Course 3: Unsupervised Learning

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Summary

Last session

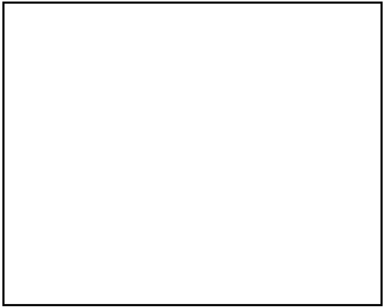
1. Supervised learning - learning from labeled examples
2. Bias/variance tradeoff
3. Overfitting and cross-validation
4. VC Dimension and curse of dimensionality

Today's session

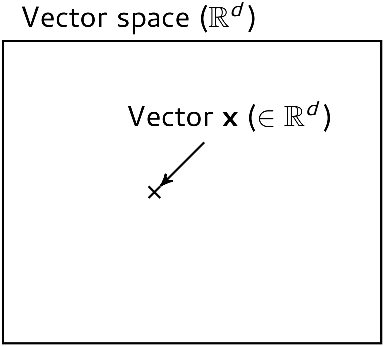
1. Learning from Unlabeled examples
2. Clustering, decomposition and dimensionality reduction

Notations

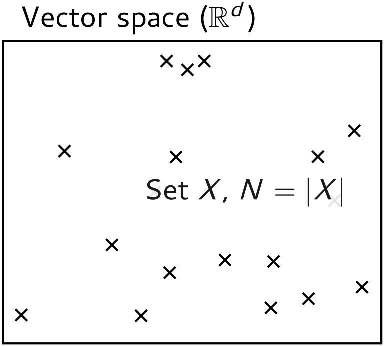
Vector space ($\mathbb{R}^d$)



Notations



Notations



Goal

Discover patterns/structure in X,

Unsupervised learning

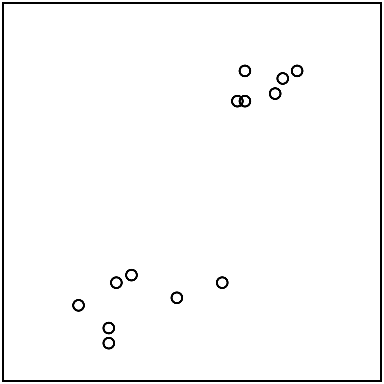
- Unsupervised = no expert, no labels,
- Two main approaches:
  - Clustering = find a partition of X in K subsets,
  - Decomposition using K vectors.
- Applications :
  - Quantization,
  - Visualization. . .

Goal

Discover patterns/structure in X,

Unsupervised learning

- Unsupervised = no expert, no labels,
- Two main approaches:
  - Clustering = find a partition of X in K subsets,
  - Decomposition using K vectors.
- Applications :
  - Quantization,
  - Visualization. . .

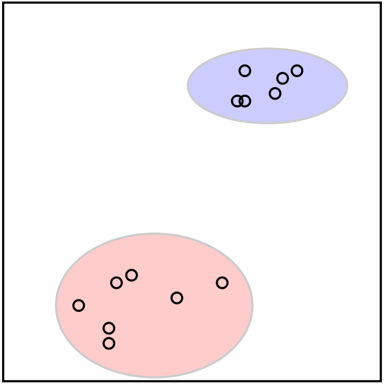


Goal

Discover patterns/structure in X,

Unsupervised learning

- Unsupervised = no expert, no labels,
- Two main approaches:
  - Clustering = find a partition of X in K subsets,
  - Decomposition using K vectors.
- Applications :
  - Quantization,
  - Visualization...



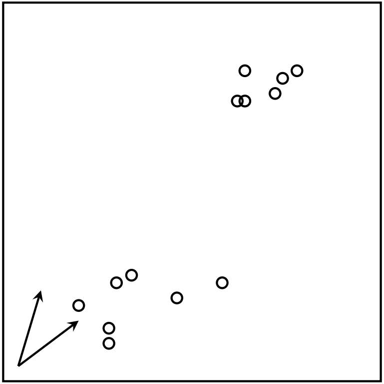
Unsupervised learning

Goal

Discover patterns/structure in X,

Unsupervised learning

- Unsupervised = no expert, no labels,
- Two main approaches:
  - Clustering = find a partition of X in K subsets,
  - Decomposition using K vectors.
- Applications :
  - Quantization,
  - Visualization...



