UNIVERSITÉ DE GRENOBLE

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Informatique et Mathématiques appliquées**

Arrêté ministériel : 7 août 2006

Présentée par
Cao Tri DO

Thèse dirigée par **Ahlame DOUZAL-CHOUAKRIA**,
codirigée par **Michèle ROMBAUT** et
co-encadré par **Sylvain MARIÉ**

préparée au sein du
Laboratoire d'Informatique de Grenoble (LIG)
dans **l'école doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique (MSTII)**

Metric Learning for Time Series Analysis

Thèse soutenue publiquement le **date de soutenance**,
devant le jury composé de:

Patrick GALLINARI
Laboratoire LIP6, Président du jury
Stéphane CANU
Laboratoire LITIS, Rapporteur
Marc SEBBAN
Laboratoire LAHC, Rapporteur
Gustavo CAMPS-VALLS
Laboratoire IPL, Examinateur
Prénom NOM
Labo de bidule, Examinateur
Ahlame DOUZAL-CHOUAKRIA
Laboratoire LIG, Directeur de thèse
Michèle ROMBAUT
Laboratoire GIPSA-Lab, Co-Directeur de thèse

UNIVERSITÉ DE GRENOBLE

ÉCOLE DOCTORALE MSTII

Description de complète de l'école doctorale

T H È S E

pour obtenir le titre de

docteur en sciences

de l'Université de Grenoble-Alpes

Mention : Informatique et Mathématiques appliquées

Présentée et soutenue par

Cao Tri DO

Metric Learning for Time Series Analysis

Thèse dirigée par Ahlame DOUZAL-CHOUAKRIA

préparée au Laboratoire d'Informatique de Grenoble (LIG)

soutenue le date de soutenance

Jury :

Rapporteurs :	Stéphane CANU	-	Laboratoire LITIS
	Marc SEBBAN	-	Laboratoire LAHC
Directeur :	Ahlame DOUZAL-CHOUAKRIA	-	Laboratoire LIG
Co-Directeur :	Michèle ROMBAUT	-	Laboratoire GIPSA-Lab
Encadrant :	Sylvain MARIÉ	-	Schneider Electric
Président :	Patrick GALLINARI	-	Laboratoire LIP6
Examinateur :	Gustavo CAMPS-VALLS	-	Laboratoire IPL
	Prénom NOM	-	Labo de bidule

Todo list

Acknowledgements

I would like to thanks:

- my directors
- my GIPSA collegues
- my AMA collegues
- my Schneider collegues
- my parents

Contents

List of Figures

List of Tables

Table of Acronyms

LIG	*Laboratoire d'Informatique de Grenoble*
AMA	*Apprentissage, Méthode et Algorithme*
GIPSA-Lab	*Grenoble Images Parole Signal Automatique Laboratoire*
AGPiG	*Architecture, Géométrie, Perception, Images, Gestes*
A4S	*Analytic for Solutions*
k**-NN**	k*-nearest neighbors*
SVM	*Support Vector Machines*
SVR	*Support Vector Regression*
d_E	*Euclidean distance*
corr	*Pearson correlation*
cort	*Temporal correlation*
dtw	*Dynamic Time Warping*
IoT	*Internet of Things*
Acc	*Classification accuracy*
Err	*Classification error rate*
MAE	*Mean Absolute Error*
RMSE	*Root Mean Square Error*
FAQ	Frequently Asked Questions / Foire Aux Questions

Introduction

Motivation

- Qu'est-ce qu'une série temporelle ? (réponse d'un système dynamique complexe (= pas de modèle du système)
- Motiver l'intérêt des séries temporelles dans les applications aujourd'hui: données de plus en plus présentes dans de nombreux domaines divers et variés
- Les séries temporelles sont impliquées dans des problèmes de classification, régression et clustering
- Pourquoi sont-elles challenging ? (délais, dynamique)
- On fait face à la fois, à un problème de small et big data

Problem statement (with words)

- Dans de nombreux algorithmes de classification ou de régression (kNN, SVM), la comparaison des individus (séries temporelles) reposent sur une notion de distance entre individus (séries temporelles).
- Contrairement aux données statiques, les données temporelles peuvent être comparés sur la base de plusieurs modalités (valeurs, forme, distance entre spectre, etc.) et à différentes échelles. La « métrique idéale », càd, celle qui permettra de résoudre au mieux le problème de classification/régression peut donc impliquer plusieurs modalités.
- Objectif de notre travail : Apprendre une métrique adéquate tenant compte de plusieurs modalités et de plusieurs échelles en vue d'une classification/régression kNN

PhD contributions

- Définition d'un nouvel espace de représentation: la représentation par paires
- Apprentissage d'une métrique multimodale et multi-échelle en vue d'une classification kNN à vaste marge de séries temporelles monovariées.
- Extension/Transposition du problème d'apprentissage de métrique (Metric Learning) dans l'espace des paires

- Comparaison de la méthode proposée avec des métriques classiques sur un vaste jeu de données (30 bases) de la littérature dans le cadre de la classification univariée de séries temporelles
- Extension du framework d'apprentissage de métrique au problème de régression de séries temporelles univariés
- Extension du framework d'apprentissage de métrique au problème de classification/régression de séries temporelles multivariés.
- Donner une solution interprétable.
- Donner un algorithme à la fois pour les small et big data.

Organisation du manuscrit

Présenter les différents chapitres

Note pour Ahlame, Michèle et Sylvain: Pour ajouter des commentaires dans le fichier .TEX, merci de les ajouter sous cette forme:

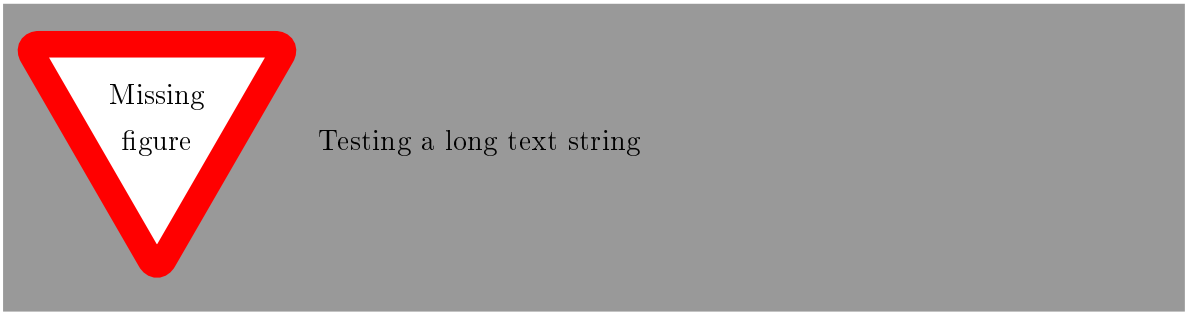
- dans le texte :

  Comment [CTD1]: Initial in [CAO] then your comment in the bracket

- dans la marge :

Comment [CTD2]: Initial in [CAO] then your comment in the bracket

If you think that they are missing figures, you can add them with a description with this command line :



Notations

$\mathbf{x}_i$	a time series
y_i	a label (discrete or continous)
$\mathbf{X} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$	a set of $n \in \mathbb{N}$ labeled time series
d_E	Euclidean distance
L_q	Minkovski q-norm
$\|\|\mathbf{x}\|\|_q$	q-norm of the vector $\mathbf{x}$
d_A	Value-based distance
$corr$	Pearson correlation
$cort$	Temporal correlation
d_F	Euclidean distance between the Fourier spectrum
D	Distance
$\mathbf{x}_{ij}$	a pair of time series $\mathbf{x}_i$ and $\mathbf{x}_j$
y_{ij}	the pairwise label of $\mathbf{x}_{ij}$
t	time stamp/index with $t = 1, ..., T$
T	length of the time series (supposed fixed)
f	frequential index
F	length of the Fourier transform
ξ	Relaxation term
p	number of metric measure considered in the metric learning process
r	order of the temporal correlation
k	number of nearest neighbors
$K(\mathbf{x}_i, \mathbf{x}_j)$	Kernel function between $\mathbf{x}_i$ and $\mathbf{x}_j$
$\phi(\mathbf{x}_i)$	embedding function from the original space to the Hilbert space
C	Hyper-parameter of the SVM (trade-off)
α	
λ	

Part I

Work positioning

The first part of the manuscript aims at positioning the work context. Our objective is the classification and regression of time series. The first chapter presents classic machine learning technics for static data. In particular, we focus on k-Nearest Neighbors classification and Support Vector Machine approach. In the second chapter, we recall metrics used in the literature to compare time series and present the concept of metric learning.

Chapter 1

Related work

Sommaire

In this chapter, we recall some concepts of machine learning. First, we review the principle, the learning framework and the evaluation protocol in supervised learning. Then, we present the algorithms used in our work: k-Nearest Neighbors (k-NN) and Support Vector Machine (SVM).

1.1 Classification, Regression

In this section, we review some terminology in machine learning. First, we recall the principle of machine learning. Then, we detail how to design a framework for supervised learning. After that, we present model evaluation. Finally, we review data normalization.

1.1.1 Machine learning principle

The idea of machine learning (also refer as Pattern Learning or Pattern Recognition) is to imitate with algorithms executed on computers, the ability of living beings to learn from examples. For instance, to teach a child how to read letters, we show him during a training phase, labeled examples of letters ('A', 'B', 'C', etc.) written in different styles and fonts. We don't give him a complete and analytic description of the topology of the characters but

labeled examples. Then, during a testing phase, we want the child to be able to recognize and to label correctly the letters that have been seen during the training, and also to generalize to new instances [G. 06].

Let $X = \{\mathbf{x}_i, y_i\}_{i=1}^n$ be a training set of n samples $\mathbf{x}_i \in \mathbb{R}^p$ and y_i their corresponding labels. The aim of machine learning is to learn a relation (model) f between the samples $\mathbf{x}_i$ and their labels y_i based on examples. This relationship can include static relationships, correlations, dynamic relationship, etc. After the training phase based on labeled examples $(\mathbf{x}_i, y_i)$, the model f has to be able to generalize on the testing phase, i.e., to give a correct prediction y_j for new instances $\mathbf{x}_j$ that haven't been seen during the training.

When y_i are class labels (e.g., class 'A', 'B', 'C' in the case of child's reading), learning the model f is a classification problem; when y_i is a continuous value (e.g., the energy consumption in a building), learning f is a regression problem. Both problems corresponds to supervised learning as $\mathbf{x}_i$ and y_i are known during the training phase [Bis06]; [G. 06]; [OE73]. For both problems, when a part of the labels y_i are known and an other part of y_i is unknown during training, learning f is a semi-supervised problem . Note that when the labels y_i are totally unknown, learning f refers to a clustering problem (unsupervised learning) [JMF99]; [CHY96], out of the scope of this work.

[biblio semi-supervisé]

1.1.2 Model selection

A key objective of learning algorithms is to build models f with good generalization abilities, i.e., models f that correctly predict the class labels y_j of new unknown samples $\mathbf{x}_j$. Fig. 1.2 shows a general approach for solving machine learning problems. In general, a dataset can be divided into 3 sub-datasets (illustrated in Fig. 1.1):

- A **training set** $X = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ consisting of n samples $\mathbf{x}_i$ whose labels y_i are known. The training set is used to build the supervised model f. When the learning algorithm needs to tune hyper-parameters, the training set X is divided into two subsets :
  - A **learning set** which is used to build the supervised model f for each value of the hyper-parameter.
  - A **validation set** which is used to evaluate the supervised model f for each value of the hyper-parameter. The model f with the lowest error on the validation set is kept.
- A **test set** $X_{Test} = \{(\mathbf{x}_j, y_j)\}_{j=1}^m$, which consists of m samples $\mathbf{x}_j$ whose labels y_j are also known but the model f is applied to predict the label $\hat{y_j}$ of samples $\mathbf{x}_j$. The test is used to evaluate the performance of the learnt model between $\hat{y_j}$ and y_j.
- An **operational set** $X_{op} = \{(\mathbf{x}_l, y_l)\}_{l=1}^L$, which consists of L samples $\mathbf{x}_l$ whose labels y_l are totally unknown. The operational set is in general a new dataset on which the learnt algorithm is applied.

laisser Sylvain choisir le mot

id	Attribute 1	Attribute 2	Attribute 3	True Class
1	Yes	Large	125	No
2	No	Medium	135	Yes
3	No	Small	256	No
4	Yes	Medium	320	No
5	Yes	Small	128	Yes
6	No	Large	852	Yes
7	No	Medium	963	Yes

Learning Set

Validation Set

Training Set

id	Attribute 1	Attribute 2	Attribute 3	True Class	Predicted Class
8	No	Large	566	No	?
9	No	Medium	456	Yes	?
10	Yes	Medium	321	No	?
11	No	Small	243	No	?
12	Yes	Small	863	Yes	?
13	No	Large	213	Yes	?
14	Yes	Large	132	Yes	?

Test Set

id	Attribute 1	Attribute 2	Attribute 3	True Class	Predicted Class
15	Yes	Large	874	?	?
16	No	Medium	541	?	?
17	No	Medium	236	?	?
18	No	Large	652	?	?
19	Yes	Small	324	?	?
20	Yes	Small	214	?	?
21	Yes	Medium	222	?	?

Operational Set

Figure 1.1: Division of a dataset into 3 datasets: training, test and operational.

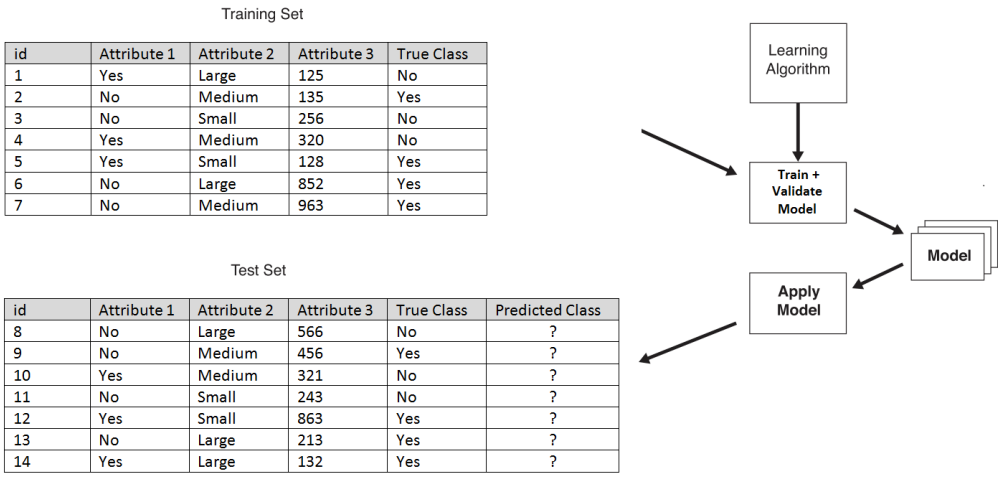


Figure 1.2: General framework for building a supervised (classification/regression) model. Example with 3 features and 2 classes ('Yes' and 'No').

There exists two types of errors committed by a classification or regression model f: training error and generalization error. **Training error** is the error on the training set and **generalization error** is the error on the testing set. A good supervised model f must not

only fit the training data X well, it must also accurately classify records it has never seen before (test set X_{Test}). In other words, a good model f must have low training error as well as low generalization error. This is important because a model that fits the training data too much can have a poorer generalization error than a model with a higher training error. Such a situation is known as model overfitting (Fig. 1.3).

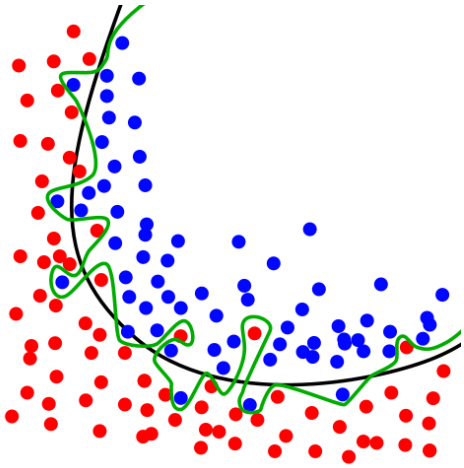


Figure 1.3: An example of overfitting in the case of classification. The objective is to separate blue points from red points. Black line shows a classifier f_1 with low complexity where as green line illustrates a classifier f_2 with high complexity. On training examples (blue and red points), the model f_2 separates all the classes perfectly but may lead to poor generalization on new unseen examples. Model f_1 is often preferred.

In most cases, learning algorithms requires to tune some hyper-parameters. For that, the training set can be divided into 2 sets: a learning and a validation set. Suppose we have two hyper-parameters to tune: C and γ. We make a grid search for each combination (C, γ) of the hyper-parameters, that is in this case a 2-dimensional grid (Fig. 1.4). For each combination (a cell of the grid), the model is learnt on the learning set and evaluated on the validation set. At the end, the model with the lowest error on the validation set is retained. This process is refered as the model selection.

