LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Introduction to Machine Learning

Unsupervised Learning

Bernd Bischl, Christoph Molnar, Daniel Schalk, Fabian Scheipl

Department of Statistics – LMU Munich

Motivation

UNSUPERVISED LEARNING

- Supervised machine learning deals with *labeled* data, i.e., we have input data x and the outcome y of past events.
- Here, the aim is to learn relationships between x and y.
- Unsupervised machine learning deals with data that is *unlabeled*, i.e., there is no real output y.
- Here, the aim is to search for patterns within the inputs x.

MOTIVATION FOR CLUSTERING

Consider multivariate data with n observations (e.g. customers) and p features (e.g. characteristics of customers).
Task: divide data into groups (clusters), such that

- the observations in each cluster are as "similar" as possible (homogeneity within each cluster), and
- the clusters are as "far away" as possible from other clusters (heterogeneity between different clusters).

CLUSTERING VS. CLASSIFICATION

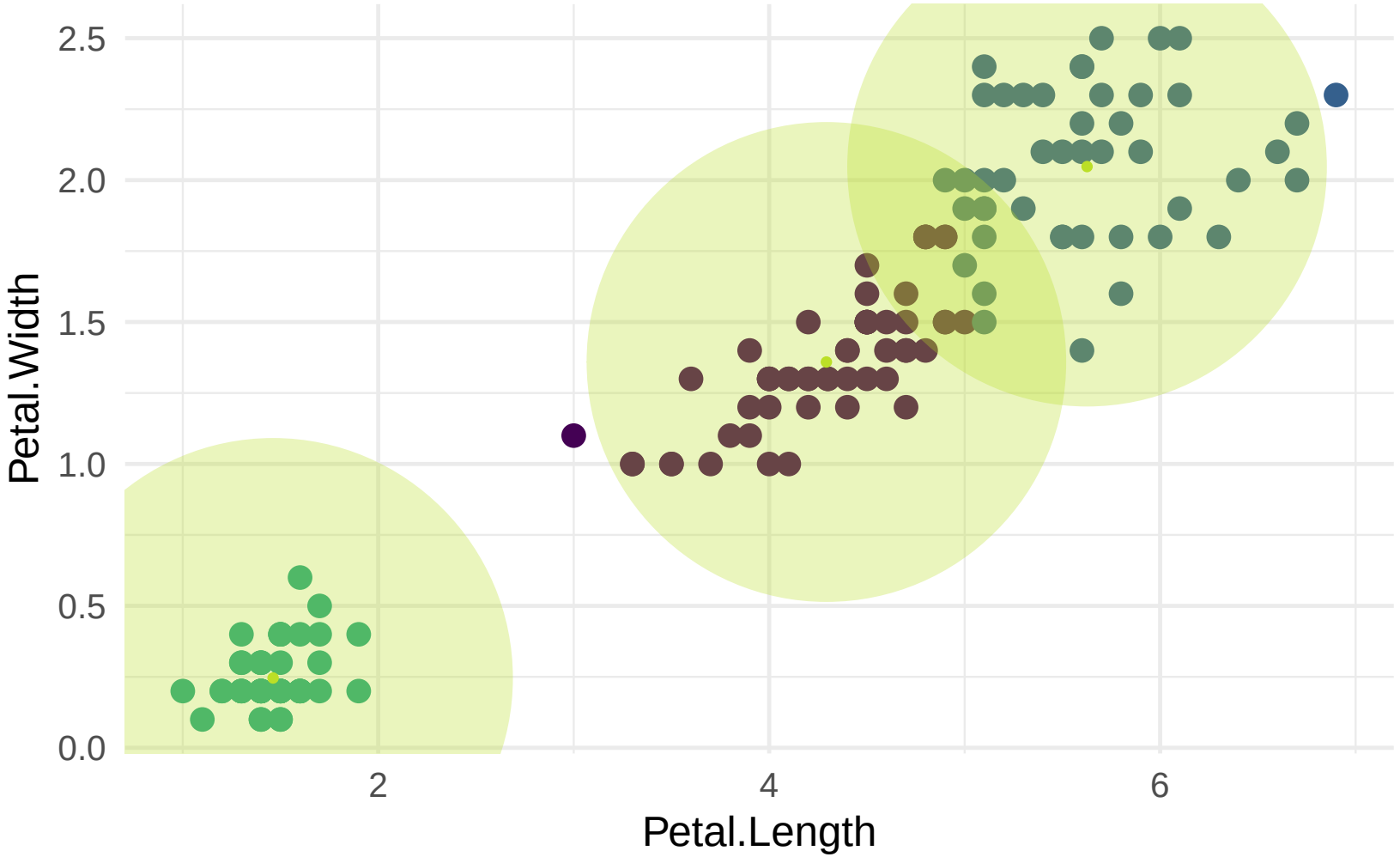
- In classification, the groups are known and we try to learn what differentiates these groups (i.e., learn a classification function) to properly classify future data.
- In clustering, we look at data, where groups are unknown and try to find similar groups.

