0 - Introduction

Math 728 D - Machine Learning & Data Science - Spring 2019

What is this about?

Central theme: “understanding” large data sets ... more precisely

... extraction of quantifiable information from large data sets ...

Some categories:

Machine learning

(UL) Unsupervised Learning

Processing “matrix data” - point clouds - ; probabilistic Exploration of large structures e.g. by random walks, sampling strategies; partitioning data into groups of similar objects - clustering - dimension reduction (projection into lower dimensional spaces while nearly preserving mutual distances), dictionary learning, learning probability densities, etc. ...

(SL) Supervised Learning

Learning from “labelled data” - training sets -; central tasks: regression (learning conditional expectations), classification (assigning labels from a finite list).

Data Assimilation

Inferring from data using additional prior information about the “searched states”, e.g. given in terms of mathematical models (parametric families of partial differential equations)

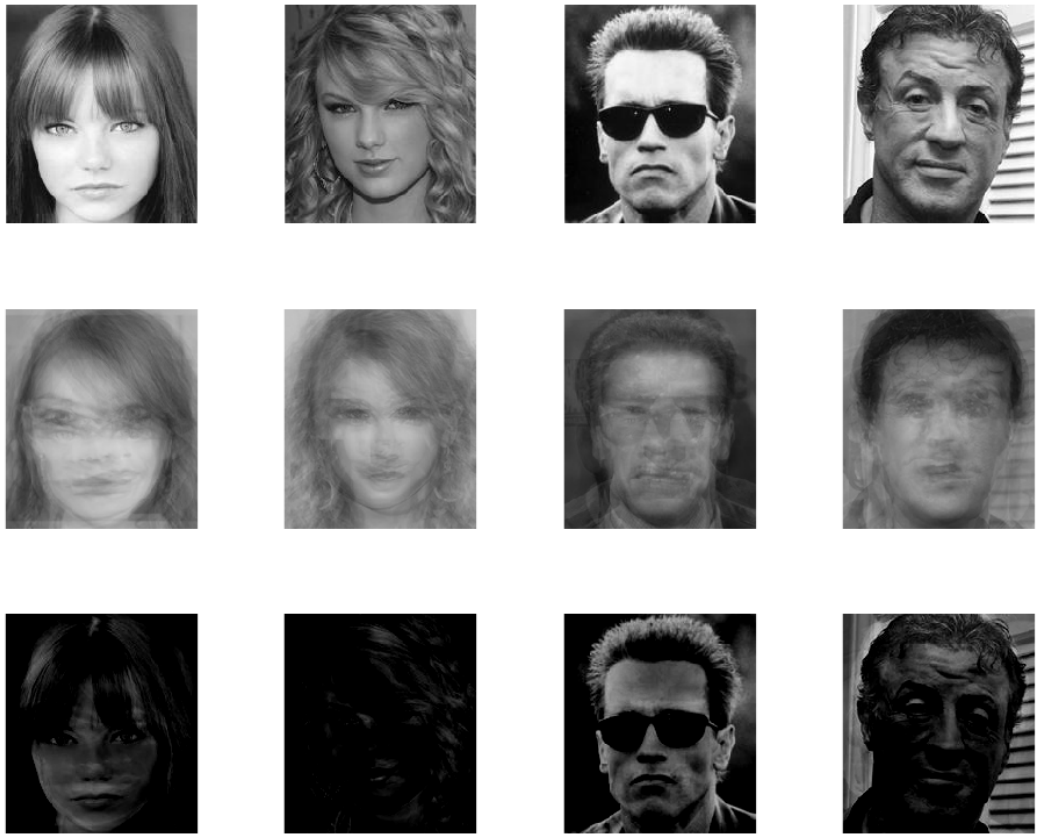
Common Challenges

high dimensionality, viz. recovering functions of a large number of variables - curse of dimensionality -, ill-posedness,...

Matrix-Data - Point Clouds

Face recognition, Fingerprints: Machine learning methods underly modern facial recognition and fingerprint matching algorithms. The data sets are enormous, but the SVD can extract a much smaller kernel of vital information. When a new face image or fingerprint must be identified, the SVD provides an optimal set of “questions”’ to ask (in the form of basis vectors), which can identify the obiect.