An alternative is cross-validation with v folds, illustrated in Fig. 1.5. In this approach, we partition the training data into v equal-sized subsets. The objective is to evaluate the error for each combination of hyper-parameters. For each run, one fold is chosen for validation, while the $v - 1$ remaining folds are used as the learning set. We repeat the process for each fold, thus v times. Each fold gives one validation error and thus we obtain v errors. The total error for the current combination of hyper-parameters is obtained by summing up the errors for all v folds. When $v = n$, the size of training set, this approach is called leave-one-out or Jackknife. Each test set contains only one sample. The advantage is much data are used as possible for training. Moreover, the validation sets are exclusive and they cover the entire data set. The drawback is that it is computationally expensive to repeat the procedure n times. Furthermore, since each validation set contains only one record, the variance of the estimated performance metric is usually high. This procedure is often used when n, the size training set,

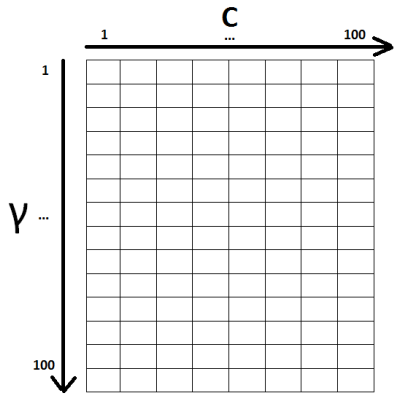


Figure 1.4: Example of a 2 dimensional grid search for parameters C and γ. It defines a grid where each cell of the grid contains a combination (C, γ). Each combination is used to learn the model and is evaluated on the validation set.

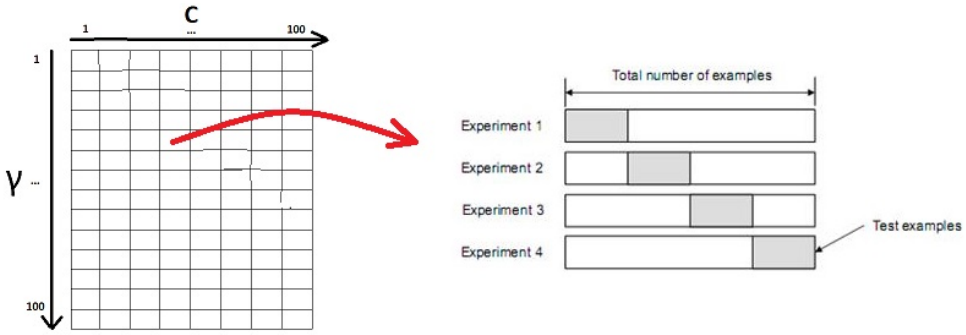


Figure 1.5: v-fold Cross-validation for one combination of parameters. For each of v experiments, use $v-1$ folds for training and a different fold for Testing, then the training error for this combination of parameter is the mean of all testing errors. This procedure is illustrated for $v = 4$.

is small. There exists other methods such as sub-sampling or bootstraps [OE73]; [G. 06]. We only use cross-validation in our experiments.

Comment [AD3]: ne pas dire ici. Michèle dit oui

1.1.3 Model evaluation

1.1.3.a Classification evaluation

The performance of a classification model is based on the counts of test samples $\mathbf{x}_j$ correctly and incorrectly predicted by the model f. These counts are tabulated in a table called the confusion matrix. Table 1.1 illustrates the concept for a binary classification problem. Each cell f_{ij} the table stands for the number of samples from class i predicted to be of class j.

Based on this matrix, the number of correct predictions made by the model is $\sum_{i=1}^{C} f_{ii}$, where C is the number of classes. Equivalently, the ratio of incorrect predictions is $1 - \sum_{i=1}^{C} f_{ii}$.

		Predicted class	
		Class = 1	Class = 0
Actual Class	Class = 1	f_{11}	f_{10}
	Class = 0	f_{01}	f_{00}

Table 1.1: Confusion matrix for a 2-class problem.

To summarize the information, it generally more convenient to use performance metrics such as the classification accuracy (Acc) or error rate (Err). This allows to compare several models with a single number. Note that $Err = 1 - Acc$.

$$Acc = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{\sum_{i=1}^{C} f_{ii}}{\sum_{i,j=1}^{C} f_{ij}} \tag{1.1}$$

$$Err = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}} = \frac{\sum_{i,j=1, i \neq j}^{C} f_{ij}}{\sum_{i,j=1}^{C} f_{ij}} \tag{1.2}$$

Using these performance metrics allows to compare the performance of different classifiers f. It allows to determine in particular whether one learning algorithm outperforms another on a particular learning task on a given test dataset X_{Test}. However, depending on the size of the test dataset, the difference in error rate Err between two classifiers may not be statistically significant. Snedecor & Cochran proposed in 1989 a statistical test based on measuring the difference between two learning algorithms [Coc77]. It has been used by many researchers [Die97]; [DHB95].

Let consider 2 classifiers f_A and f_B. We test these classifiers on the test set X_{Test} and denote p_A and p_B their respective error rates. The intuition of this statistical test is that when algorithm A classifies an example $\mathbf{x}_j$ from the test set X_{Test}, the probability of misclassification is p_A. Thus, the number of misclassification of m test examples is a binomial random variable with mean mp_A and variance $p_A(1 - p_A)m$. The binomial distribution can be approximated by a normal distribution when m has a reasonable value (Law of large numbers). The difference between two independent normally distributed random variables is also normally distributed. Thus, the quantity $p_A - p_B$ is a normally distributed random variable. Under the null hypothesis (the two algorithm should have the same error rate), this will have a mean of zero and a standard error se of:

$$se = \sqrt{\frac{2p(1-p)}{m}} \tag{1.3}$$

where $p = \frac{p_A + p_B}{2}$ is the average of the two error probabilities. From this analysis, we obtain the statistic:

$$z = \frac{p_A - p_B}{\sqrt{2p(1-p)/m}} \tag{1.4}$$

which has (approximatively) a standard normal distribution. We can reject the null hypothesis if $|z| > Z_{0.975} = 1.96$ (for a 2-sided test with probability of incorrectly rejecting the null hypothesis of 0.05).

1.1.3.b Regression evaluation

As the concept of classes is restricted to classification problems, the performance of a regression model f is based on metrics that measure the difference between the predicted label $\hat{y}_j$ and the known label y_j. The Mean Absolute Error function (MAE) computes the mean absolute error, a risk metric corresponding to the expected value of the absolute error loss or L1-norm loss.

$$MAE(\hat{y}, y) = \frac{1}{m} \sum_{j=1}^{m} |\hat{y}_j - y_j| \tag{1.5}$$

A commonly used performance metrics is the Root Mean Squared Error function ($RMSE$) that computes the root of the mean square error, a risk metric corresponding to the expected value of the squared (quadratic) error loss.

$$RMSE(\hat{y}, y) = \sqrt{\frac{1}{m} \sum_{j=1}^{m} (\hat{y}_j - y_j)^2} \tag{1.6}$$

Many works relies on the R^2 function, the coefficient of determination. It provides a measure of how well future samples are likely to be predicted by the model.

$$R^2(\hat{y}, y) = 1 - \frac{\sum_{j=1}^{m} (\hat{y}_j - y_j)^2}{\sum_{j=1}^{m} (\bar{y} - y_j)^2} \tag{1.7}$$

where $\bar{y} = \sum_{j=1}^{m} y_j$ is the mean over the known labels y_j.

1.1.4 Data normalization

Real dataset are often subjected to noise or data scaling. Before applying any learning protocol, it is often necessary to pre-process the data: data scaling, data filtering (e.g., de-noising), outlier removal, etc. We focus on data normalization (data scaling) in our work.

Part 2 of Sarle's Neural Networks FAQ (1997) [1] explains the importance of data normalization for neural network but they can be applied to any learning algorithms. The main advantage of normalization is to avoid attributes in greater numeric ranges to dominate those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation. For example, in the case of Support Vector Machine (SVM), because kernel values usually depend on the inner products of feature vectors, i.e. the linear kernel and the polynomial kernel, large attribute values might cause numerical problems [HCL08].

In most cases, it is recommended to scale each attribute to the range [-1; +1] or [0; 1]. Many normalization methods have been proposed such as Min/Max normalization, Z-normalization or normalization of the log normalization . Let $X = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$ be a training set, $\mathbf{x}_i$ being a sample described by p features $\mathbf{X}_1, \ldots, \mathbf{X}_p$. We define μ_j and σ_j as the mean and the standard deviation of a variable $\mathbf{X}_j$, applying the Z-normalized variable $\mathbf{X}_j^{norm}$ is given by:

références normalization

$$\mathbf{X}_j^{norm} = \frac{\mathbf{X}_j - \mu_j}{\sigma_j} \tag{1.8}$$

Note that the underlying assumption supposes that the variable $\mathbf{X}_j$ is normally distributed: data evolves between $[-\infty; +\infty]$ and are coming from a Gaussian process. In some cases, the data are skewed such as monetary amounts, incomes or distance measures. These data are often log-normally distributed, e.g., the log of the data is normally distributed (Fig. 1.6). The underlying idea is to take the log of the data ($\mathbf{X}_j^{log}$) to restore the symmetry, and then, to apply a Z-normalization of this transformation:

$$\mathbf{X}_j^{log} = \ln(\mathbf{X}_j); \tag{1.9}$$

$$\mathbf{X}_j^{log,norm} = \frac{\mathbf{X}_j^{log} - \mu_j^{log}}{\sigma_j^{log}} \tag{1.10}$$

$$\mathbf{X}_j^{norm} = \exp(\mathbf{X}_j^{log,norm}) \tag{1.11}$$

where ln denotes the Natural Logarithm function, μ_j^{log} and σ_j^{log} the mean and the standard deviation of a variable $\mathbf{X}_j^{log}$.

Finally, we recall some precautions to the practitioner in the learning protocol, experimented by Hsu & al. in the context of SVM [HCL08]. First, training and testing data must be scaled using the same method. Second, training and testing data must not be scaled separately. Third, the whole dataset must not be scaled together at the same time. These often leads to poorer results. A proper way to do normalization is to scale the training data, store the parameters of the normalization (i.e. μ_i and σ_i for Z-normalization), then apply the same normalization to the testing data.

[1] `http://www.faqs.org/faqs/ai-faq/neural-nets/`

[2] source: `http://www.r-statistics.com/2013/05/log-transformations-for-skewed-and-wide-distributions-from-practical-data-science-with-r/`

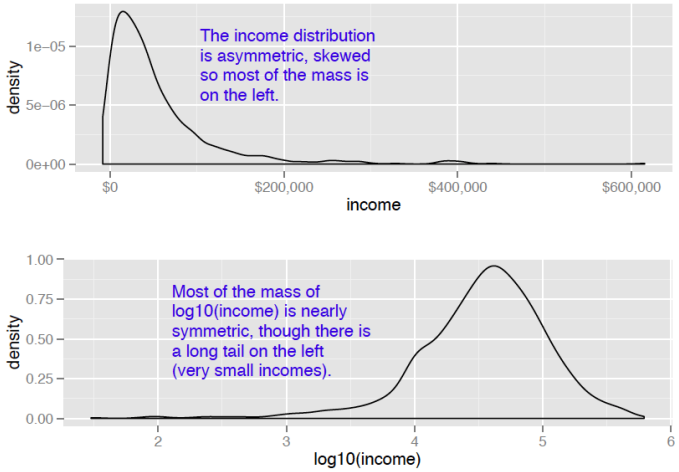


Figure 1.6: A nearly log-normal distribution, and its log transform [2]

1.2 Machine learning algorithms

Many algorithms have been proposed in the context of supervised learning, such as the Deep Neural Network, the Decision Tree or the Relevance Vector Machine (RVM). Our proposition uses Support Vector Machine (SVM) in the context of k-Nearest Neighbors (k-NN) classification. We limit the section to present these two algorithms.

Comment [AD4]: dit d'enlever 1ère phrase mais Michèle dit de garder

1.2.1 k-Nearest Neighbors (k-NN) classifier

A simple approach to classify samples is to consider that "close" samples have a great probability to belong to the same class. Given a test sample $\mathbf{x}_j$, one can decide that $\mathbf{x}_j$ belong to the class y_i of its nearest neighbor $\mathbf{x}_i$ in the training set.

Comment [AR5]: Ahlam trouve que ce n'est pas clair. A refaire

More generally, we can consider the k nearest neighbors of $\mathbf{x}_j$. The class y_j of the test sample $\mathbf{x}_j$ is assigned with a voting scheme among them, i.e., using the majority of the class of nearest neighbors. This algorithm is refer as the k-Nearest Neighbors algorithm (k-NN) [SJ89]; [CH67]. Fig. 1.7 illustrates the concept for a neighborhood of $k = 3$ and $k = 5$.

In the k-NN algorithm, the notion of "closeness" between samples $\mathbf{x}_i$ is based on the computation of a metric [3] D. For static data, usually used metrics are the Euclidean distance, the Minkowski distance or the Mahalanobis distance. Considering a training set X of n samples, solving the 1-NN classification problem is equivalent to solve the optimization problem:
For a new sample $\mathbf{x}_j$, $\forall i \in \{1...n\}$,

$$y_j = y_{i^*} \tag{1.12}$$

[3] A clarification of the terms metric, distance, dissimilarity, etc. will be given in Chapter 2. For now, we refer all of them as metrics.

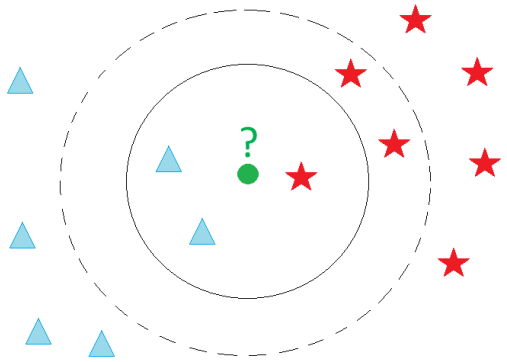


Figure 1.7: Example of k-NN classification. The test sample (green circle) is classified either to the first class (red stars) or to the second class (blue triangles). If $k = 3$ (solid line circle) it is assigned to the second class because there are 2 triangles and only 1 star inside the inner circle. If $k = 5$ (dashed line circle) it is assigned to the first class (3 stars vs. 2 triangles inside the outer circle).

where $i^* = \underset{i \in \{1...n\}}{\text{argmin}} D(\mathbf{x}_i, \mathbf{x}_j)$.

The k-NN algorithm can be extended to estimate continous labels (regression problems). The procedure is similar. The label y_j is defined as :

$$y_j = \frac{1}{k} \sum_{i=1}^{k} y_i \tag{1.13}$$

where i corresponds to the index of the k-nearest neighbors [Alt92]. There exists other variants of the k-NN algorithms. In a weighed k-NN, the approach consists in weighting the k-NN decision by assigning to each neighbor $\mathbf{x}_i$ from an unknown sample $\mathbf{x}_j$, a weight w_i defined as a function of the distance $D(\mathbf{x}_i, \mathbf{x}_j)$ [Dud76]. To cope with uncertainty or imprecision in the labeling of the training data $\mathbf{x}_i$, other authors propose in a fuzzy k-NN to determine the membership degree in each class of an unseen sample $\mathbf{x}_j$ by combining the memberships of its neighbors [KGG85]. Denoeux propose a framework based on Dempster-Shafer theory where the k-NN rule takes into account the non-representativity of training data, the weighting rule and uncertainty in the labeling [Den95].

Comment [AD6]: Expliquer d'avantage

Despite its simplicity, the k-NN algorithm has been shown to be successful on time series classification problems [BMP02]; [Xi+06]; [Din+08]. However, the k-NN algorithm presents some disadvantages, mainly due to its computational complexity, both in space (storage of the training samples $\mathbf{x}_i$) and time (search of the neighbors) [OE73]. Suppose we have n labeled training samples in p dimensions, and find the closest neighbors to a test sample $\mathbf{x}_j$ ($k = 1$). In the most simple approach, we look at each stored samples $\mathbf{x}_i$ ($i = 1...n$) one by one, calculate its metric to $\mathbf{x}_i$ (D($\mathbf{x}_i$,$\mathbf{x}_j$)) and retain the index of the current closest one. For the standard Euclidean distance, each metric computation is $O(p)$ and thus the search is $O(pn)$. Moreover, using standard metrics (such as the Euclidean distance) uses all the p dimensions

in its computation and thus assumes that all dimensions have the same effect on the metric. This assumption may be wrong and can impact the classification performances.

1.2.2 Support Vector Machine (SVM) algorithm

Support Vector Machine (SVM) is a classification method introduced in 1992 by Boser, Guyon, and Vapnik [BGV92]; [CV95] to solve at first linearly separable problems. The SVM classifier have demonstrate high accuracy, ability to deal with high-dimensional data, good generalization properties and interpretation for various applications from recognizing handwritten digits, to face identification, text categorization, bioinformatics and database marketing [Wan02]; [YL99]; [HHP01]; [SSB03]; [CY11]. SVMs belong to the category of kernel methods, algorithms that depends on the data only through dot-products [SS13]. It allows thus to solve non-linear problem. This section gives a brief overview of the mathematical key points and interpretation of the method. For more informations, the reader can consult [SS13]; [CY11]; [CV95].