Why do we need clustering?

- Discovery: looking for new insights in the data (e.g. finding groups of customers that buy a similar product).
- Derive a reduced representation of the full data set.

CLUSTERING TASK

Goal: Group data into similar clusters (or estimate fuzzy membership probabilities)



CLUSTERING: CUSTOMER SEGMENTATION

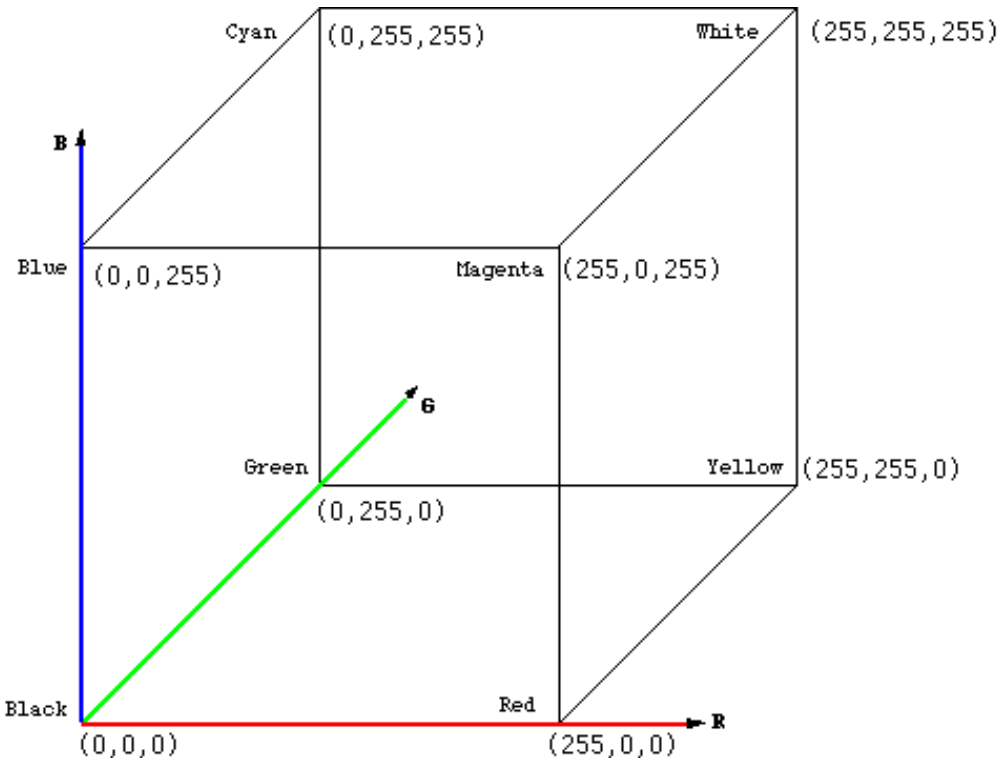
- In marketing, customer segmentation is an important task to understand customer needs and to meet with customer expectations.
- Customer data is partitioned in terms of similarities and the characteristics of each group are summarized.
- Marketing strategies are designed and prioritized according to the group size.

Example Use Cases:

- Personalized ads (e.g., recommend articles).
- Music/Movie recommendation systems.

CLUSTERING: IMAGE COMPRESSION

- An image consists of pixels arranged in rows and columns.
- Each pixel contains **RGB** color information, i.e., a mix of the intensity of 3 **primary colors**: **R**ed, **G**reen and **B**lue.
- Each primary color takes intensity values between 0 and 255.



Source: By Ferlixwangg CC BY-SA 4.0, from Wikimedia Commons.

CLUSTERING: IMAGE COMPRESSION

An image can be compressed by reducing its color information, i.e., by replacing similar colors of each pixel with, say, k distinct colors.

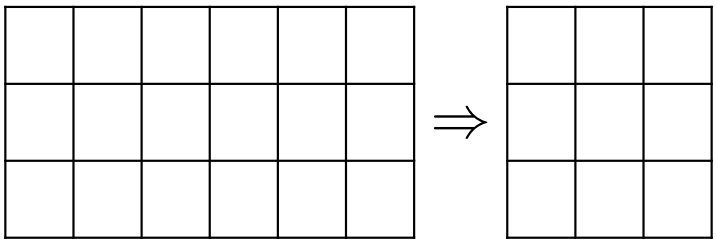
Example:

Original Image

Image using 16 Colors

DIMENSIONALITY REDUCTION TASK

Goal: Describe data with fewer features (reduce number of columns).
$\Rightarrow$ there will always be an information loss.



Unsupervised Methods:

- Principle Component Analysis (PCA).
- Factor Analysis (FA).
- Feature filter methods.

Supervised Methods:

- Linear Discriminant Analysis (LDA).
- Feature filter methods.