FRIEDRICH-ALEXANDER-
UNIVERSITÄT
ERLANGEN-NÜRNBERG

SCHOOL OF ENGINEERING

Lecture Pattern Analysis

Part 16: Short Recap and Remarks on the Exam

Christian Riess
IT Security Infrastructures Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg
June 13, 2021

Introduction and High-Level Review

- We finished Parts 1 and 2 of Pattern Analysis
  - Part 1 focuses on representations of a sample space
  - Part 2 focuses on simplifications via clustering and manifold learning
- High-level remarks on some (potentially implicit) aspects are listed below

1. Besides the actual algorithms, PA also has a “latent space” of topics
   - Local operators in the feature space
   - Tradeoffs on model complexity and flexibility (e.g., fixed kernel neighborhoods vs. data-driven random forest neighborhoods)
   - Model selection
2. Tools/tricks can oftentimes be re-used in different situations:
   - Space partitioning: Random Forest splits can be used in classification, regression, DE, ML (btw., the same is true for kernels and the k-NN operator)
   - A reference distribution is the backbone of the Gap Statistics, and we also used it to approximate a density via regression

High-Level Review (Continued)

3. Different optimization criteria or algorithm variants can lead to different results:
   - Kernel density estimates can be zero somewhere, K-NN density estimates not
   - The clusters from k-means, mean shift, GMMs can have very different shapes
   - The gap statistics is a good match for k-means clusters, but not for mean shift
   - PCA/MDA/ISOMAP projections preserve long distances, LE preserves local neighborhoods. The latter makes non-linear projections more robust
4. Algorithm assumptions are important
   - Kernel density estimation assumes a number of samples in the kernel window
   - Random Forest training must decorrelate the trees, at least to some extent
   - Individual trees in a random forest must perform better than random guessing (50% in a 2-class problem)
5. Computational requirements (space/time) are important
   - Kernel density estimation must store and lookup all samples for a query, or pre-compute the whole d-dim. density
   - Random Forests variants are oftentimes more efficient (and also more expressive), even more since they are trivially parallelized

Hints on the Exam

- 60 minutes, 60 points, just a pen (no books, cheat sheets, ...)
- Most questions require 1–3 sentences as answer
- Few questions require a sketch, very few are multiple choice
- Questions will be a combination of three levels of mental productivity:
  1. Reproduction Questions, for example
     - Write down the objective function for calculating the mean shift vector
     - State the algorithmic steps of the ISOMAP algorithm
     - Name 3 options for randomization in Random Forest training
  2. Explanation Questions, for example
     - How does a Random Forest achieve a smooth classification boundary?
     - What complicates working in high-dimensional spaces?
  3. Comparison / Analysis Questions, for example
     - Was the density in Fig. X created from a box kernel or Gaussian kernel? Why?
     - Sketch a sample distribution where Laplacian Eigenmaps with kernel-based affinities might fail, but Manifold Forests might work

Hints on the Preparation

- It may pay to practice already during the preparation short answers to some questions
- There are two sources of preparation material online:
  - The PA 2021 class material (our studOn class):
    This is the reference for the exam, including everything that is on studOn
  - The PA 2018 class material (on video.fau.de):
    Blackboard lecture; large overlap with 2021, but please check for differences
- The exam will not require you to write code
- The exam will not have questions on content that only occurs in the supplemental literature (Bishop, Hastie/Tibshirani/Friedman) but not in the lecture/exercises/joint meetings
- Many learner types benefit from learning groups
  Corona tries to screw this up to the greatest extent possible, but try to reach out to colleagues