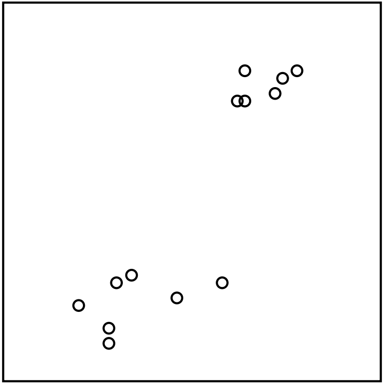
Unsupervised learning

Goal

Discover patterns/structure in X,

Unsupervised learning

- Unsupervised = no expert, no labels,
- Two main approaches:
  - Clustering = find a partition of X in K subsets,
  - Decomposition using K vectors.
- Applications :
  - Quantization,
  - Visualization...



Example: clustering using L_2 norm (1/6)

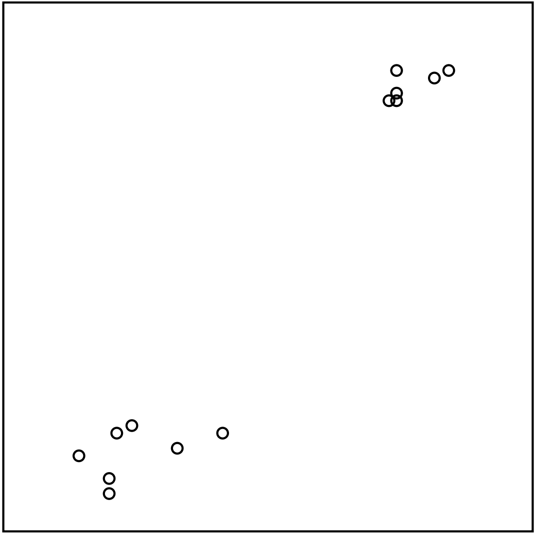
An example to perform clustering is to rely on distances to centroids. We define K *cluster centroids* $\Omega_k, \forall k \in [1..K]$

Definitions

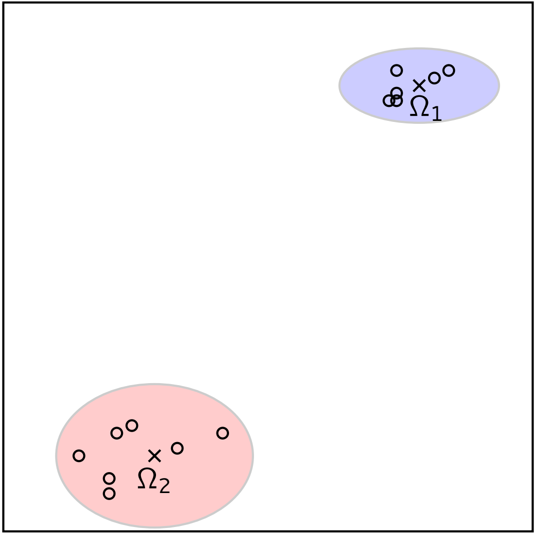
We denote $q : \mathbb{R}^d \to [1..K]$ a function that associates a vector $\mathbf{x}$ with the index of (one of) its closest centroid $q(\mathbf{x})$. Formally:

- $\forall k \in [1..K], \Omega_k \in \mathbb{R}^d$
- $\forall \mathbf{x} \in X, \forall j \in [1..K], \|\mathbf{x} - \Omega_{q(\mathbf{x})}\|_2 \leq \|\mathbf{x} - \Omega_j\|_2$
- Error $E(q) \triangleq \sum_{\mathbf{x} \in X} \|\mathbf{x} - \Omega_{q(\mathbf{x})}\|_2$
- $X = \bigcup_k \underbrace{\{\mathbf{x} \in X, q(\mathbf{x}) = k\}}_{\text{cluster } k}$

Example: clustering using L_2 norm (2/6)