We first present an intuition of maximum margin concept. We give the primal formulation of the SVM optimization problem. Then, by transforming the latter formulation into its dual form, the kernel trick can be applied to learn non-linear classifiers. Finally, we detail how we can interpret the obtained coefficients and how SVMs can be extended for regression problems.

1.2.2.a Intuition

Let $\{\mathbf{x}_i, y_i\}_{i=1}^n$ be a set of n samples $\mathbf{x}_i \in \mathbb{R}^p$ and their labels $y_i = \pm 1$ (2 class-problem). The objective is to learn a hyperplane, whose equations are $\mathbf{w}^T\mathbf{x} + b = 0$, that can separate samples of class +1 from the ones of class -1. When the problem is linearly separable such as in Fig. 1.8, there exists an infinite number of hyperplanes.

Comment [AD7]: Mettre dans les figures des + et - pour les classes

Vapnik & al. [CV95] propose to choose the separating hyperplane that maximizes the margin, e.g. the hyperplane that leaves as much distance as possible between the hyperplane and the closest samples $\mathbf{x}_i$ of each class, called the support vectors. This distance is equal to $\frac{1}{||\mathbf{w}||_2}$. We denote $||\mathbf{w}||_2$, the L2-norm of the vector $\mathbf{w}$ and $||\mathbf{w}||_1$ the L1-norm of $\mathbf{w}$:

$$||\mathbf{w}||_2 = \sqrt{\mathbf{w}^T\mathbf{w}} = \sqrt{\sum_{h=1}^{p} w_h^2} \tag{1.14}$$

$$||\mathbf{w}||_1 = \sum_{h=1}^{p} |w_h| \tag{1.15}$$

where $\mathbf{w} = [w_1, \ldots, w_p]$ denotes the weight vector.
The hyperplanes passing through the support vectors of each class are referred as the canonical hyperplanes, and the region between the canonical hyperplanes is called the margin band (Fig. 1.9).

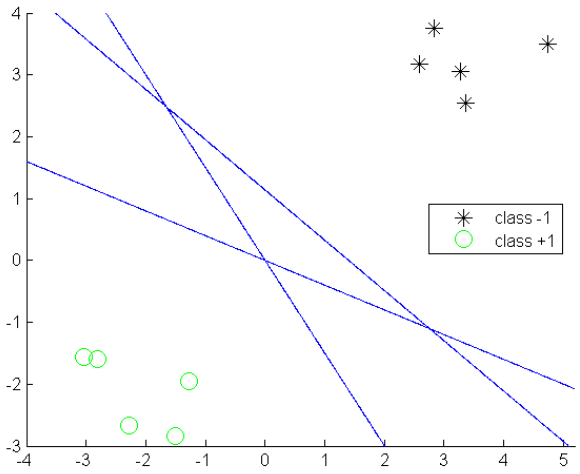


Figure 1.8: Example of linear classifiers in a 2-dimensional plot. For a set of points of classes +1 and -1 that are linearly separable, there exists an infinite number of separating hyperplanes corresponding to $\mathbf{w}^T\mathbf{x} + b = 0$.

1.2.2.b Primal formulation

Finding $\mathbf{w}$ and b by maximizing the margin $\frac{1}{||\mathbf{w}||_2}$ is equivalent to minimizing the norm of $\mathbf{w}$ such that all samples from the training set are correctly classified:

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \frac{1}{2}||\mathbf{w}||_2^2 \tag{1.16}$$

$$\textbf{s.t. } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \tag{1.17}$$

This is a constrained optimization problem in which we minimize an objective function (Eq. 1.16) subject to constraints (Eq. 1.17). This formulation is referred as the primal hard margin problem. When the problem is not linearly separable, slack variables $\xi_i \geq 0$ are introduced to relax the optimization problem:

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \left(\overbrace{\frac{1}{2}||\mathbf{w}||_2^2}^{Regularization} + C \overbrace{\sum_{i=1}^{n} \xi_i(\mathbf{w}; b; x_i; y_i)}^{Loss} \right) \tag{1.18}$$

$$\textbf{s.t. } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \tag{1.19}$$

$$\xi_i \geq 0 \tag{1.20}$$

where $C > 0$ is a penalty hyper-parameter.

This formulation is referred as the primal soft margin problem. It is a quadratic programming optimization problem subjected to constraints. Thus, it is a convex problem: any local

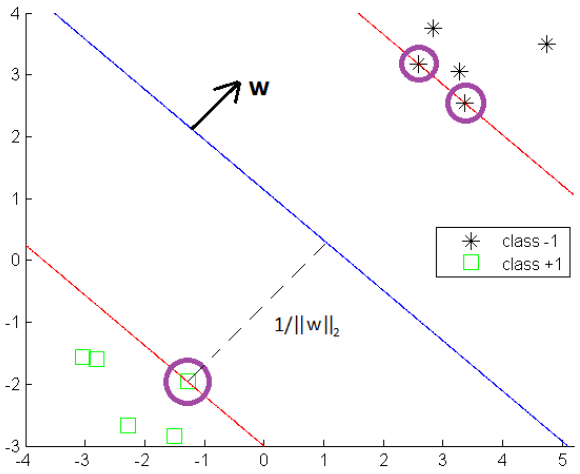


Figure 1.9: The argument inside the decision function of a classifier is $\mathbf{w}^T\mathbf{x}+b$. The separating hyperplane corresponding to $\mathbf{w}^T\mathbf{x}+b=0$ is shown as a line in this 2-dimensional plot. This hyperplane separates the two classes of data with points on one side labeled $y_i=+1$ ($\mathbf{w}^T\mathbf{x}+b\geq 0$) and points on the other side labeled $y_i=-1$ ($\mathbf{w}^T\mathbf{x}+b<0$). Support vectors are circled in purple and lies on the hyperplanes $\mathbf{w}^T\mathbf{x}+b=+1$ and $\mathbf{w}^T\mathbf{x}+b=-1$

solutions is a global solution. The objective function in Eq. 1.18 is made of two terms. The first one, the regularization term, penalizes the complexity of the model and thus, controls the ability of the algorithm to generalize on new samples. The second one, the loss term, is an adaptation term to the data. The hyper-parameter C is a trade-off between the regularization and the loss term. When C tends to $+\infty$, the problem is equivalent to the primal hard margin problem. The hyper-parameter C is learnt during the training phase.

For SVM, the two common loss functions ξ_i are $\max(1-y_i\mathbf{w}^T\mathbf{x}_i,0)$ and $[\max(1-y_i\mathbf{w}^T\mathbf{x}_i,0)]^2$. The former is referred to as L1-Loss and the latter is L2-Loss function. L2-loss function will penalize more slack variables ξ_i during training. Theorically, it should lead to less error in training and poorer generalization in most of the case.

Two common regularizers are $||\mathbf{w}||_1$ and $||\mathbf{w}||_2$. The former is referred to as L1-Regularizer while the latter is L2-Regularizer. L1-Regularizer is used to obtain sparser models than L2-Regularizer. Thus, it can be used for variable selection.

From this, for a binary classification problem, to classify a new sample $\mathbf{x}_j$, the decision function is:

$$f(\mathbf{x}_j)=sign(\mathbf{w}^T\mathbf{x}_j+b) \tag{1.21}$$

1.2.2.c Dual formulation

From the primal formulation, using a L2-Regularizer, it is possible to have an equivalent dual form. This latter formulation allows samples $\mathbf{x}_i$ to appear in the optimization problem through dot-products only. The kernel trick can be applied to extend the methods to learn non-linear classifiers.

First, to simplify the calculation development, let consider the hard margin formulation in Eqs. 1.18, 1.19 and 1.20 with a L1-Loss function. As a constrained optimization problem, the formulation is equivalent to the minimization of a Lagrange function $L(\mathbf{w}, b)$, consisting of the sum of the objective function and the n constraints multiplied by their respective Lagrange multipliers $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^T$:

$$\underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \left(L(\mathbf{w}, b) = \frac{1}{2}(\mathbf{w}^T\mathbf{w}) - \sum_{i=1}^{n} \alpha_i (y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1) \right) \tag{1.22}$$

$$\textbf{s.t. } \forall i = 1...n :$$

$$\alpha_i \geq 0 \tag{1.23}$$

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 \geq 0 \tag{1.24}$$

$$\alpha_i(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1) = 0 \tag{1.25}$$

where $\alpha_i \geq 0$ are the Lagrange multipliers. In optimization theory, Eq. 1.23, 1.24 and 1.25 are called the Karush-Kuhn-Tucker (KKT) conditions. It corresponds to the set of conditions which must be satisfied at the optimum of a constrained optimization problem. The KKT conditions will play an important role in the interpretation of SVM in Section 1.2.2.e.

At the minimum value of $L(\mathbf{w}, b)$, we assume the derivatives with respect to b and $\mathbf{w}$ are set to zero:

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0$$

that leads to:

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{1.26}$$

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \tag{1.27}$$

By substituting $\mathbf{w}$ into $L(\mathbf{w}, b)$ in Eq. 1.22, we obtain the dual formulation (*Wolfe dual*):

$$\underset{\boldsymbol{\alpha}}{\text{argmax}} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i . \mathbf{x}_j) \right) \tag{1.28}$$

$$\textbf{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{1.29}$$

$$\alpha_i \geq 0 \tag{1.30}$$

The dual objective in Eq. 1.28 is quadratic in the parameters α_i. Adding the constraints in Eqs. 1.29 and 1.30, it is a constrained quadratic programming optimization problem (QP). Note that while the primal formulation is minimization, the equivalent dual formulation is maximization. It can be shown that the objective functions of both formulations reach the same value when the solution is found [CY11].
In the same spirit, considering the soft margin primal problem, it can be shown that it leads to the same formulation [CY11] (Eqs. 1.28 and 1.29), except that the Lagrange multipliers α_i are upper bounded by the trade-off C in the soft margin formulation:

$$0 \leq \alpha_i \leq C \tag{1.31}$$

The constraints in Eq. 1.31 are called the Box constraints [CY11]. From the optimal value of α_i, denoted α_i^*, it is possible to compute the weight vector $\mathbf{w}^*$ and the bias b^* at the optimality:

$$\mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i \tag{1.32}$$

$$b^* = \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i - y_i) \tag{1.33}$$

At the optimality point, only a few number of datapoints have $\alpha_i^* > 0$ as shown as in Fig. 1.10. These samples are the vector supports. All other datapoints have $\alpha_i^* = 0$, and the decision function is independent of them. Thus, the representation is sparse.

From this, to classify a new sample $\mathbf{x}_j$, the decision function for a binary classification problem is:

$$f(\mathbf{x}_j) = sign(\sum_{i=1}^{n} \alpha_i^* y_i (\mathbf{x}_i . \mathbf{x}_j) + b^*) \tag{1.34}$$

1.2.2.d Kernel trick

The concept of kernels was introduced by Aizerman & al. in 1964 to design potential functions in the context of pattern recognition [ABR64]. The idea was re-introduced in 1992 by Boser &

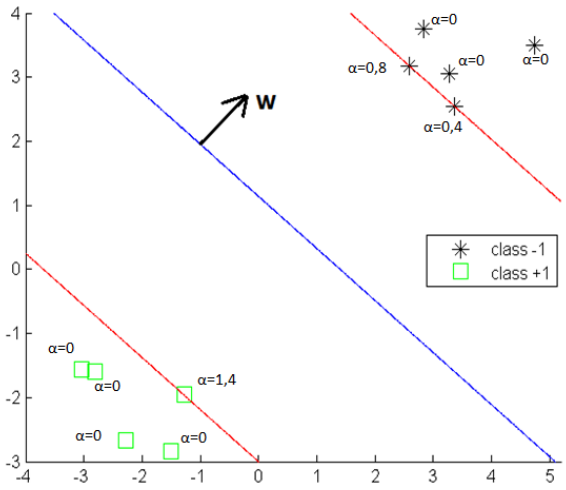


Figure 1.10: Obtained hyperplane after a dual resolution (full blue line). The 2 canonical hyperplanes (dash blue line) contains the support vectors whose $\alpha_i > 0$. Other points have their $\alpha_i = 0$ and the equation of the hyperplane is only affected by the support vectors.

al. for Support Vector Machine (SVM) and has been received a great number of improvements and extensions to symbolic objects such as text or graphs [BGV92].

From the dual objective in Eq. 1.28, we note that the samples $\mathbf{x}_i$ are only involves in a dot-product. Therefore, we can map these samples $\mathbf{x}_i$ into a higher dimensional hyperspace, called the feature space, through the replacement:

$$(\mathbf{x}_i.\mathbf{x}_j) \rightarrow \Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j) \tag{1.35}$$

where Φ is the mapping function. The intuition behind is that for many datasets, it is not possible to find a hyperplan that can separate the two classes in the input space if the problem is not linearly separable. However, by applying a transformation Φ, data might become linearly separable in a higher dimensional space (feature space). Fig. 1.11 illustrates the idea: in the original 2-dimensional space (left), the two classes can't be separated by a line. However, with a third dimension such that the $+1$ labeled points are moved forward and the -1 labeled moved back the two classes become separable.

In most of the case, the mapping function Φ does not need to be known since it will be defined by the choice of a kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j)$. We call Gram matrix G, the matrix containing all $K(\mathbf{x}_i, \mathbf{x}_j)$:

$$G = (K(\mathbf{x}_i, \mathbf{x}_j))_{1 \leq i,j \leq n} = \begin{pmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & \dots & K(\mathbf{x}_1, \mathbf{x}_n) \\ \dots & & \dots \\ K(\mathbf{x}_n, \mathbf{x}_1) & \dots & K(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix}$$

Defining a kernel has to follow rules. One of these rules specifies that the kernel function

has to define a proper inner product in the feature space. Mathematically, the Gram matrix has to be semi-definite positive (Mercer's theorem) [SS13]. These restricted feature spaces, containing an inner product are called Hilbert space.

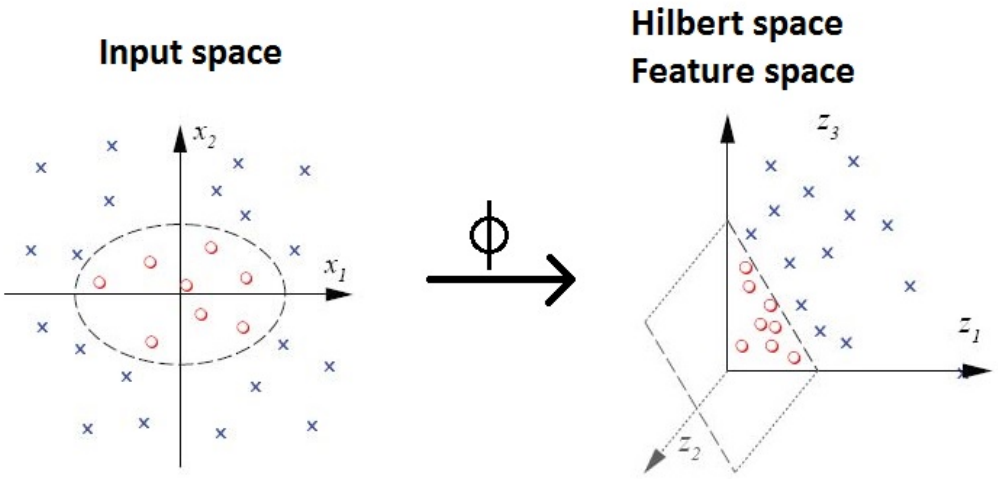


Figure 1.11: Left: in two dimensions the two classes of data (cross and circle) are mixed together, and it is not possible to separate them by a line: the data is not linearly separable. Right: using a Gaussian kernel, these two classes of data become separable by a hyperplane in feature space, which maps to the nonlinear boundary shown, back in input space.[4]

Many kernels have been proposed in the literature such as the polynomial, sigmoid, exponential or wavelet kernels [SS13]. The most popular ones that we will use in our work are respectively the Linear and the Gaussian (or Radial Basis Function (RBF)) kernels:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i.\mathbf{x}_j \tag{1.36}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{||\mathbf{x}_j - \mathbf{x}_i||_2^2}{2\sigma^2}) = \exp(-\gamma||\mathbf{x}_j - \mathbf{x}_i||_2^2) \tag{1.37}$$

where $\gamma = \frac{1}{2\sigma^2}$ is the parameter of the Gaussian kernel and $||\mathbf{x}_j - \mathbf{x}_i||_2$ is the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. Note that the Linear kernel is the identity transformation. In practice, for large scale problem (when the number of dimensions p is high), using a Linear kernel is sufficient [FCH08].

The Gaussian kernel computed between a sample $\mathbf{x}_j$ and a support vector $\mathbf{x}_i$ is an exponentially decaying function in the input space. The maximum value of the kernel ($K(\mathbf{x}_i, \mathbf{x}_j)$=1) is attained at the support vector (when $\mathbf{x}_i = \mathbf{x}_j$). Then, the value of the kernel decreases uniformly in all directions around the support vector, with distance and ranges between zero and one. It can thus be interpreted as a similarity measure. Geometrically speaking, it leads to hyper-spherical contours of the kernel function as shown in Fig. 1.12 [5]. The parameter γ controls the decreasing speed of the sphere. In practice, this parameter is learnt during the

[4] source: http://users.sussex.ac.uk/~christ/crs/ml/lec08a.html

[5] https://www.quora.com/Support-Vector-Machines/What-is-the-intuition-behind-Gaussian-kernel-in-SVM

training phase.

