Big Data Analysis Application and Practice (XAI605)

Introduction

2023 Spring

Instructor: Sejun Park

Instructor

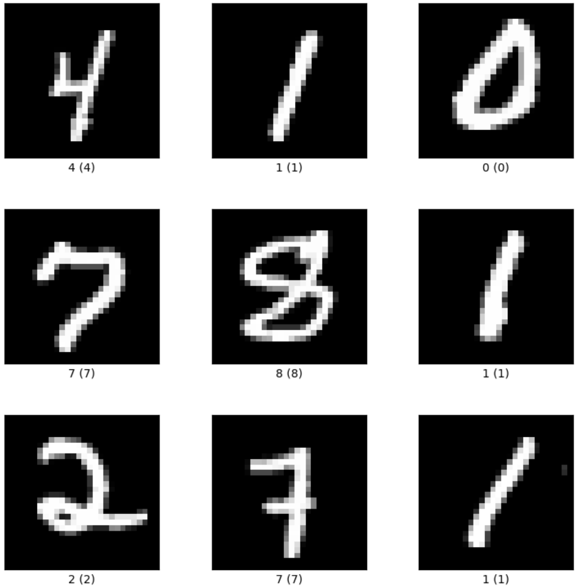
- Assistant Professor, Department of AI, Korea University
  - B.S. (2014) in EE and Math at KAIST
  - Ph.D. (2020) in EE at KAIST
- Research interest: “mathematical” machine learning problems
  - Developing algorithms with provable guarantees
    - e.g., on the correctness of automatic differentiation, efficient computation of gradient in neural networks
  - Analyzing expressivity, optimization properties, and generalization properties of machine learning models
    - e.g., universal approximation, memorization, generalization properties of SGD

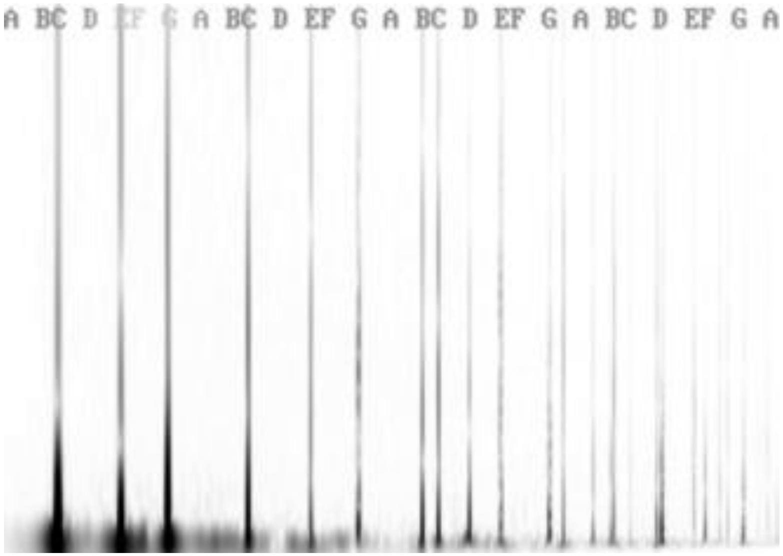
What we will study in this course

- **Course title:** Big Data Analysis Application and Practice
  - **“Analysis”**: dimensionality reduction (and data visualization) methods
    - Goal: embed high-dimensional data in a space with a small dimension
    - From classical methods (e.g., PCA, LDA) to modern ones (e.g., t-SNE, UMAP)
  - **“Big Data”**: fast algorithms
    - e.g., algorithms with running time at most proportional to #data points
  - **“Application and Practice”**: practice sessions
    - In these sessions, you will **implement learned algorithms by yourself**
- **Grading**
  - 6 practice sessions (60%, 10% for each), 2 exams (40%, 20% for each)

