tensorflow

What's in your dataset?



Dealing with missing values

- Complete case analysis
- Imputation with feature statistic (mean, median, mode)
- Model-based imputation (k-nearest neighbors, linear regression, etc.)
- Multiple imputation by chained equations (MICE)

Feature transformation

- Data normalization
  - Z-score standardizatin
  - Min-max scaling
  - Row normalization
- Binning continuous features into categories
- Encoding categorical variables
  - One-hot / dummy encoding
  - Label encoding
  - Target encoding

One-hot encoding

color
red
green
blue
red

color_red	color_blue	color_green
1	0	0
0	0	1
0	1	0
1	0	0

Label encoding

color
red
green
blue
red

color
0
1
2
0

target encoding

color	target
red	1
green	0
blue	0
red	1

Encoding	target
1.00	1
0.5	0
0.5	0
1.00	1

Adapted from: https://towardsdatascience.com/know-about-categorical-encoding-even-new-ones-c266227b9cbd

Feature extraction

- Kernel-based feature extraction
- Text-specific feature construction
  - Bag of words (tokenizing, counting and normalizing)
  - Term frequency –inverse document frequency (tf-idf)
- Domain-knowledge based feature creation
- Using ML to learn new features from your data

Feature selection

- Statistical-based feature selection
  - Correlation coefficients
  - Hypothesis tests
- Model-based feature selection
- Iterative feature selection