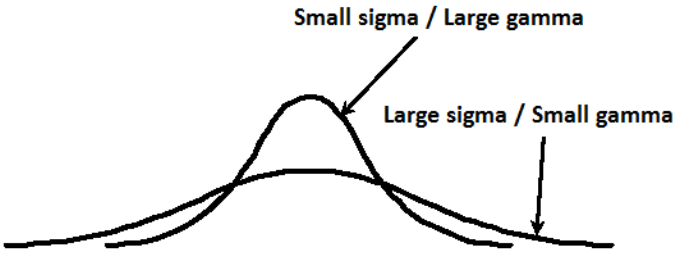


Figure 1.12: Illustration of the Gaussian kernel in the 1-dimensional input space for a small and large γ.

By applying the kernel trick to the soft margin formulation in Eqs. 1.28, 1.29 and 1.31, the following optimization problem allows to learn non-linear classifiers:

$$\underset{\boldsymbol{\alpha}}{\text{argmax}} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \right) \tag{1.38}$$

$$\textbf{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{1.39}$$

$$0 \leq \alpha_i \leq C \tag{1.40}$$

The decision function f becomes:

$$f(\mathbf{x}_j) = sign(\sum_{i=1}^{n} \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}_j) + b^*) \tag{1.41}$$

Note that in this case, we can't recover the weight vector $\mathbf{w}^*$. Let n_{SV} be the number of support vectors ($n_{SV} \leq n$). To recover b^*, we recall that for support vectors $\mathbf{x}_i$:

$$y_j \left(\sum_{i=1}^{n_{SV}} \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}_j) + b^* \right) = 1 \tag{1.42}$$

From this, we can solve b^* using an arbitrarily chosen support vector $\mathbf{x}_i$:

$$b^* = \frac{1}{y_j} - \sum_{i=1}^{n_{SV}} \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}_j) \tag{1.43}$$

1.2.2.e Interpretation

Interpretation in the primal

We recall that $\mathbf{x}_i$ is a sample in p dimensions: $\mathbf{X}_1, \ldots, \mathbf{X}_p$. Geometrically, the vector $\mathbf{w}$ represents the direction of the hyperplane (Fig. 1.13). The bias b is equal to the distance

of the hyperplane to the origin point $\mathbf{x} = \mathbf{0}$[6]. The orthogonal projection of a sample $\mathbf{x}_i$ on the direction $\mathbf{w}$ is $P_{\mathbf{w}}(\mathbf{x}_i) = \mathbf{w}^T\mathbf{x}_i$. In the soft margin problem, the slack variables ξ_i of the samples $\mathbf{x}_i$ that lies within the two canonical hyperplanes are equal to zero. Outside of these canonical hyperplanes, the slack variables $\xi_i > 0$ are equal to the distance to the hyperplane.

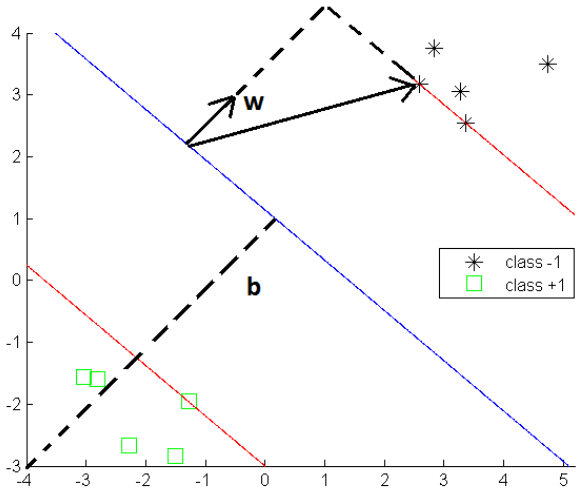


Figure 1.13: Geometric representation of SVM.

In the primal, the weight vector $\mathbf{w} = [w_1, \ldots, w_p]^T$ contains as many elements as there are dimensions in the dataset, i.e., $\mathbf{w} \in \mathbb{R}^p$. The magnitude of each element in $\mathbf{w}$ denotes the importance of the corresponding variable for the classification problem. If the element of $\mathbf{w}$ for some variable is 0, these variables are not used for the classification problem.

In order to visualize the above interpretation of the weight vector $\mathbf{w}$, let us examine several hyperplanes $\mathbf{w}^T\mathbf{x} + b = 0$ shown in Fig. 1.14 with $p = 2$. Fig. 1.14(a) shows a hyperplane where elements of $\mathbf{w}$ are the same for both variables $\mathbf{X}_1$ and $\mathbf{X}_2$. The interpretation is that both variables contribute equally for classification of objects into positive and negative. Fig. 1.14(b) shows a hyperplane where the element of $\mathbf{w}$ for $\mathbf{X}_1$ is 1, while that for $\mathbf{X}_2$ is 0. This is interpreted as that $\mathbf{X}_1$ is important but $\mathbf{X}_2$ is not. An opposite example is shown in Fig. 1.14(c) where $\mathbf{X}_2$ is considered to be important but $\mathbf{X}_1$ is not. Finally, Fig. 1.14(d) provides a 3-dimensional example ($p = 3$) where an element of $\mathbf{w}$ for $\mathbf{X}_3$ is 0 and all other elements are equal to 1. The interpretation is that $\mathbf{X}_1$ and $\mathbf{X}_2$ are important but $\mathbf{X}_3$ is not.

Another way to interpret how much a variable contributes in the vector $\mathbf{w}$ is to express the contribution in percentage. To do that, if the variables $\mathbf{X}_j$ of the time series are normalized before learning the SVM model, they evolves in the same range. Thus, the ratio $\frac{w_j}{||\mathbf{w}||_2}.100$ defines the percentage of contribution for each variable $\mathbf{X}_j$ in the SVM model.

Interpretation in the dual

As a constrained optimization, the dual form satisfies the Karush-Kuhn-Tucker (KKT) condi-

[6] $\mathbf{0}$ stands for the null vector: $\mathbf{0} = [0, \ldots, 0]^T$

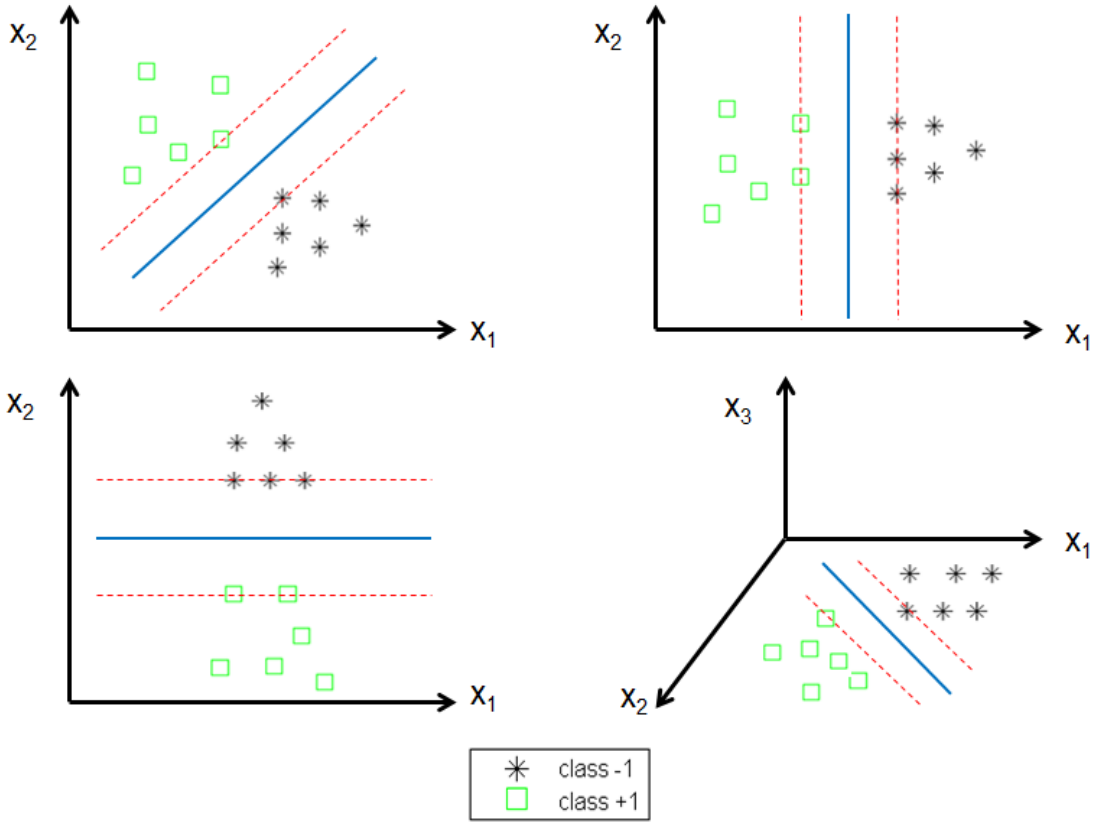


Figure 1.14: Example of several SVMs and how to interpret the weight vector $\mathbf{w}$

tions (Eqs. 1.23, 1.24 and 1.25). We recall Eq. 1.25:

$$\alpha_i(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1) = 0$$

From this, for every datapoint $\mathbf{x}_i$, either $\alpha_i^* = 0$ or $y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1$. Any datapoint with $\alpha_i^* = 0$ do not appear in the sum of the decision function f in Eq. 1.34 or 1.41. Hence, they play no role for the classification decision of a new sample $\mathbf{x}_j$. The others $\mathbf{x}_i$ such that $\alpha_i^* > 0$ corresponds to the support vector. Looking at the distribution of α_i^* allows also to have either a better understanding of the datasets, or either to detect outliers. The higher is the coefficient α_i^* for a sample $\mathbf{x}_i$, the more the sample $\mathbf{x}_i$ impacts on the decision function f. However, unusual high value of α_i^* among the samples can lead to two interpretations: either this point is a critical point to the decision, either this point is an outlier. In the soft margin formulation, by constraining α_i^* to be inferior to C (Box constraints) the effect of outliers can be reduced and controlled.

1.2.2.f Extensions of SVM

SVM has received many interests in recent years. Many extensions has been developed such

as ν-SVM, asymmetric soft margin SVM or multiclass SVM [KU02]; [CS01]. One interesting extension is the extension of Support Vector Machine to regression problems, also called Support Vector Regression (SVR). The objective is to find a linear regression model $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$. To preserve the property of sparseness, the idea is to consider an ϵ-insensitive error function. It gives zero error if the absolute difference between the prediction $f(\mathbf{x}_i)$ and the target y_i is less than ϵ where $\epsilon > 0$ penalize samples that are outside of a ϵ-tube as shown as in Fig. 1.15.

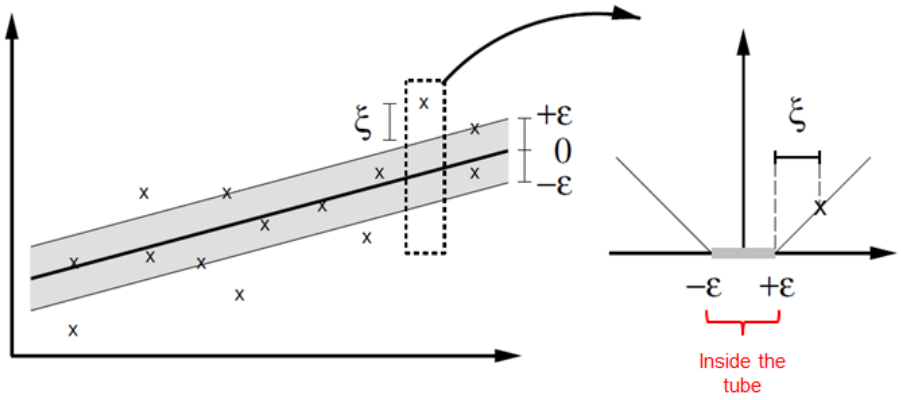


Figure 1.15: Illustration of SVM regression (left), showing the regression curve with the ϵ-insensitive "tube" (right). Samples $\mathbf{x}_i$ above the ϵ-tube have $\xi_1 > 0$ and $\xi_1 = 0$, points below the ϵ-tube have $\xi_2 = 0$ and $\xi_2 > 0$, and points inside the ϵ-tube have $\xi = 0$.

An example of ϵ-insensitive error function E_ϵ is given by:

$$E_\epsilon(f(\mathbf{x}_i) - y_i) = \begin{cases} 0 & \text{if} \quad |f(\mathbf{x}_i) - y_i| < \epsilon \\ |f(\mathbf{x}_i) - y_i| - \epsilon & \text{otherwise} \end{cases} \tag{1.44}$$

The soft margin optimization problem in its primal form is formalized as:

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \left(\overbrace{\frac{1}{2}||\mathbf{w}||_2^2}^{Regularization} + C \overbrace{\sum_{i=1}^{n}(\xi_{i_1} + \xi_{i_2})}^{Loss} \right) \tag{1.45}$$

$$\textbf{s.t.} \ \forall i = 1 \dots n :$$

$$y_i - (\mathbf{w}^T\mathbf{x}_i + b) \geq \epsilon - \xi_{i_1} \tag{1.46}$$

$$(\mathbf{w}^T\mathbf{x}_i + b) - y_i \geq \epsilon - \xi_{i_2} \tag{1.47}$$

$$\xi_{i_1} \geq 0 \tag{1.48}$$

$$\xi_{i_2} \geq 0 \tag{1.49}$$

The slacks variables are divided into 2 slacks variables, one for samples above the decision

function $f\ (\xi_{i_1})$, and one for samples under the decision function $f\ (\xi_{i_2})$. As for SVM, it is possible to have a dual formulation:

$$\underset{\boldsymbol{\alpha}}{\text{argmax}} \left(\sum_{i=1}^{n} y_i(\alpha_{i_1} - \alpha_{i_2}) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_{i_1} - \alpha_{i_2})(\alpha_{j_1} - \alpha_{j_2})(\mathbf{x}_i.\mathbf{x}_j) \right) \tag{1.50}$$

$$\textbf{s.t. } \forall i = 1 \dots n :$$

$$\sum_{i=1}^{n} \alpha_{i_1} = \sum_{i=1}^{n} \alpha_{i_2} \tag{1.51}$$

$$0 \leq \alpha_{i_1} \leq C \tag{1.52}$$

$$0 \leq \alpha_{i_2} \leq C \tag{1.53}$$

As in SVM, we obtain three possible decision functions for a new sample $\mathbf{x}_j$, respectively in its primal, dual, and non-linear form:

$$f(\mathbf{x}_j) = \mathbf{w}^T \mathbf{x}_j + b \tag{1.54}$$

$$f(\mathbf{x}_j) = \sum_{i=1}^{n} (\alpha^*_{i_1} - \alpha^*_{i_2})(\mathbf{x}_i.\mathbf{x}_j) + b \tag{1.55}$$

$$f(\mathbf{x}_j) = \sum_{i=1}^{n} (\alpha^*_{i_1} - \alpha^*_{i_2}) K(\mathbf{x}_i, \mathbf{x}_j) + b \tag{1.56}$$

More informations about the calculation development can be found in [Bis06].

1.2.3 Other classification algorithms

Partie non encore rédigée. A faire à la fin.

- Positionner les travaux par rapport aux autres méthodes d'apprentissage supervisé
- S'intéresser au Deep neural network (à la mode en ce moment)
- RVM, Decision Tree,
- Ne pas trop développer
- Dans notre cas, on ne s'intéressera pas à ce type d'algorithmes (type deep learning) car il ne repose pas sur une notion de distance et les features qui sont trouvés ne sont pas interprétables

1.3 Conclusion of the chapter

This chapter has presented two machine learning algorithms used in our proposition: the k-Nearest Neighbors (k-NN) and the Support Vector Machine (SVM). We review the different steps in a machine learning framework: data normalization, model selection and model evaluation. In the following, we consider the k-NN as our classifier. The SVM will be used in our work for its large margin concept.

Our objective being the learning of a metric that optimizes the performances of the k-NN classifier, we review in the next section some metrics proposed for time series as well as metric learning concept for static data.

Chapter 2

Time series metrics and metric learning

Sommaire

In this chapter, we first present the definition of time series. Then, we recall the general properties of a metric and introduce some metrics proposed for time series. In particular, we focus on amplitude-based, behavior-based and frequential-based metrics. As real time series are subjected to varying delays, we recall the concept of alignment and dynamic programming. Then, we present some proposed combined metrics for time series. Finally, we review the concept of metric learning.

2.1 Definition of a time series

Time series are frequently data that can be found in various emerging applications such as sensor networks, smart buildings, social media networks or Internet of Things (IoT) [Naj+12]; [Ngu+12]; [YG08]. They are involved in many learning problems such as recognizing a human movement in a video, detect a particular operating mode, etc. [PAN+08]; [Ram+08].