Example: clustering using L_2 norm (2/6)



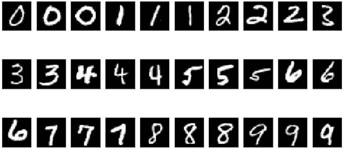
Clustering using L_2 norm (3/6)

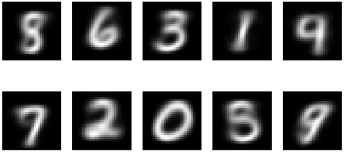
MNIST Dataset

- "Toy" dataset (=small and easy)
- 60000 + 10000 handwritten digits

Clustering MNIST

Using K-means algorithm with $K = 10$





Clustering using L_2 norm (4/6)

Quantizing MNIST

- Replace $\mathbf{x}$ by $\Omega_{k(\mathbf{x})}$
- Compression factor $\kappa = 1 - K/N$

Clustering using L_2 norm (5/6)

Optimal clustering

- Define $E_{opt_K}(q^*) \triangleq \arg \min_{q:\mathbb{R}^d \to [1..K]} E(q)$,
- Finding an optimal clustering is an NP-hard problem.

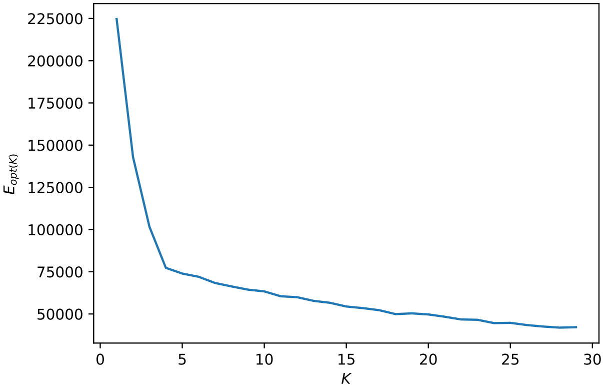
Properties

- $0 = E_{opt_N}(q^*) \leq E_{opt_{N-1}}(q^*) \leq \cdots \leq E_{opt_1}(q^*) = var(X)$,
  - Proof: monotonicity by particularization, extremes with identity function (left) and variance (right).
- $0 \leq \kappa \leq \frac{N-1}{N}$.

Clustering using L_2 norm (6/6)

Choosing K

- Finding a compromise between error and compression,
- Simple practical method : "elbow".

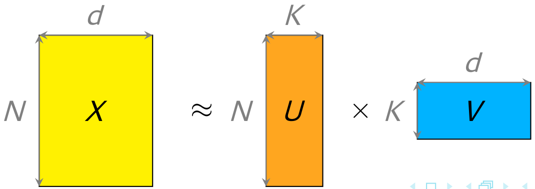


Example 2: Sparse Dictionary Learning (1/4)

Definitions

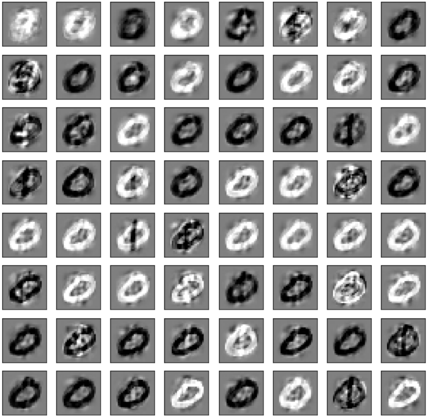
Dictionary learning solves the following matrix factorization problem:

- The set X is considered as a matrix $X \in \mathcal{M}_{N \times d}(\mathbb{R})$,
- We consider decompositions using a dictionary $V \in \mathcal{M}_{K \times d}(\mathbb{R})$ and a code $U \in \mathcal{M}_{N \times k}(\mathbb{R})$, with the lines of V being with norm 1,
- Error $E(U, V) \triangleq \|X - UV\|_2 + \alpha \|U\|_1$
- Training: find U^*, V^* that minimizes $E(U^*, V^*)$
- α is a sparsity control parameter that enforces codes with soft (ℓ_1) sparsity



