Machine Learning in High Dimension
IA317

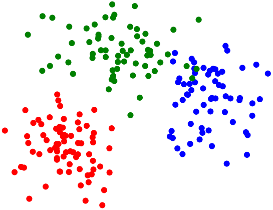
Thomas Bonald & Charlotte Laclau

2023 – 2024

Machine learning

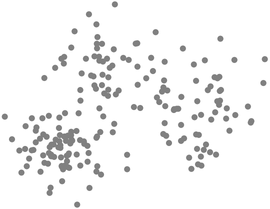
Supervised learning

- Classification
- Regression



Unsupervised learning

- Similarity
- Clustering
- Anomaly detection



High dimension

Data $=$ n samples, each with d features

$$X \in \mathbb{R}^{n \times d}$$

High dimension $d >> 1$ (possibly larger than n)

Examples

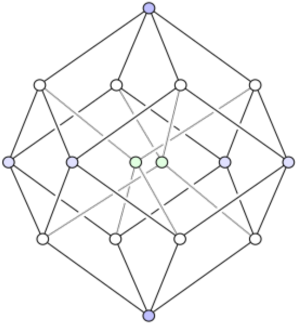
- Textual data (bags of words)
- Medical data
- Customer data

Data might be **heterogeneous**
(e.g., mix of numerical features and categorical features).

Curse of dimensionality

In high dimension:

- samples tend to be **isolated**
- distances tend to be **equal**
- computations are **expensive**



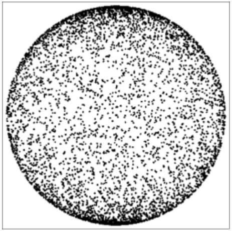
Example (numerical features)

- For $X, Y \sim \mathcal{N}(0, I_d)$,

$$||X - Y||^2 \sim 2\chi^2(d)$$

- Pairwise distance $D = ||X - Y|| \sim \sqrt{2}\chi(d)$:

$$\mathrm{E}(D) = \sqrt{2}\frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})} = O(\sqrt{d}) \quad \mathrm{var}(D) = k - \mathrm{E}(D)^2 = O(1)$$



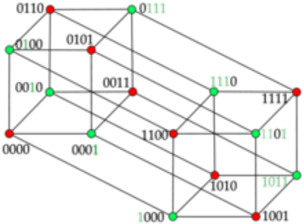
Example (binary features)

- For $X, Y \sim \mathcal{U}(\{0,1\}^d)$ and the Hamming distance,

$$d(X,Y) \sim \mathcal{B}(d, \frac{1}{2}) \approx \mathcal{N}(\frac{d}{2}, \frac{d}{4}) \quad \text{when } d \to +\infty$$

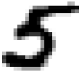
- Pairwise distance $D = ||X - Y||$:

$$\mathrm{E}(D) = O(\sqrt{d}) \quad \mathrm{var}(D) = O(1)$$



Real data: MNIST

$X \in \{0, \ldots, 255\}^{n \times d}$
$n = 10,000$ samples
$d = 28 \times 28 = 784$



Classification by **nearest neighbors** → accuracy ≈ 92%

Outline

1. **Nearest neighbors**
2. **Locally sensitive hashing**
3. **Dimension reduction**
4. **Ensemble methods**
5. **Naive Bayes***
6. **Sparse regression***
7. **Anomaly detection**

Each block = 1 lecture + 1 lab (2 graded *)

Not covered:

- Deep learning (see IA307)
- NLP (see IA312)
- Kernel methods (see IA326)
- Graph methods (see SD212)

Information & Evaluation

- **Moodle**
  `https://moodle.r2.enst.fr/`
  For general information, slides, notebooks, etc.
- **Attendance**
  Presence to the labs is mandatory
  A single absence over the 7 labs is tolerated
- **Evaluation**
  2 graded labs (20% each)
  Final quiz (60%)