In **clustering** problems, one would like to organize similar time series together into homogeneous groups. In **classification** problems, the aim is to assign time series to one of several predefined categories (e.g., different types of defaults in a machine). In **regression** problems, the objective is to predict a continuous value from observed time series (e.g., forecasting the measurement of a power meter from pressure and temperature sensors). Due to their temporal and structured nature, time series constitute complex data to be analyzed by classic machine learning approaches.

For physical systems, a time series of length T can be seen as a signal, sampled at a frequency f_e, in a temporal window $[0; \frac{T}{f_e}]$. From a mathematical perspective, a time series is a collection of a finite number of normalized observations made sequentially at discrete time instants $t = 1, ..., Q$. Note that when $f_e = 1$, $Q = T$.

Let $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{iQ})$ be a univariate time series of length Q. Each observation x_{it} is bounded (i.e., the infinity is not a valid value: $x_{it} \neq \pm\infty$). The time series $\mathbf{x}_i$ is said to be univariate if the collection of observations x_{it} comes from the observations of one variable (i.e., the temperature measured by one sensor). When the observations are made at the same time from p variables (several sensors such as the temperature, the pressure, etc.), the time series is said multivariate. One possible representation is $\mathbf{x}_i = (\mathbf{x}_{i,1}, ..., \mathbf{x}_{i,p}) = (x_{i1,1}, ..., x_{iQ,1}, x_{i1,2}, ..., x_{i1,p}, ..., x_{iQ,p})$, where $\mathbf{x}_{i,j} = (\mathbf{x}_{i1,j}, \ldots, \mathbf{x}_{iQ,j})$. For simplification purpose, we consider in the following univariate time series.

Some authors propose to extract representative features from time series. Fig. 2.1 illustrates a model for time series proposed by Chatfield in [Cha04]. It states that a time series can be decomposed into 3 components: a trend, a cycle (periodic component) and a residual (irregular variations).

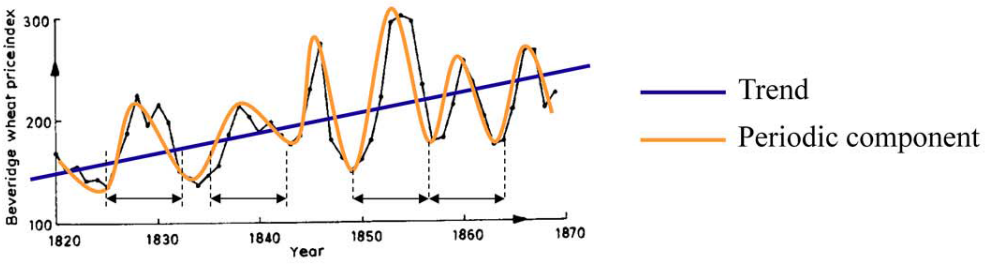


Figure 2.1: The Beveridge wheat price index is the average in nearly 50 places in various countries measured in successive years from 1500 to 1869. [1]

According to Chatfield, most time series exhibit a variation at a fixed period of time (seasonality) such as for example the seasonal variation of temperature. Beyond this cycle, there exists either or both a long term change in the mean (trend) that can be linear, quadratic, and a periodic (cyclic) component. In practice, these 3 features are rarely sufficient for the classification or regression of real time series.

[1] This time series can be downloaded from `http://www.york.ac.uk/depts/maths/data/ts/ts04.dat`

Other authors made the hypothesis of time independency between the observations x_{it}. They consider time series as a static vector data and use classic machine learning algorithms [Lia+12]; [CT01]; [HWZ13]; [HHK12]. Our work focus on classification and regression problems, and on time series comparison through metrics.

2.2 Properties of a metric

A mapping $D : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}^+$ over a vector space $\mathbb{R}^p$ is called a metric or a distance if for all vectors $\forall \mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_l \in \mathbb{R}^p$, it satisfies the properties:

Comment [CTD8]: je préfère garder l'espace pour + de visibilité

1. $D(\mathbf{x}_i, \mathbf{x}_j) \geq 0$ (positivity)
2. $D(\mathbf{x}_i, \mathbf{x}_j) = D(\mathbf{x}_j, \mathbf{x}_i)$ (symmetry)
3. $D(\mathbf{x}_i, \mathbf{x}_j) = 0 \Leftrightarrow \mathbf{x}_i = \mathbf{x}_j$ (distinguishability)
4. $D(\mathbf{x}_i, \mathbf{x}_j) + D(\mathbf{x}_j, \mathbf{x}_l) \geq D(\mathbf{x}_i, \mathbf{x}_l)$ (triangular inequality)

A mapping D that satisfies at least properties 1, 2, 3 is called a dissimilarity, and the one that satisfies at least properties 1, 2, 4 a pseudo-metric. Note that for a metric, a dissimilarity and a pseudo metric, if a time series $\mathbf{x}_i$ is expected to be closer to $\mathbf{x}_j$ than to $\mathbf{x}_l$, then $D(\mathbf{x}_i, \mathbf{x}_j) \leq D(\mathbf{x}_i, \mathbf{x}_l)$. On the contrary, the mapping is called a similarity S when the time series $\mathbf{x}_i$ is expected to be closer to $\mathbf{x}_j$ than to $\mathbf{x}_l$ and then $S(\mathbf{x}_i, \mathbf{x}_j) \geq S(\mathbf{x}_i, \mathbf{x}_l)$. To simplify the discussion in the following, we refer to pseudo-metric and dissimilarity as metrics, pointing out the distinction only when necessary.

2.3 Unimodal metrics for time series

Defining and evaluating metrics for time series has become an active area of research for a wide variety of problems in machine learning [Din+08]; [Naj+12]. In the following, we suppose that time series have the same lengths Q and have been regularly sampled at the frequency f_e. Let $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{iQ})$ and $\mathbf{x}_j = (x_{i1}, x_{i2}, ..., x_{iQ})$ be two univariate time series of length Q.

A large number of distance measures have been proposed in the literature [MV14]. Contrary to static data, time series may exibit modalities and specificities due to their temporal nature (e.g., value, shape, frequency, delay, temporal locality). In this section, we review 3 categories of time series metrics used in our work: amplitude-based, frequential-based and behavior-based.

2.3.1 Amplitude-based metrics

The most usual comparison measures are amplitude-based metrics, where time series are compared in the temporal domain on their amplitudes regardless of their behaviors or frequential characteristics. Among these metrics, there are the commonly used Euclidean distance that compares elements observed at the same time [Din+08]:

$$d_E(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{t=1}^{Q}(x_{it} - x_{jt})^2} \tag{2.1}$$

Note that the Euclidean distance is a particular case of the Minkowski L_p norm ($p = 2$). An other amplitude-based metric is the Mahalanobis distance [PL12], defined as a dissimilarity measure between two random vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ of the same distribution with the covariance matrix $\mathbf{M}$:

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M}^{-1} (\mathbf{x}_i - \mathbf{x}_j) \tag{2.2}$$

If the covariance matrix $\mathbf{M}$ is the identity matrix, the Mahalanobis distance is equal to the Euclidean distance. If the covariance matrix $\mathbf{M}$ is diagonal, then the resulting distance measure is called a normalized Euclidean distance:

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{l=1}^{Q} \frac{(x_{il} - x_{jl})^2}{m_l}} \tag{2.3}$$

where m_l is the variance of the x_{il} and x_{jl} over the sample set. In the following of the work, we consider the standard Euclidean distance d_E as the amplitude-based distance d_A.

In the example of Fig. 2.2, the aim is to determined which time series ($\mathbf{x}_2$ or $\mathbf{x}_3$) is the closest to $\mathbf{x}_1$. The amplitude-based distance d_A states that $\mathbf{x}_2$ is closer to $\mathbf{x}_1$ than $\mathbf{x}_3$ since $d_A(\mathbf{x}_1, \mathbf{x}_2) = 7.8816 < d_A(\mathbf{x}_1, \mathbf{x}_3) = 31.2250$.

2.3.2 Frequential-based metrics

The second category, commonly used in signal processing, relies on comparing time series based on their frequential properties (e.g. Fourier Transform, Wavelet, Mel-Frequency Cepstral Coefficients [SS12]; [TC98]; [BM67]). In our work, we limit the frequential comparison to Discrete Fourier Transform [Lhe+11], but other frequential properties can be used as well. Thus, for time series comparison, first the time series $\mathbf{x}_i$ are transformed into their Fourier representation $\tilde{\mathbf{x}}_i = [\tilde{x}_{i1}, ..., \tilde{x}_{iF}]$, with $\tilde{x}_{if}$ the complex component at frequential index f. The Euclidean distance is then used between their respective complex number modules $\tilde{x}_{if}$, noted $|\tilde{x}_{if}|$:

$$d_F(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{f=1}^{F}(|\tilde{x}_{if}| - |\tilde{x}_{jf}|)^2} \tag{2.4}$$

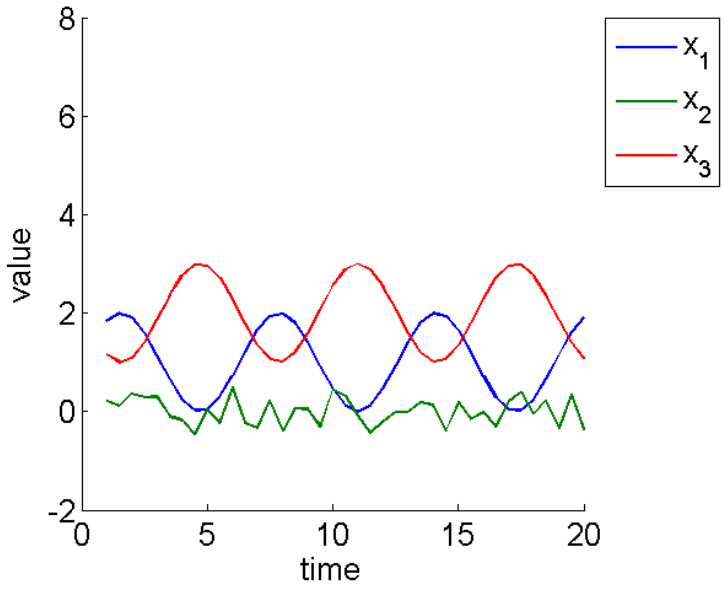
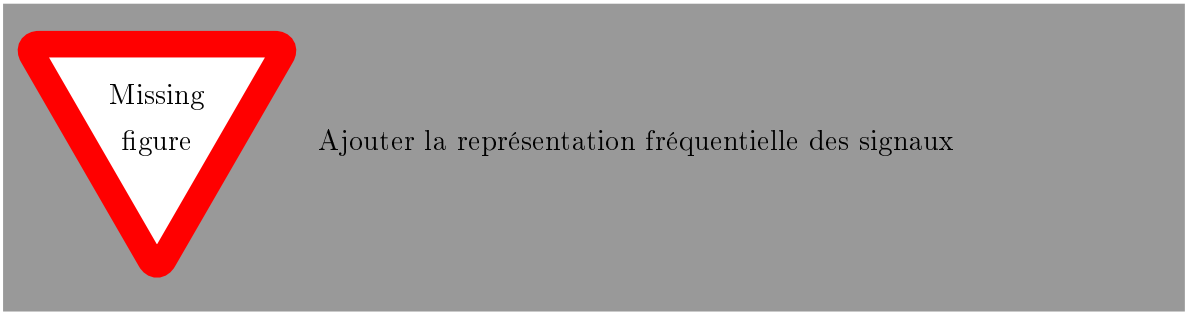


Figure 2.2: 3 toy time series. Time series in blue and red are two sinusoïdal signals. Time series in green is a random signal.

In the example of Fig. 2.2, the frequential-based distance d_F states that the time series $\mathbf{x}_3$ is closer to $\mathbf{x}_1$ than $\mathbf{x}_2$ since $d_F(\mathbf{x}_1, \mathbf{x}_3) = 0.8519 < d_F(\mathbf{x}_1, \mathbf{x}_2) = 0.9250$. This can be illustrated in the Frequency domain (Fig. **??**)



2.3.3 Behavior-based metrics

The third category of metrics aims to compare time series based on their shape or behavior despite the range of their amplitudes. By time series of similar behavior, it is generally intended that for all temporal window $[t, t']$, they increase or decrease simultaneously with the same growth rate. On the contrary, they are said of opposite behavior if for all $[t, t']$, if one time series increases, the other one decreases and (vise-versa) with the same growth rate in absolute value. Finally, time series are considered of different behaviors if they are not similar, nor opposite. Many applications refer to the Pearson correlation [AT10]; [Ben+09] for

behavior comparison. A generalization of the Pearson correlation is introduced in [DCA11]:

$$cort_r(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum (x_{it} - x_{it'})(x_{jt} - x_{jt'})}{\sqrt{\sum (x_{it} - x_{it'})^2}\sqrt{\sum (x_{jt} - x_{jt'})^2}} \tag{2.5}$$

where $|t-t'| \leq r$, $r \in [1, ..., Q-1]$. The parameter r can be tuned or fixed a priori. It measures the importance of noise in data. For non-noisy data, low orders r is generally sufficient. For noisy data, the practitioner can either use de-noising data technics (Kalman or Wiener filtering [Kal60]; [Wie42]), or fix a high order r.

The temporal correlation *cort* computes the sum of growth rate between $\mathbf{x}_i$ and $\mathbf{x}_j$ between all pairs of values observed at $[t, t']$ for $t' \leq t+r$ (r-order differences). The value $cort_r(\mathbf{x}_i, \mathbf{x}_j) = +1$ means that $\mathbf{x}_i$ and $\mathbf{x}_j$ have similar behavior. The value $cort_r(\mathbf{x}_i, \mathbf{x}_j) = -1$ means that $\mathbf{x}_i$ and $\mathbf{x}_j$ have opposite behavior. Finally, $cort_r(\mathbf{x}_i, \mathbf{x}_j) = 0$ expresses that their growth rates are stochastically linearly independent (different behaviors).

When $r = Q - 1$, it leads to the Pearson correlation. As $cort_r$ is a similarity measure, it can be transformed into a dissimilarity measure:

$$d_B(\mathbf{x}_i, \mathbf{x}_j) = \frac{1 - cort_r(\mathbf{x}_i, \mathbf{x}_j)}{2} \tag{2.6}$$

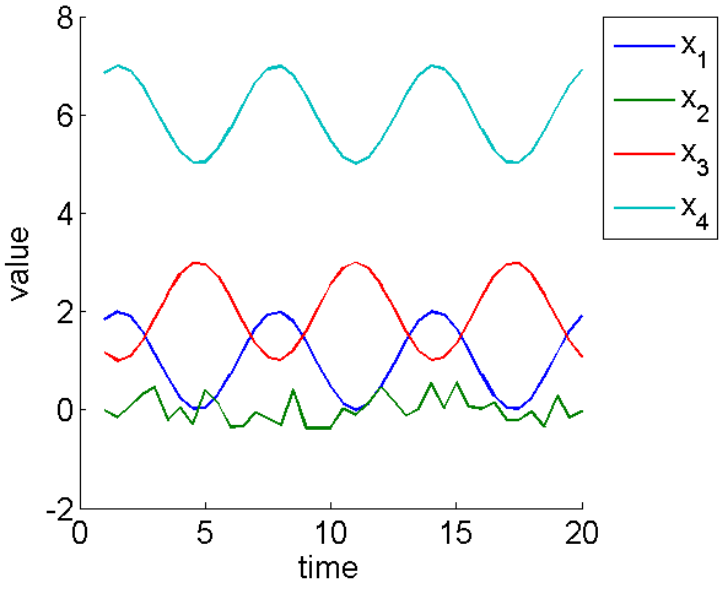


Figure 2.3: The signal from Fig. 2.2 and a signal $\mathbf{x}_4$ which is signal $\mathbf{x}_1$ and an added translation. Based on behavior comparison, $\mathbf{x}_4$ is the closest to $\mathbf{x}_1$.

Considering Fig. 2.3, we obtain:

$$d_B(\mathbf{x}_1, \mathbf{x}_2) = 0.477$$
$$d_B(\mathbf{x}_1, \mathbf{x}_3) = 1$$
$$d_B(\mathbf{x}_1, \mathbf{x}_4) = 0$$

2.3.4 Other metrics and Kernels for time series

A faire à la fin, pas urgent

- Il existe dans la littérature de nombreuses autres métriques pour les séries temporelles (laisser la porte ouverte).
- Certaines métriques sont utilisées dans le domaine temporelle
- D'autres métriques sont utilisés dans d'autres représentations (Wavelet, etc.)
- Certaines combinent la représentation temporelles et fréquentielles (Représentation spectrogramme en temps-fréquence)
- Se baser sur l'article "TSclust : An R Package for Time Series Clustering".
- Fermer le cadre : dans la suite de notre travail, on ne va pas les utiliser mais elles pourront être intégrées dans le framework qui suivra au chapitre suivant

2.4 Time series alignment and dynamic programming approach

In some applications, time series needs to be compared at different time t (i.e. energy data [Naj+12]) whereas in others, comparing time series on the same time t is essential (i.e. gene expression [DCN07]). When time series are asynchronous (i.e. varying delays or dynamic changes), they must be aligned before any analysis process. The asynchronous effects can be of various natures: time shifting (phase shift in signal processing), time compression or time dilatation. For example, in the case of voice recognition (Fig. 2.4), it is straightforward that a same sentence said by two different speakers will produce different time series: one speaker may speak faster than the other; one speaker may take more time on some vowels, etc.

