Lecture Pattern Analysis

Part 05: Introduction: Simplifications of the Feature Space

Christian Riess

IT Security Infrastructures Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg

May 7, 2021

Introduction

- So far, we looked at different options to represent a set of samples:
  - Local operators from fixed neighborhood relationships:
    Non-parametric Density Estimation via Kernels and k-NN
  - Local operators from learning-based sample space partitioning:
    Trees and Random Forests
- Compared to our kernels, tree-based methods
  - do not need to store all samples to respond to a query
  - subdivide the sample space dynamically with an objective function
- However, beyond the representation itself, these methods do not provide much information about the distributions of samples
- In the upcoming second part of the lecture, we will look at representational simplifications to make the distributions interpretable:
  - **Clustering** segments the data into few meaningful groups
  - **Manifold Learning** reduces the sample space dimensionality while preserving the structure of the data

Clustering

- The goal is to assign identical labels to similar samples
- Difference to classification/regression: Clustering is unsupervised
- Hence, clustering applications oftentimes explore data, e.g.:
  - Which gene expressions cause which type of cancer[1]?
  - Which other products attract customers who buy coffee when it is discounted?
- We will investigate these specific algorithms:
  - k-means
  - Gaussian Mixture Models
  - Mean Shift
- We will also address the model selection problem, i.e., the selection of the hyperparameters

[1] See Hastie/Tibshirani/Friedman Sec. 14.3.8 for a k-means example

Manifold Learning

- The goal is to represent the data manifold in a lower dimensional space, i.e., to perform a structure-preserving mapping to a lower dimension
- Oftentimes, manifold learning is directly integrated into a PR pipeline, e.g.,
  - as pre-processing step to reduce the dimensionality of the input, e.g., the 100s of spectral bands in remote sensing are highly correlated
  - within the feature extraction step to make the classifier input “denser”
- For deep neural networks, manifold learning is oftentimes used to visualize that good features have been learned
- We will investigate these specific algorithms:
  - PCA (known, I guess?)
  - Multi-dimensional Scaling
  - ISOMAP
  - Laplacian Eigenmaps
- If time permits, we can also touch applications of spectral graph processing