Images, Averages, Image-Average



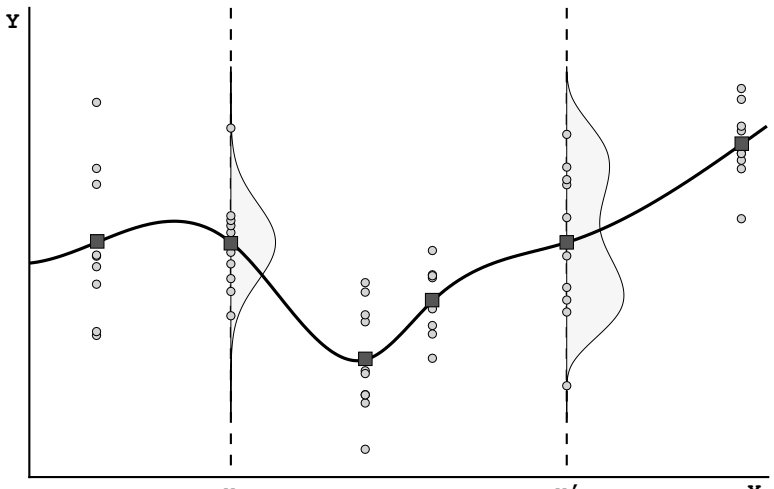
Mathematical Tools/Concepts

Data come as vectors (poins) $\mathbf{x} \in \mathbb{R}^d$ for very large d. For instance, a digital image is a large array of grey shade or color values. This draws heavily on the following mathematical subjects:

- Linear Algebra, especially matrix factorization such as Singular Value Decomposition (SVD) and related other spectral factorizations (e.g. Laplacian eigenmaps)

  Applications: principal component analysis, model reduction, discrete optimization, regularization, learning statistical mixtures of probability densities,...
- High-dimensional geometry, how to measure distances in high dimensions? finding nearest neighbors, sampling strategies...
- Probabilistic concepts, tail bounds, concentration inequalities, dimension reduction, to reduce complexity...
- Constrained and unconstrained continuous optimization, stochastic optimization,...

Regression



Risk functional: $\mathcal{E}(f) := \int_Z (y - f(x))^2 dP \rightsquigarrow$

P unknown probability measure on $Z := X \times Y$

Factorization into conditional and marginal densities $dP(x, y) = dP(y|x)dP_X(x)$

Goal: estimate the regression function

$$f_P(x) := \int_Y y dP(y|x) = \mathbb{E}(y|x)$$

$$\mathcal{E}(f) = \mathcal{E}(f_\rho) + \|f - f_p\|^2_{L_2(X,p_X)}, \quad \|\cdot\| := \|\cdot\|_{L_2(X,p_X)}$$

Task: construct an estimator $\hat{f}_Z$ that approximates f_p well in $L_2(X, p_X)$ using independent, identically distributed (iid) random samples drawn from the underlying unknown probability distribution

Typical question: how well is an expectation $\mathbb{E}[X]$ of a random variable X approximated by sample means $\frac{1}{N}\sum_{j=1}^{N} X_j$, under which assumptions on the probability distribution?

Classification

Example:

$Z = X \times Y$, the set Y of labels is discrete and finite;

samples (training data) $=$ set of medical images, each being classified as showing healthy (label 1) or cancerous (label (0)) tissue, i.e., $Y = \{0, 1\}$

each image is represented by a large vector $x \in X = \mathbb{R}^N$ (grey or color values)

Issues, Mathematical Tools/Concepts

Some background on the following is needed:

- Basics from probability theory and statistics, conditional probability, frequentists and Baysian concepts, concentration inequalities, tail bounds,...
- Discrete and continuous concepts meet,...
- Constrained and unconstrained continuous optimization, stochastic optimization, regularization...
- Principles for measuring complexity (covering numbers, VC-dimension,...)
- Approximating functions of many variables, Curse of Dimensionality,...
- Representation formats for functions of many variables: dictionary representations, decision trees, tensor/low rank represenations, multilayer compositional representations - deep neural networks;