To cope with delays and dynamic changes, dynamic programming approach has been introduced [BC94]. An alignment $\boldsymbol{\pi}$ of length $|\boldsymbol{\pi}| = m$ between two time series $\mathbf{x}_i$ and $\mathbf{x}_j$ of length Q is defined as the set of m ($Q \leq m \leq 2Q-1$) couples of aligned elements of $\mathbf{x}_i$ to m elements of $\mathbf{x}_j$:

$$\boldsymbol{\pi} = ((\pi_i(1), \pi_j(1)), (\pi_i(2), \pi_j(2)), \ldots, (\pi_i(m), \pi_j(m))) \tag{2.7}$$

Comment [MR9]: Modifier figure. enlever 'one' et mettre la même échelle temporelle

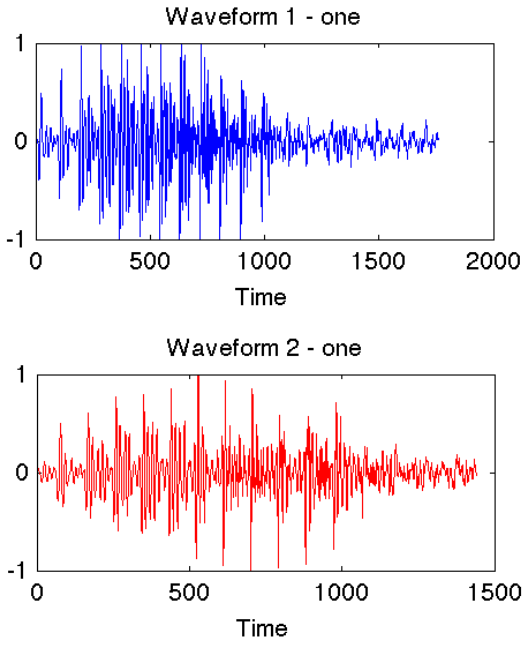


Figure 2.4: Example of a same sentence said by two different speakers. Time series are shifted, compressed and dilatated in the time.

where the applications π_i and π_j defined from $\{1, ..., m\}$ to $\{1, ..., Q\}$ obey the following boundary monotonicity conditions:

$$1 = \pi_i(1) \leq \pi_i(2) \leq ... \leq \pi_i(m) = Q \tag{2.8}$$
$$1 = \pi_j(1) \leq \pi_j(2) \leq ... \leq \pi_j(m) = Q \tag{2.9}$$

Comment [AD10]: Ahla pas fan des notations

$\forall l \in \{1, ..., m\}$,

$$\pi_i(l+1) \leq \pi_i(l) + 1 \tag{2.10}$$
$$\text{and} \quad \pi_j(l+1) \leq \pi_j(l) + 1 \tag{2.11}$$
$$\text{and} \quad (\pi_i(l+1) - \pi_i(l)) - (\pi_j(l+1) - \pi_j(l)) \geq 1. \tag{2.12}$$

Intuitively, an alignment $\boldsymbol{\pi}$ defines a way to associate elements of two time series. Alignments can be described by paths in the $Q \times Q$ grid that crosses the elements of $\mathbf{x}_i$ and $\mathbf{x}_j$ (Fig. 2.5). We denote $\boldsymbol{\pi}$ a valid alignment and A, the set of all possible alignments between $\mathbf{x}_i$ and $\mathbf{x}_j$ ($\boldsymbol{\pi} \in A$). To find the best alignment $\boldsymbol{\pi}^*$ between two time series $\mathbf{x}_i$ and $\mathbf{x}_j$, the Dynamic Time Warping (DTW) algorithm has been proposed [KR04]; [SC].

DTW requires to choose a cost function φ to be optimised, such as a dissimilarity function (d_A, d_B, d_F, etc.). Classical DTW uses the Euclidean distance d_A (Eq. 2.1) as the cost

function [BC94]. The warp path $\boldsymbol{\pi}$ is optimized for the chosen cost function φ:

$$\boldsymbol{\pi}^* = \underset{\boldsymbol{\pi} \in A}{\operatorname{argmin}} \frac{1}{|\boldsymbol{\pi}|} \sum_{(t,t') \in \boldsymbol{\pi}} \varphi(x_{it}, x_{jt'}) \tag{2.13}$$

When the cost function φ is a similarity measure, the optimization involves maximization instead of minimization. When other constraints are applied on $\boldsymbol{\pi}$, Eq. (2.13) leads to other variants of DTW (Sakoe-Shiba [SC78], Itakura parallelogram [RJ93]). Finally, the warped signals $\mathbf{x}_{i,\boldsymbol{\pi}}$ and $\mathbf{x}_{j,\boldsymbol{\pi}}$ are defined as:

$$\mathbf{x}_{i,\boldsymbol{\pi}} = (x_{i\pi_i(1)}, ..., x_{i\pi_i(m)}) \tag{2.14}$$

$$\mathbf{x}_{j,\boldsymbol{\pi}} = (x_{j\pi_j(1)}, ..., x_{j\pi_j(m)}) \tag{2.15}$$

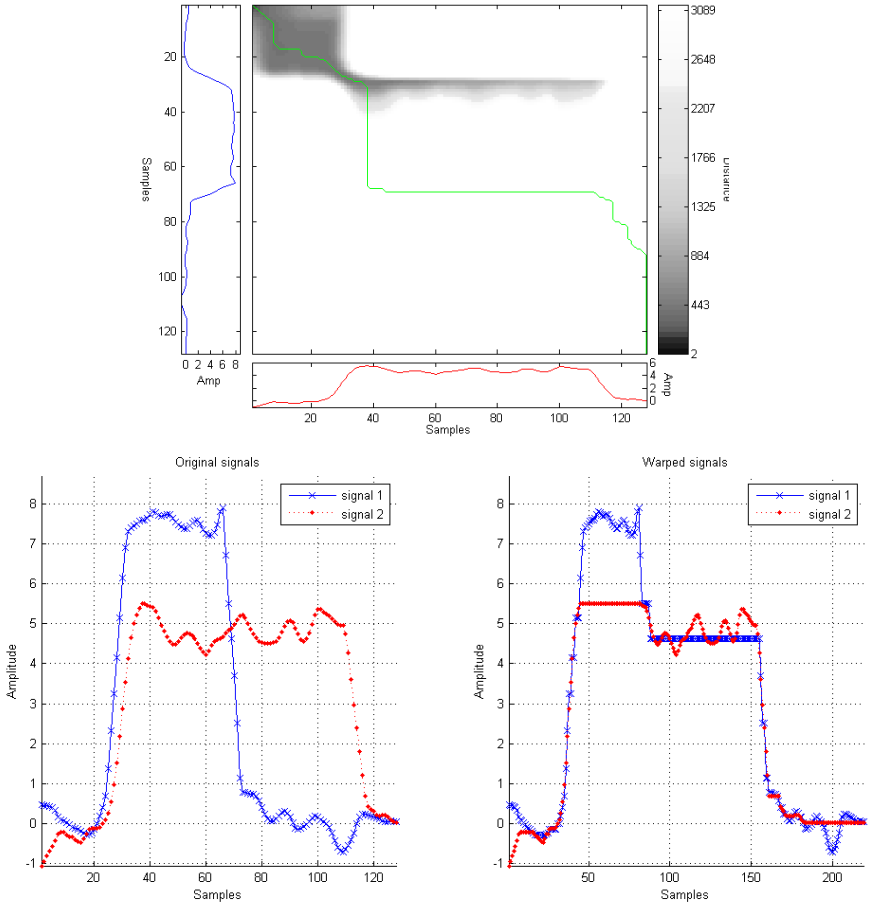


Figure 2.5: Example of DTW grid between 2 time series $\mathbf{x}_i$ and $\mathbf{x}_j$ (top) and the signals before and after warping (bottom). On the DTW grid, the two signals can be represented on the left and bottom of the grid. The optimal path $\boldsymbol{\pi}^*$ is represented in green line and show to associate elements of $\mathbf{x}_i$ to element of $\mathbf{x}_j$. Background show in grey scale the value of the considered metric (amplitude-based distance d_A in classical DTW)

The previous metric (amplitude-based d_A, behavior-based d_B) can be then computed on the warped signals $\mathbf{x}_{i,\boldsymbol{\pi}^*}$ and $\mathbf{x}_{j,\boldsymbol{\pi}^*}$. In the following, we suppose that the best alignment $\boldsymbol{\pi}^*$ is found. For simplification purpose, we refer $\mathbf{x}_{i,\boldsymbol{\pi}^*}$ and $\mathbf{x}_{j,\boldsymbol{\pi}^*}$ as $\mathbf{x}_i$ and $\mathbf{x}_j$.

2.5 Combined metrics for time series

In most classification problems, it is not known a priori if time series of a same class exhibits same characteristics based on their amplitude, behavior or frequential components alone. In some cases, several components (amplitude, behavior and/or frequential) may be implied.

A first technic considers a classifier for each p metric and combines the decision of the p resulting classifiers. This methods is referred as post-fusion , not considered in our work. Other propositions show the benefit of involving both behavior and amplitude components through a combination function. They combines the unimodal metrics together to obtain a single metric used after that in a classifier. This is called pre-fusion. The most classical combination functions combines the unimodal metrics (mainly d_A and d_B) through linear and geometric functions:

$$D_{Lin}(\mathbf{x}_i, \mathbf{x}_j) = \alpha d_B(\mathbf{x}_i, \mathbf{x}_j) + (1-\alpha) d_A(\mathbf{x}_i, \mathbf{x}_j) \tag{2.16}$$

$$D_{Geom}(\mathbf{x}_i, \mathbf{x}_j) = (d_B(\mathbf{x}_i, \mathbf{x}_j))^{\alpha} (d_A(\mathbf{x}_i, \mathbf{x}_j))^{1-\alpha} \tag{2.17}$$

where $\alpha \in [0; 1]$ defines the trade-off between the amplitude d_A and the behavior d_B components, and is thus application dependent. In general, it is learned through a grid search procedure. Without being restrictive, these combinations can be extended to take into account more unimodal metrics.
More specific work on d_A and *cort* propose to combine the two components through a sigmoid combination function [DCA11]:

$$D_{Sig}(\mathbf{x}_i, \mathbf{x}_j) = \frac{2 d_A(\mathbf{x}_i, \mathbf{x}_j)}{1 + \exp(\alpha cort_r(\mathbf{x}_i, \mathbf{x}_j))} \tag{2.18}$$

where α is a parameter that defines the compromise between behavior and amplitude components. When α is fixed to 0, the metric only includes the value proximity component. For $\alpha \geq 6$, the metric completely includes the behavior proximity component.

Fig.2.6 illustrates the value of the resulting combined metrics (D_{Lin}, D_{Geom} and D_{Sig}) in 2-dimensional space using contour plots for different values of the trade-off α. For small value of α ($\alpha = 0$), the three metrics only includes d_A. For high value of α ($\alpha = 1$), D_{Lin} and D_{Geom} only includes d_B. For $\alpha = 6$, D_{Sig} doesn't include completely *cort*. Note that these combinations are fixed and defined independently from the analysis task at hand. Moreover, in the case of D_{Sig}, only two variables are taking into account in these combined metrics and the component $cort_r$ can be seen as a penalizing factor of d_A. It doesn't represent a real compromise between value and behavior components. Finally, by adding metrics, the grid

search to find the best parameters can become time consuming.

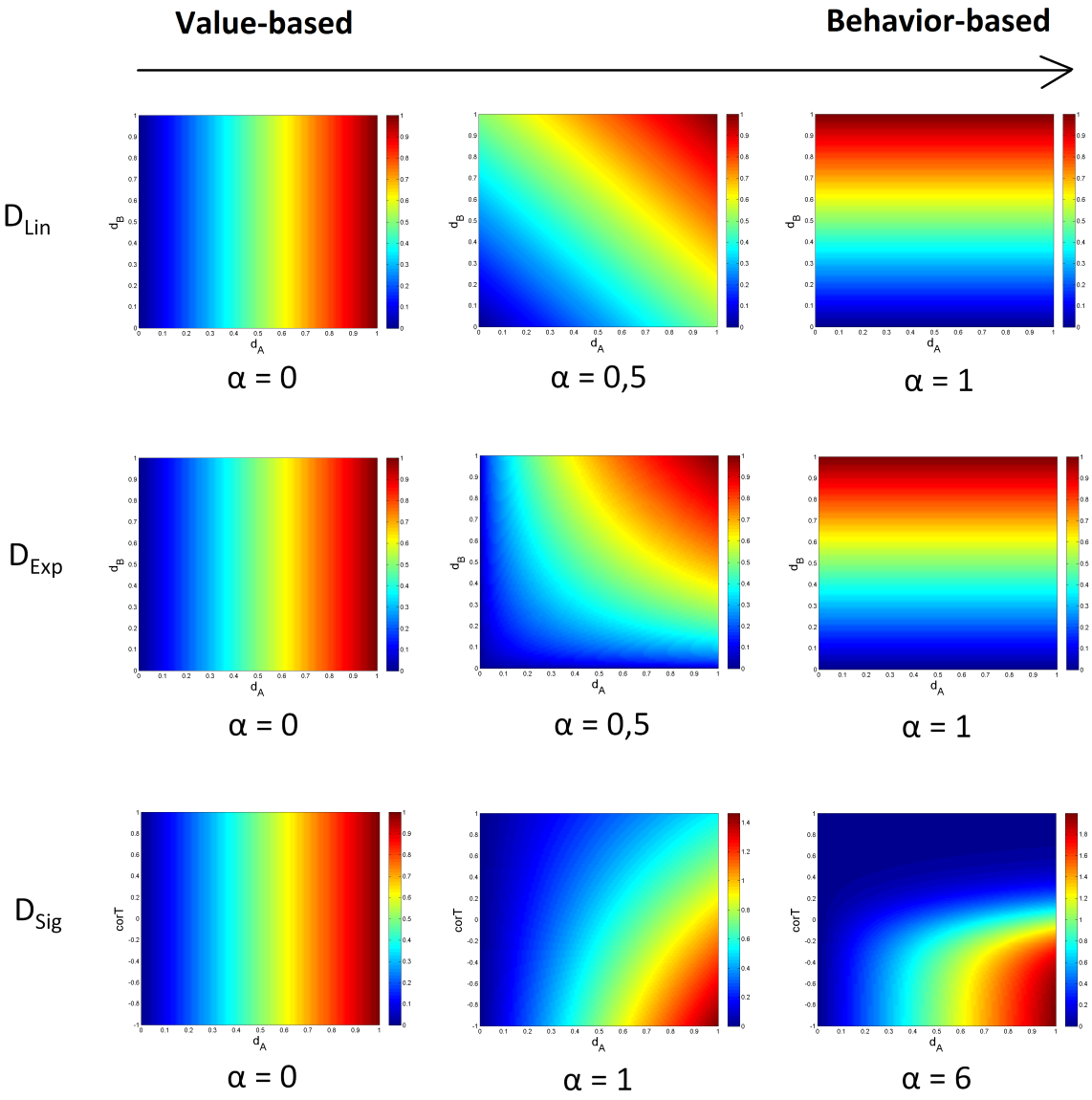


Figure 2.6: Contour plot of the resulting combined metrics: D_{Lin} (1^{st} line), D_{Geom} (2^{nd} line) and D_{Sig} (3^{rd} line), for different value of α (D_{Sig}: $\alpha = 0; 1; 6$ and D_{Lin} and D_{Geom}: $\alpha = 0; 0.5; 1$). For D_{Sig}, the first and second dimensions are respectively the amplitude-based metrics d_A and the temporal correlation $corT$; for D_{Lin} and D_{Geom}, they correspond to d_A and the behavior-based metric d_B.

2.6 Metric learning

As our objective is to learn a metric in order to optimize the performance of the k-NN classifier, we review first metric learning concepts. Then, we focus on the framework proposed by

Weinberger & Saul for Large Margin Nearest Neighbor (LMNN) classification [WS09].

2.6.1 Review on metric learning work

In the case of static data, many work have demonstrated that k-NN classification performances depends highly on the considered metric and can be improved by learning an appropriate metric [She+02]; [Gol+04]; [CHL05]. Metric Learning can be defined as a process that aims to learn a distance from labeled examples by making closer samples that are expected to be similar, and far away those expected to be dissimilar.

A faire, avec papier PRL et papier Aurélien Bellet

2.6.2 Large Margin Nearest Neighbors (LMNN)

Let $\mathbf{X} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ be a set of N static vector samples, $\mathbf{x}_i \in \mathbb{R}^p$, p being the number of descriptive features and y_i the class labels. Weinberger & Saul proposed in [WS09] an approach to learn a dissimilarity metric D for a large margin k-NN in the case of static data.