Motivation

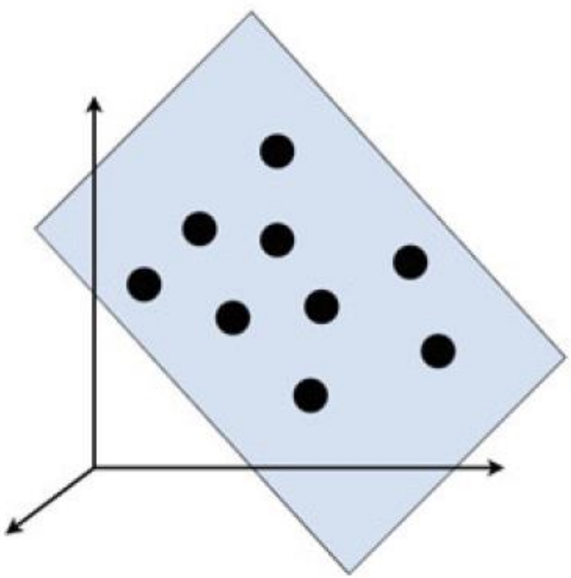
- Each feature of a data point does not carry an equal amount of information
  - e.g., background pixels in image, Fourier features in audio

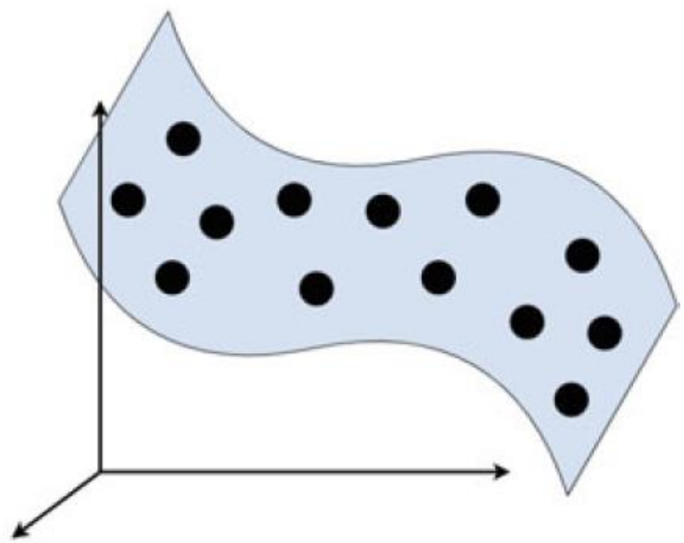




Motivation

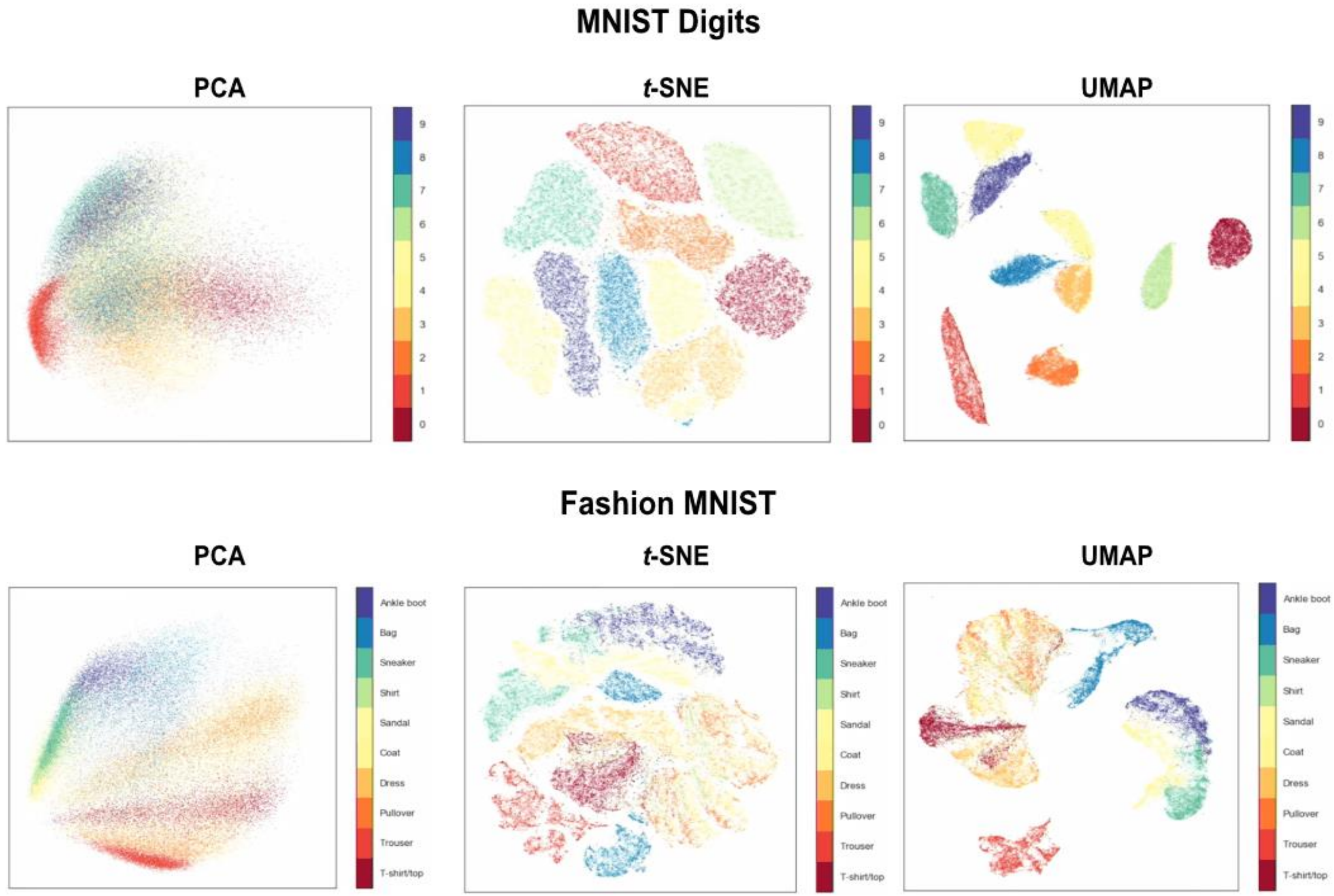
- Manifold hypothesis
  - The *d*-dimensional data points of a dataset usually do not cover the entire space, but they lie on a **specific lower-dimensional structure**





Why dimensionality reduction

- **Data visualization**
  - We often want to “see” the distribution of the data points



Why dimensionality reduction

- **Noise reduction for supervised learning**
  - Smaller input dimension can lead us to better generalization
  - Namely, dimensionality reduction can relax the “overfitting”
    - n: #training data
    - d: dimension of each data
    - p: the regression function is p-times differentiable

Optimal convergence rate of regression $\lesssim n^{-p/(2p+d)}$

Schedule

week	period	freq.	studying contents
1	03.02 - 03.08	1	Introduction
2	03.09 - 03.15	1	Classical data visualization algorithms
3	03.16 - 03.22	1	Practice session
4	03.23 - 03.29	1	Classical data visualization algorithms
5	03.30 - 04.05	1	Practice session
6	04.06 - 04.12	1	Classical data visualization algorithms
7	04.13 - 04.19	1	Practice session
8	04.20 - 04.26	1	Mid-term exam
9	04.27 - 05.03	1	Modern data visualization algorithms
10	05.04 - 05.10	1	Practice session
11	05.11 - 05.17	1	No class
12	05.18 - 05.24	1	Modern data visualization algorithms
13	05.25 - 05.31	1	Practice session
14	06.01 - 06.07	1	Modern data visualization algorithms
15	06.08 - 06.14	1	Practice session
16	06.15 - 06.21	1	Final exam

Warning

- This course will be very mathy
  - We will observe the mathematical objective of each algorithm
  - We will often analyze the computational complexity of each algorithm
  - You should be able to read/write mathematical statements
- I am assuming you
  - Know math including calculus, optimization, linear algebra, algorithm, …
  - Can use python, matlab and can google so that you can implement algorithms learned in the class
    - Sample problem: implement the singular value decomposition and PCA
- You should bring your own laptop for each practice session