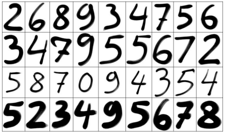
MLDM project
Handwritten Digits Recognition by Nearest-Neighbor Classification

Marc Sebban

Laboratoire Hubert Curien, UMR CNRS 5516
University of Jean Monnet Saint-Étienne (France)

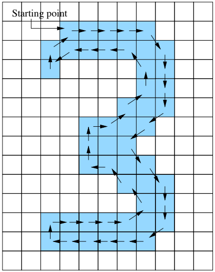


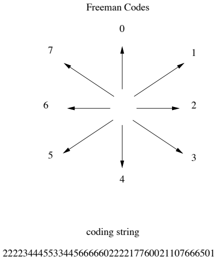
The MLDM project in a nutshell

- **Duration**: 3 months 1/2 by groups of 3 students.
- **# ECTS=6** ≈ **120 hours** that is about 8 hours per week and per student involved in the project.
- **Goal**: Develop/optimize a plateform of handwritten digit recognition and extract knoweldge
- **Algorithms**:
  - Nearest neighbor algorithm.
  - Sequence mining algorihm.
  - Metric Learning algorithm LMNN (Large Margin Nearest Neighbor [Weinberger et al. 2006]). Code available in Matlab and Python.
  - Deep Learning (for learning features) + Nearest Neighbor

Training set

- Create a labeled database of handwritten digits drawn in black and white (graphical interface).
- This dataset can be merged with state of the art databases (like MNIST).
- Represent the digits in:
  - a structured way by using Freeman's codes.
  - a numerical way by using deep learning (CNN or auto-encoders).





Classification Algorithm

Implement a Nearest-Neighbor algorithm with the following features:

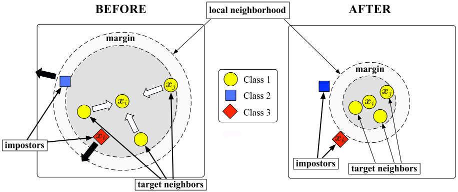
- Use the Edit distance algorithm (for stuctured data) or the Euclidean distance (for numerical data) to compute neighbors.
- Implement different algorithms to reduce the time and storage complexity of NN.
  - Remove outliers.
  - Remove irrelevant training examples.
  - Speed-up the seek of neighbors.
  - Assess the efficiency of these algorithms w.r.t. a baseline.

Improve the numerical representation by metric learning

LMNN applied on the numerical representation

Define constraints tailored to k-NN in a local way: the k nearest neighbors should be of same class ("target neighbors"), while examples of different classes should be kept away ("impostors"):

$$\begin{aligned} \mathcal{S} &= \{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j \text{ and } \mathbf{x}_j \text{ belongs to the } k\text{-neighborhood of } \mathbf{x}_i\}, \\ \mathcal{R} &= \{(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) : (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S}, y_i \neq y_k\}. \end{aligned}$$



The Mahalanobis distance

$\forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, the Mahalanobis distance is defined as follows:

$$d_{\mathbf{M}}(\mathbf{x}, \mathbf{x}') = \sqrt{(\mathbf{x} - \mathbf{x}')^T \mathbf{M} (\mathbf{x} - \mathbf{x}')},$$

where $\mathbf{M} \in \mathbb{R}^{d \times d}$ is a symmetric PSD matrix ($\mathbf{M} \succeq 0$).

Hard Formulation

$$\begin{aligned} \min_{\mathbf{M} \succeq 0} \quad & \sum_{(\mathbf{x_i}, \mathbf{x_j}) \in \mathcal{S}} d^2_{\mathbf{M}}(\mathbf{x_i}, \mathbf{x_j}) \\ \text{s.t.} \quad & d^2_{\mathbf{M}}(\mathbf{x_i}, \mathbf{x_k}) - d^2_{\mathbf{M}}(\mathbf{x_i}, \mathbf{x_j}) \geq 1 \quad \forall (\mathbf{x_i}, \mathbf{x_j}, \mathbf{x_k}) \in \mathcal{R}. \end{aligned}$$

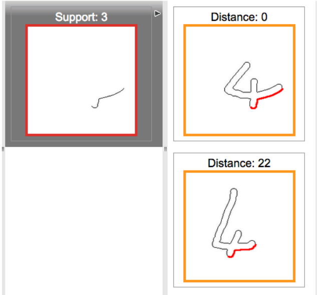
`https://pypi.python.org/pypi/metric-learn`

Sequence Mining Algorithm

- Extract pieces of digits that are representative of each class (from 0 to 9).
- Use a frequent sequence mining algorithm

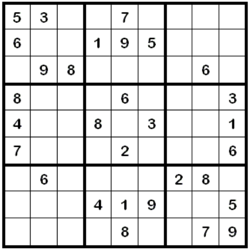
References: `http://www.philippe-fournier-viger.com/spmf/`

`http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.332.4745&rep=rep1&type=pdf`)



Subsidiary task

Implement a game that you can play with your recognition system



5	3			7				
6			1	9	5			
	9	8					6	
8				6				3
4			8		3			1
7				2				6
	6					2	8	
			4	1	9			5
				8			7	9

$$\begin{array}{r} 58 \\ +\ 26 \\ \hline \end{array}$$

Key dates

Key Dates

- **January 11th midnight, 2019**: send the project to the following address: `marc.sebban@univ-st-etienne.fr`. The archive will contain:
    - the code of the platform
    - a report (Latex) written in the form of a 8 pages scientific paper, thus with a title, and presenting the work in an abstract, explaining the aim and the contribution of the paper, the experimental setup, the results and with a conclusion. (see https://2017.icml.cc/Conferences/2017/StyleAuthorInstructions).
- **January 15th, 2019**: Defense of the project. An oral presentation of the project and an on-line demo of the platform will be required.