Large Margin Nearest Neighbor (LMNN) approach is based on two intuitions: first, each training sample $\mathbf{x}_i$ should have the same label y_i as its k nearest neighbors; second, training samples with different labels should be widely separated. For this, the concept of **target** and **imposters** for each training sample $\mathbf{x}_i$ is introduced. The training sample $\mathbf{x}_i$ is referred as a **center point**. Target neighbors of $\mathbf{x}_i$, noted $j \rightsquigarrow i$, are the k closest $\mathbf{x}_j$ of the same class ($y_j = y_i$), while imposters of $\mathbf{x}_i$, denoted, $l \nrightarrow i$, are the $\mathbf{x}_l$ of different class ($y_l \neq y_i$) that invade the perimeter defined by the farthest targets of $\mathbf{x}_i$. Mathematically, for a sample $\mathbf{x}_i$, an imposter $\mathbf{x}_l$ is defined by an inequality related to the targets $\mathbf{x}_j$: $\forall l, \exists j \in j \rightsquigarrow i/$

$$D(\mathbf{x}_i, \mathbf{x}_l) \leq D(\mathbf{x}_i - \mathbf{x}_j) + 1 \tag{2.19}$$

Geometrically, an imposter $\mathbf{x}_l$ is a sample that invades the target neighborhood plus one unit margin as illustrated in Fig. 2.7. The target neighborhood is defined with respect to an initial metric. Without prior knowledge, L2-norm is often used. Metric Learning by LMNN aims to minimize the number of impostors invading the target neighborhood. By adding a margin of safety of one, the model is ensured to be robust to small amounts of noise in the training sample (large margin). The learned metric D pulls the targets $\mathbf{x}_j$ and pushes the imposters $\mathbf{x}_l$ as shown in Fig. 2.7.

LMNN approach learns a Mahalanobis distance D for a robust k-NN. We recall that the k-NN decision rule will correctly classify a sample if its k nearest neighbors share the same label (Section 1.2.1). The objective of LMNN is to increase the number of samples with this property by learning a linear transformation $\mathbf{L}$ of the input space ($\mathbf{x}_i = \mathbf{L}.\mathbf{x}_i$) before applying the k-NN classification:

$$D(\mathbf{x}_i, \mathbf{x}_j) = ||\mathbf{L}(\mathbf{x}_i, \mathbf{x}_j)||_2^2 \tag{2.20}$$

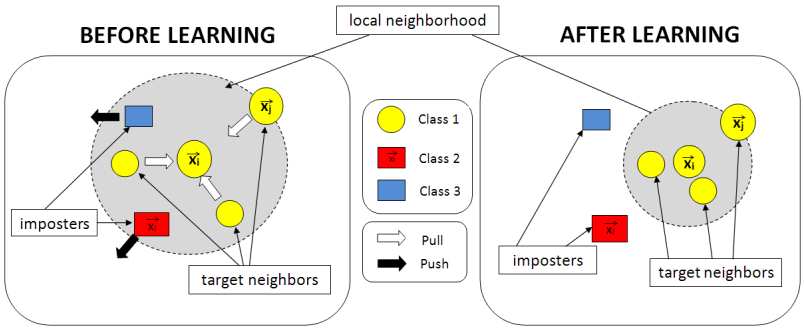


Figure 2.7: Pushed and pulled samples in the $k = 3$ target neighborhood of $\mathbf{x}_i$ before (left) and after (right) learning. The pushed (vs. pulled) samples are indicated by a white (vs. black) arrows (Weinberger & Saul [WS09]).

Commonly, the squared distances can be expressed in terms of the square matrix:

$$\mathbf{M} = \mathbf{L}'\mathbf{L} \tag{2.21}$$

It is proved that any matrix $\mathbf{M}$ formed as below from a real-valued matrix $\mathbf{L}$ is positive semidefinite (i.e., no negative eigenvalues) [WS09]. Using the matrix $\mathbf{M}$, squared distances can be expressed as:

$$D^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)\mathbf{M}(\mathbf{x}_i - \mathbf{x}_j) \tag{2.22}$$

The computation of the learned metric $D_{\mathbf{M}}$ can thus be seen as a two steps procedure: first, it computes a linear transformation of the samples $\mathbf{x}_i$ given by the transformation $\mathbf{L}$; second, it computes the Euclidean distance in the transformed space:

$$D^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = D^2(\mathbf{L}\mathbf{x}_i, \mathbf{L}\mathbf{x}_j) \tag{2.23}$$

Learning the linear transformation $\mathbf{L}$ is thus equivalent to learn the corresponding Mahalanobis metric D parametrized by $\mathbf{M}$. This equivalence leads to two different approaches to metric learning: we can either estimate the linear transformation $\mathbf{L}$, or estimate a positive semidefinite matrix $\mathbf{M}$. LMNN solution refers on the latter one.

Mathematically, it can be formalized as an optimization problem involving two competiting terms for each sample $\mathbf{x}_i$: one term penalizes large distances between nearby inputs with the same label (pull), while the other term penalizes small distances between inputs with different labels (push). For all samples $\mathbf{x}_i$, this implies a minimization problem:

$$\begin{aligned}
&\underset{\mathbf{M},\xi}{\operatorname{argmin}} \underbrace{\sum_{i,j \rightsquigarrow i} D^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j)}_{pull} + C \underbrace{\sum_{i,j \rightsquigarrow i, l \nrightarrow i} \frac{1 + y_{il}}{2} . \xi_{ijl}}_{push} \\
&\text{s.t. } \forall j \rightsquigarrow i, l \nrightarrow i, \\
&D^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_l) - D^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \geq 1 - \xi_{ijl} \\
&\xi_{ijl} \geq 0 \\
&\mathbf{M} \succeq 0
\end{aligned} \tag{2.24}$$

where C is a trade-off between the push and pull term and $y_{il} = -1$ if $y_i = y_l$ (same class) and $+1$ otherwise (different classes). Generally, the parameter C is tuned via cross validation and grid search. Similarly to Support Vector Machine (SVM) approach, slack variables ξ_{ijl} are introduced to relax the optimization problem.

2.6.3 Parallels between LMNN and SVM

Many connections can be made between LMNN and SVM: both are convex optimization problem based on a regularized and a loss term. In particular, Do & al. investigate this relationship and have shown that SVM can be formulated as a metric learning problem [Do+12]. The Mahalanobis distance $\mathbf{M}$ learned by LMNN can be expressed as a quadratic mapping $\boldsymbol{\phi}$. For a center point $\mathbf{x}_i$, for any sample $\mathbf{x}$, we have [Do+12]:

$$D^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}) = D^2(\mathbf{L}\mathbf{x}_i, \mathbf{L}\mathbf{x}) \tag{2.25}$$

$$D^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}) = \mathbf{w}_i^T \boldsymbol{\phi}(\mathbf{x}) + b_i \tag{2.26}$$

where $\mathbf{w}_i$ and b_i are the coefficient of the hyperplane H_i in the quadratic space $\boldsymbol{\phi}$.

Do & al. show that LMNN can be seen as a set of local SVM classifiers in the quadratic space induced by $\boldsymbol{\phi}$. For each center point $\mathbf{x}_i$, LMNN tries in its objective function to have its target neighbors $\mathbf{x}_j$ to have small value $\mathbf{w}_i^T \boldsymbol{\phi}(\mathbf{x}_j) + b_i$, i.e. be at the small distance from the hyperplane H_i. Minimizing the target neighbor distances from the hyperplane H_i makes the distance between support vectors and H_i small. Fig. 2.8 gives the equivalent point of view from the original space (Fig. 2.8(a)) into the quadratic space (Fig. 2.8(b)). The circle $\mathbf{C}_i$ with the center $\mathbf{L}\mathbf{x}_i$ in Fig. 2.8(a) corresponds to the hyperplane H_i in Fig. 2.8(b).

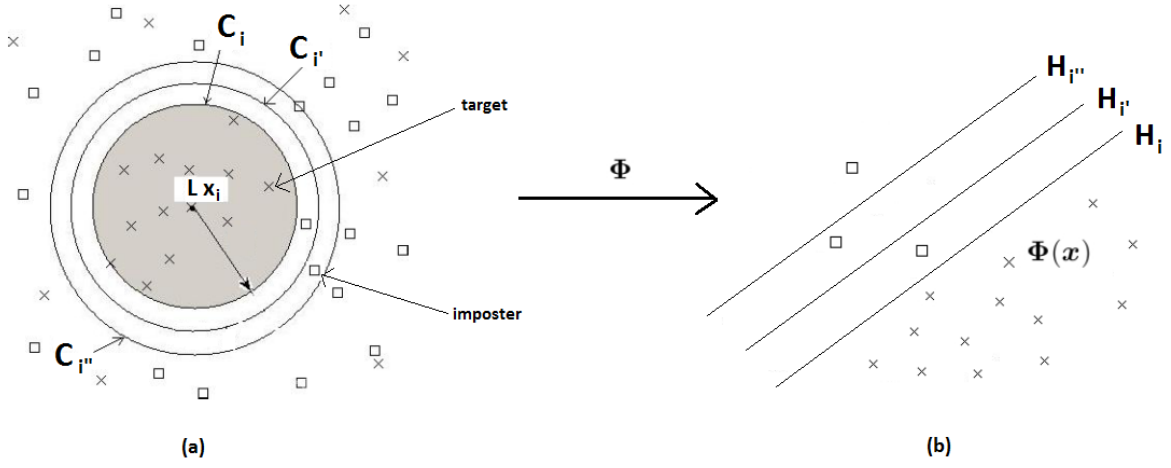


Figure 2.8: (a) Standard LMNN model view (b) LMNN model view under an SVM-like interpretation [Do+12]

Geometrically, SVM margin is defined globally with respect to a hyperplane, while LMNN margin is defined locally with respect to a center point $\mathbf{x}_i$. Fig. 2.9(a) illustrates the different

local linear models in the quadratic space. The optimization process of LMNN combines the different local SVM hyperplane by bringing each point $\phi(\mathbf{x}_i)$ around a consensus hyperplane H.

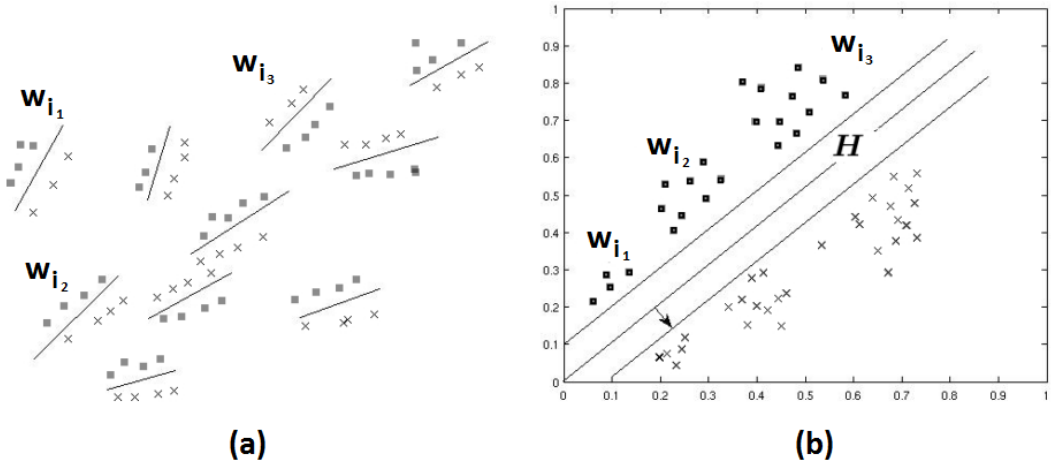


Figure 2.9: (a) LMNN in a local SVM-like view (b) LMNN in an SVM metric learning view [Do+12]

From these connections, some authors extends the LMNN approach to work in non-linear feature spaces by using the "kernel trick" . Finally, note that LMNN differs from SVM in which LMNN requires no modification for multiclass problems. ref

2.7 Conclusion of the chapter

To cope with modalities inherent to time series (amplitude, behavior, frequency, etc.), we review in this chapter several unimodal metrics for time series, in particular, the Euclidean distance d_A, the Temporal correlation d_B or the Fourier-based distance d_F. In practice, real time series may be subjected to delays and need to be re-aligned before any analysis task. For that, the Dynamic Time Warping (DTW) algorithm is used in practice. However, these metrics (d_A, d_B, d_F) only include one modality. In general, several modalities may be implied and authors proposed to combine temporal metrics together. They mainly combine the Euclidean distance d_A and the Temporal correlation d_B.

As k-NN performances is impacted by the choice of the metric, other work propose in the case of static data to learn the metric in order to optimize the k-NN classification. In the following, we extend this framework to learn a combined metric for a large margin k-NN classifier of time series.

Conclusion of Part I

In order to make the classification or regression of time series, a lot of technics exist in the literature. Our work focus on the k-NN classifier and SVM will be used in the following for its large margin concept. We note that the k-Nearest Neighbors algorithm is based on the comparison of time series through distance measures.

Considering time series as static data lead to the only comparison based on their amplitude and the same time. To take into account other specificities of time series (behavior, frequential components), other metrics (e.g., the temporal correlation d_B, the frequential-based distance d_F, etc.) and other methods (Dynamic Time Warping DTW, dichotomy) have been proposed in the literature to cope with temporal characteristics.

Learning an adequate metric is a key challenge to well classify time series. Inspired by Metric Learning work for static data, we propose in the following a framework to learn a Multi-modal and Multi-scale Metric for a robust nearest neighbor classifier of time series.

Part II

Multi-modal and Multi-scale Time series Metric Learning ($\mathrm{M}^2\mathrm{TML}$)

The first part has enlightened the importance of combining several modalities to make a better analysis (classification, regression) of time series.

In the first chapter, we present this pairwise representation and formalize the optimization problem and its adapted Support Vector Machine (SVM) equivalence. In the second chapter, we present the details of the proposed algorithm: Multi-modal and Multi-scale Time series Metric Learning ($\mathrm{M}^2\mathrm{TML}$).

Chapter 3

Pairwise space and Time series Metric Learning (TML) formalization

Sommaire

In this chapter, we formalize the problem of Time series Metric Learning (TML) which is the learning of a metric that combines several unimodal metrics for a robust k-NN classifier.
We first introduce a new space representation, the pairwise space. Secondly, we transpose the metric learning problem in the pairwise space. Finally, we propose three possible formulations: Linear programming, Quadratic programming and SVM-based approach.

3.1 Pairwise space representation

Let $d_1, ..., d_h..., d_p$ be p given metrics that allow to compare samples. For instance, in Chapter 2, we have proposed three types of metrics for time series: amplitude-based d_A, behavior-based d_B and frequential-based d_F. Our objective is to learn a metric D that combines the p metrics in order to optimize the performance of a k-NN classifier. Formally:

$$D = f(d_1, \ldots, d_p) \tag{3.1}$$

In this section, we first introduce a new space representation, the pairwise space. Then, we present how to define pairwise labels for classification and regression problem. Finally, we give some interpretations in the pairwise space.

3.1.1 Pairwise embedding

The computation of a metric d, and of course D, always takes into account a pair of samples $(\mathbf{x}_i, \mathbf{x}_j)$. We introduce a new space representation referred as the **pairwise space**. In this new space, illustrated in Fig. 3.1, a vector $\mathbf{x}_{ij}$ represents a pair of time series $(\mathbf{x}_i, \mathbf{x}_j)$ described by the p unimodal metrics d_h: $\mathbf{x}_{ij} = [d_1(\mathbf{x}_i, \mathbf{x}_j), ..., d_p(\mathbf{x}_i, \mathbf{x}_j)]^T$. We denote N the number of pairwise vectors $\mathbf{x}_{ij}$ generated by this embedding.

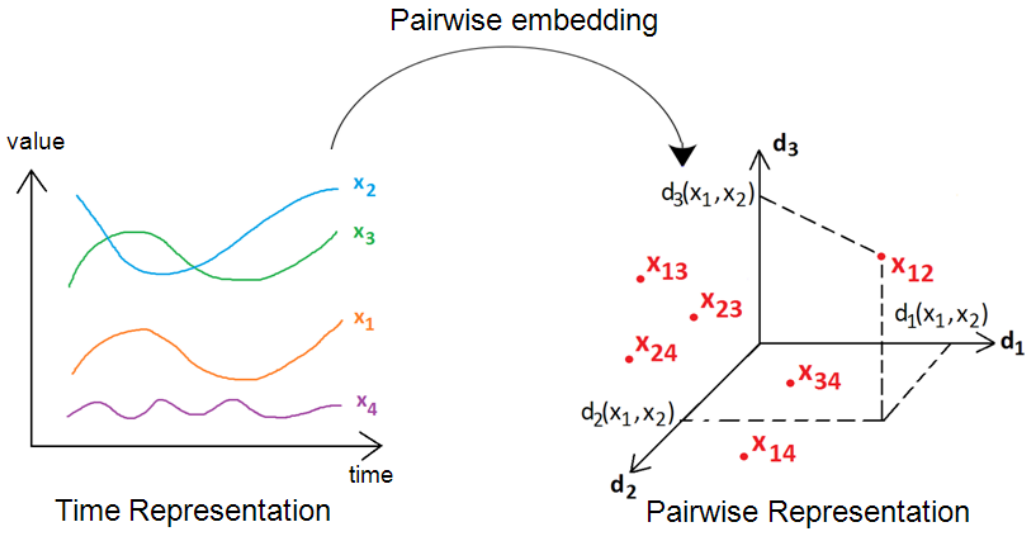


Figure 3.1: Example of embedding of time series $\mathbf{x}_i$ from the temporal space (left) into the pairwise space (right). In this example, a pair of time series $(\mathbf{x}_1, \mathbf{x}_2)$ is projected into the pairwise space as a vector $\mathbf{x}_{12}$ described by $p = 3$ basic metrics: $\mathbf{x}_{12} = [d_1(\mathbf{x}_1, \mathbf{x}_2), d_2(\mathbf{x}_1, \mathbf{x}_2), d_3(\mathbf{x}_1, \mathbf{x}_2)]^T$.