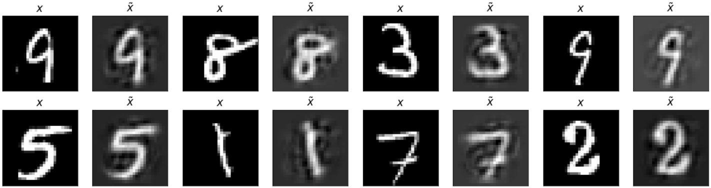
Example: Sparse Dictionary Learning (2/4)

Learning a dictionary on MNIST with $K = 64$



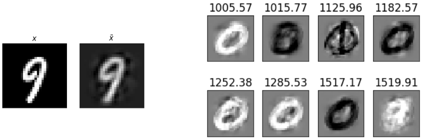
Example 2: Sparse Dictionary Learning (3/4)

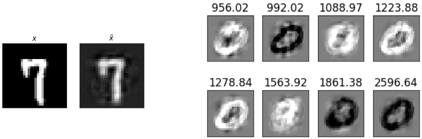
Reconstruction $\tilde{\mathbf{x}} = UV$ of $\mathbf{x}$



8

atoms with largest absolute values:





Example 2: Sparse Dictionary Learning (4/4)

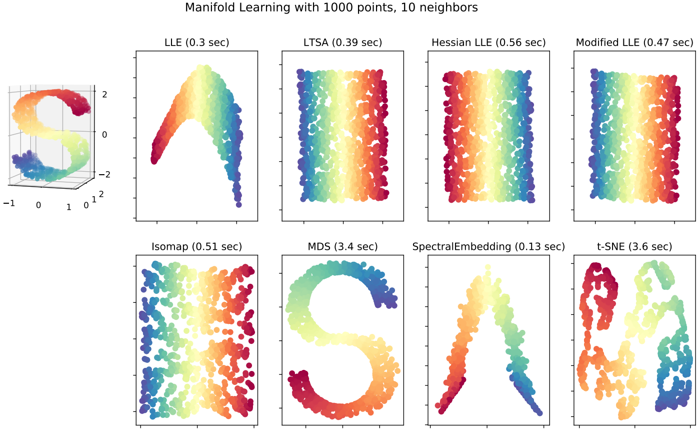
Optimal error

- $E_{opt_K}(U^*, V^*) \triangleq \arg\min_{U,V} E(U, V)$.

Some results

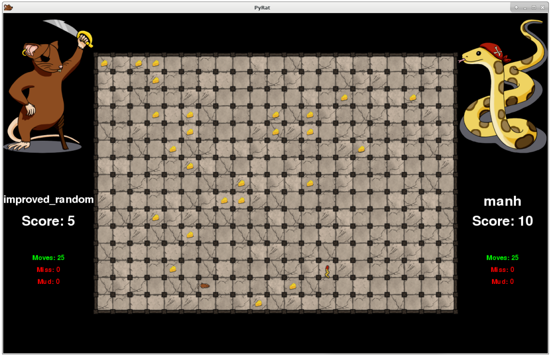
- For $\alpha = 0$ and $K \geq d$, $E_{opt_d}(U^*, V^*) = 0$,
  - One can choose any completion of a basis.
- For $K = N$, $\forall\alpha$, $E_{opt_K}(U^*, V^*) = \alpha N$,
  - If vectors of X are with norm 1, one can choose $V = X$ and $U = \mathbf{I}_N$.

Example 3: Manifold Learning



Approaches to uncover lower dimensional structure of high dimensional data. Source : Manifold module, sklearn website

Non-symmetric PyRat without walls / mud



Can you find patterns in Lost and Draw games using Unsupervised learning ?

Lab Session 3 and assignments for Session 5

TP Unsupervised Learning (TP2)

- K-means, Dictionary Learning and Manifold Learning
- Application on Digits and PyRat

Project 2 (P2)

You will choose an unsupervised learning method. You have to prepare a Jupyter Notebook on this method, including:

- A brief description of the theory behind the method,
- Advanced tests and analysis on your own PyRat Datasets.

During Session 5 (November 21st) you will have 7 minutes to present your notebook.