A combination function D of the metrics d_h can be seen as a function in this space. In the following, we propose first to use a linear combination of d_h: $D(\mathbf{x}_i, \mathbf{x}_j) = \sum_h w_h.d_h(\mathbf{x}_i, \mathbf{x}_j)$. For simplification purpose, we denote $D(\mathbf{x}_i, \mathbf{x}_j) = D(\mathbf{x}_{ij})$ and the pairwise notation gives:

$$D(\mathbf{x}_i, \mathbf{x}_j) = D(\mathbf{x}_{ij}) = \mathbf{w}^T \mathbf{x}_{ij} \tag{3.2}$$

where $\mathbf{w}$ is the vector of weights w_h: $\mathbf{w} = [w_1, \ldots, w_p]^T$.

3.1.2 Pairwise label

In the pairwise space, each vector $\mathbf{x}_{ij}$ can be labeled y_{ij} by following the rule: if $\mathbf{x}_i$ and $\mathbf{x}_j$ are similar, the vector $\mathbf{x}_{ij}$ is labeled -1; and $+1$ otherwise.

For classification problems, the concept of similarity between samples $\mathbf{x}_i$ and $\mathbf{x}_j$ is driven by the class label y_i and y_j in the original space:

$$y_{ij} = \begin{cases} -1 & \text{if } y_i = y_j \\ +1 & \text{if } y_i \neq y_j \end{cases} \tag{3.3}$$

For regression problems, each sample $\mathbf{x}_i$ is assigned to a continuous value y_i. Two approaches are possible to define the similarity concept. The first one discretizes the continuous space of values of the labels y_i to create classes. One possible discretization bins the label y_i into Q intervals as illustrated in Fig. 3.2. Each interval becomes a class which associated value can be set for example as the mean or median value of the interval. Then, the classification framework is used to define the pairwise label y_{ij}.

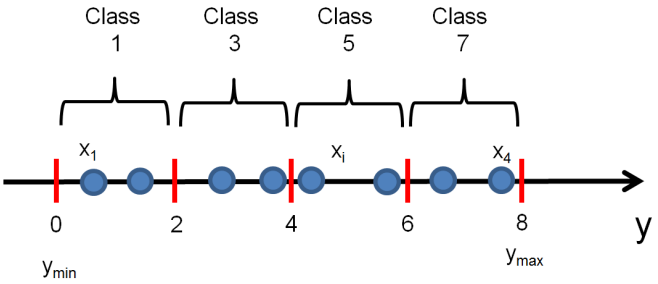


Figure 3.2: Example of discretization by binning a continuous label y into $Q = 4$ equal-length intervals. Each interval is associated to a unique class label. In this example, the class label for each interval is equal to the mean in each interval.

This approach may leads to border effects between the classes. For instance, two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ that are close to a frontier and that are on different sides of the border will be considered as different, as illustrated in Fig 3.3. Moreover, a new sample $\mathbf{x}_j$ will have its labels y_j assigned to a class and not a real continuous value.

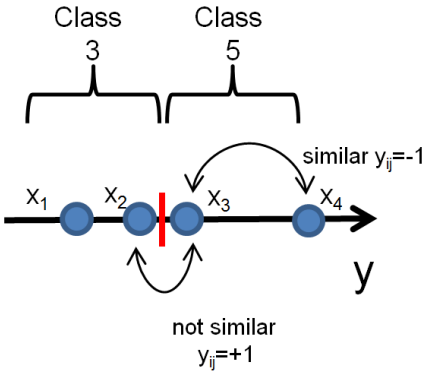


Figure 3.3: Border effect problems. In this example, $\mathbf{x}_2$ and $\mathbf{x}_3$ have closer value labels y_2 and y_3 than $\mathbf{x}_3$ and $\mathbf{x}_4$. However, with the discretization $\mathbf{x}_2$ and $\mathbf{x}_3$ don't belong to the same class and thus are consider as not similar.

The second approach considers the continuous value of y_i, computes a L1-norm between the labels $|y_i - y_j|$ and compare this value to a threshold ϵ. Geometrically, a tube of size ϵ around each value of y_i is built. Two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ are considered as similar if the absolute difference between their labels $|y_i - y_j|$ is lower than ϵ (Fig. 3.4):

$$y_{ij} = \begin{cases} -1 & \text{if } |y_i - y_j| \leq \epsilon \\ +1 & \text{otherwise} \end{cases} \tag{3.4}$$

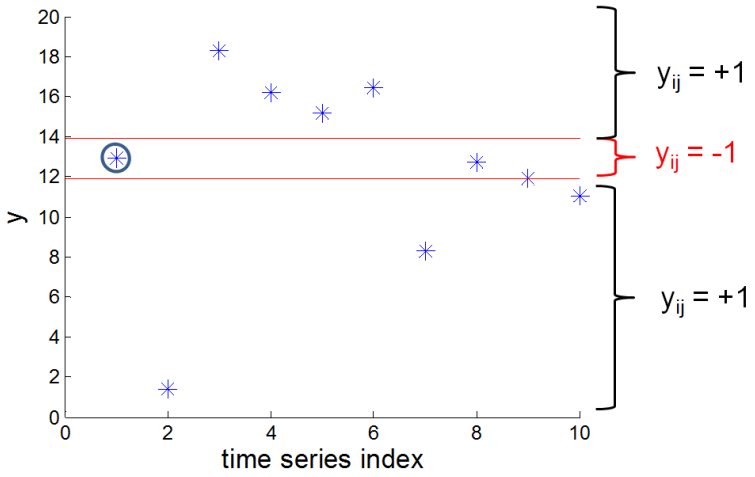


Figure 3.4: Example of pairwise label definition using an ϵ-tube (red lines) around the time series $\mathbf{x}_i$ (circled in blue). For, time series $\mathbf{x}_j$ that falls into the tube, the pairwise label is $y_{ij} = -1$ (similar) and outside of the tube, $y_{ij} = +1$ (not similar).

3.1.3 Interpretation in the pairwise space

The interpretation of the data in the pairwise space is particular since the pairwise space is not a standard Euclidean space. The interpretation in this space requires to be careful.

If $\mathbf{x}_{ij} = \mathbf{0}$ then $\mathbf{x}_j$ is identical to $\mathbf{x}_i$ according to all metrics d_h. The norm of the vector $\mathbf{x}_{ij}$ can be interpreted as a proximity measure: the lower the norm of $\mathbf{x}_{ij}$ is, the closer are the time series $\mathbf{x}_i$ and $\mathbf{x}_j$. Nevertheless, if two pairwise vectors $\mathbf{x}_{ij}$ and $\mathbf{x}_{kl}$ has their norms closed, it doesn't mean that the time series $\mathbf{x}_i$, $\mathbf{x}_j$, $\mathbf{x}_k$ and $\mathbf{x}_l$ are similar. Fig 3.5 shows an example of two pairwise vectors $\mathbf{x}_{ij}$ and $\mathbf{x}_{kl}$ that are close together in the pairwise space. However, in the temporal space, the time series $\mathbf{x}_1$ and $\mathbf{x}_3$ are not similar for example. It means that $\mathbf{x}_i$ is as similar to $\mathbf{x}_j$ as $\mathbf{x}_k$ is to $\mathbf{x}_l$.

A metric D that combines the p unimodal metrics $d_1, \ldots, d_p$ can be seen as a function of the pairwise space. It can be noticed that when the time series $\mathbf{x}_i$ are embedded in the pairwise, the information of their original class y_i is lost. Any multi-class problem is transformed in the pairwise space as a binary classification problem.

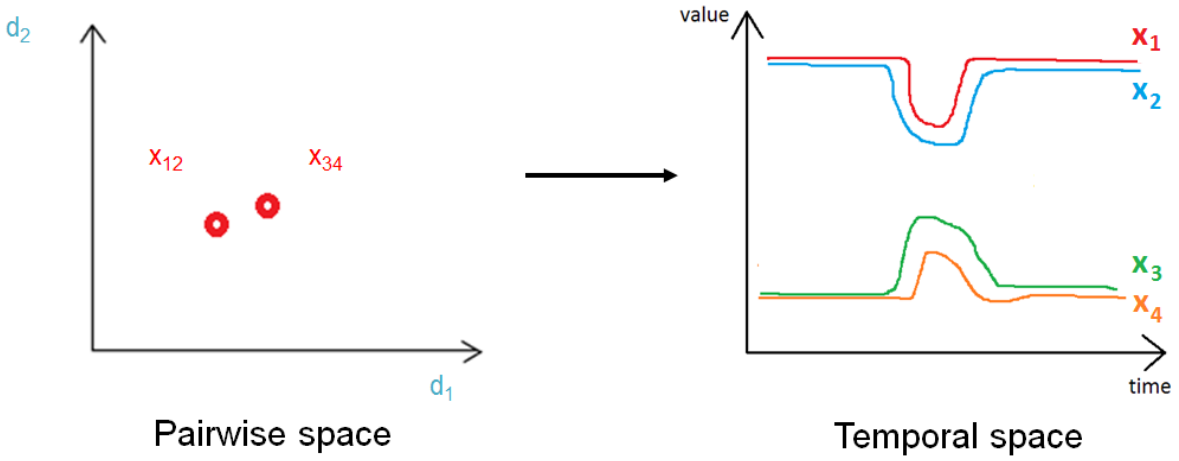


Figure 3.5: Example of two pairwise vectors $\mathbf{x}_{12}$ and $\mathbf{x}_{34}$ close in the pairwise space. However, the time series $\mathbf{x}_1$ and $\mathbf{x}_3$ are not similar in the temporal space.

In the next sections, we transpose the metric learning problem for large margin nearest neighbors in the pairwise space. We propose three formulations: Linear programming, Quadratic programming and SVM-based approach.

3.2 Linear Programming (LP) formalization

Our objective is to define a metric D as a linear combination of the unimodal metric d_h (Eq. 3.2). In the pairwise space, the metric D should:

- **pull** to the origin the k nearest neighbors pairs $\mathbf{x}_{ij}$ of same labels ($y_{ij} = -1$)
- **push** from the origin all the pairs $\mathbf{x}_{il}$ of different classes ($y_{il} = +1$)

Fig. 3.6 illustrates our idea. For each time series $\mathbf{x}_i$, we build the set of target pairs $\mathbf{x}_{ij}$ ($j \rightsquigarrow i$) and the set of pairs $\mathbf{x}_{il}$ of different class ($y_{il} = +1$). Then, we optimise the weight vector $\mathbf{w}$ so that the pairs $\mathbf{x}_{ij}$ are pulled to the origin and the pairs $\mathbf{x}_{il}$ are pushed from the origin.

Inspired from the Large Margin Nearest Neighbors (LMNN) framework proposed by Weinberger & Saul in Section 2.6.2, we transpose the metric learning problem into the pairwise space to learn a temporal metric D combining several unimodal metric d_h. In our problem, the optimal metric D is learned as the solution of a minimization problem, such that for each time series $\mathbf{x}_i$, it pulls its targets $\mathbf{x}_j$ and pushes all the samples $\mathbf{x}_l$ with a different label ($y_l \neq y_i$).

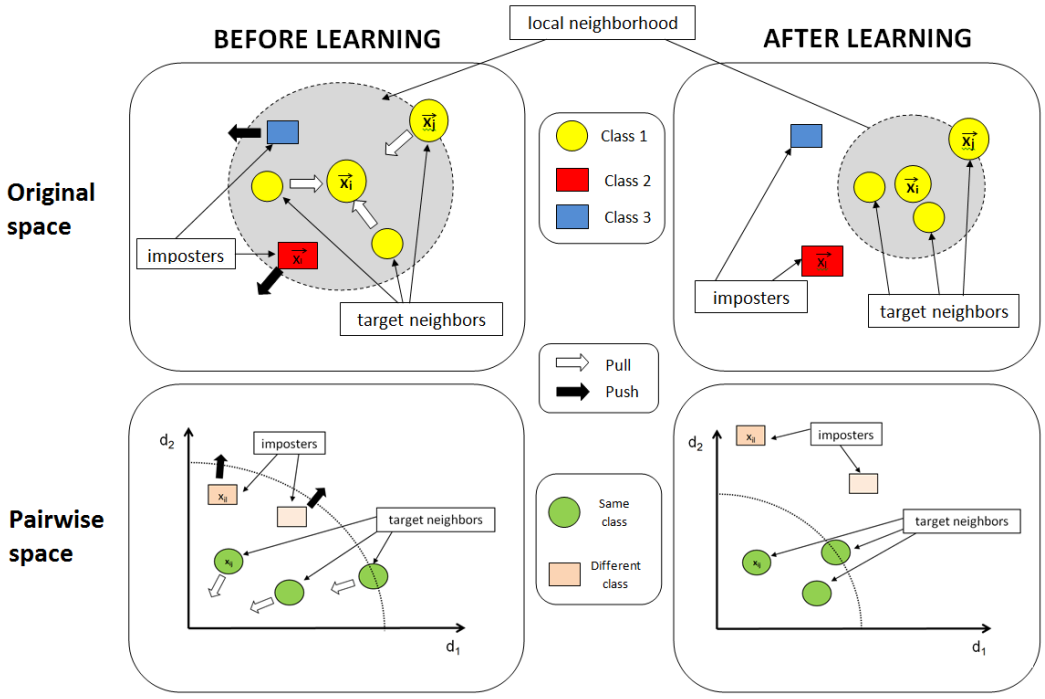


Figure 3.6: Geometric representation of the adaptation of metric learning problem from the original space (top) to the pairwise space (bottom) for a $k = 3$ target neighborhood of $\mathbf{x}_i$. Before learning (left), imposters $\mathbf{x}_l$ invade the targets perimeter $\mathbf{x}_j$. In the pairwise space, this is equivalent to have pairwise vectors $\mathbf{x}_{il}$ with a norm lower to some pairwise target $\mathbf{x}_{ij}$. The aim of metric learning is to push pairwise $\mathbf{x}_{il}$ (black arrow) and pull pairwise $\mathbf{x}_{ij}$ from the origin (white arrow).

The Time series Metric Learning (TML) problem in the pairwise space is formalized as:

$$\underset{D,\xi}{\operatorname{argmin}} \left\{ \underbrace{\sum_{i,j \rightsquigarrow i} D(\mathbf{x}_{ij})}_{pull} + C \underbrace{\sum_{i,j \rightsquigarrow i,l} \frac{1+y_{il}}{2}.\xi_{ijl}}_{push} \right\} \tag{3.5}$$

$$\text{s.t. } \forall j \rightsquigarrow i, y_l \neq y_i,$$

$$D(\mathbf{x}_{il}) - D(\mathbf{x}_{ij}) \geq 1 - \xi_{ijl} \tag{3.6}$$

$$\xi_{ijl} \geq 0 \tag{3.7}$$

where ξ_{ijl} are the slack variables and C, the trade-off between the pull and push costs. The proposed TML differs from LMNN in which the push term in TML considers all samples $\mathbf{x}_l$ with a different label from $\mathbf{x}_i$, whereas in LMNN, only the imposters are taken into consideration (those whose invade the target perimeter). Intuitively, this due to the fact that we do not want that samples $\mathbf{x}_l$ with a different class that were not at the beginning imposters, become imposters during the optimization process. By considering all the samples $\mathbf{x}_l$, we ensure that

at each step of the optimization process, if a sample $\mathbf{x}_l$ becomes an imposter, then it will violate the constraints in Eq. 3.7 and thus, its slack variables ξ_{ijl} will be penalized in the objective function (Eq. 3.5) :

- If $D(\mathbf{x}_{il}) < D(\mathbf{x}_{ij})$, then the pairs $\mathbf{x}_{il}$ is an imposter pair that invades the neighborhood of the target pairs $\mathbf{x}_{ij}$. The slack variable $\xi_{ijl} > 1$ will be penalized in the objective function (Eq. 3.5).
- If $D(\mathbf{x}_{il}) \geq D(\mathbf{x}_{ij})$ but $D(\mathbf{x}_{il}) \leq D(\mathbf{x}_{ij}) + 1$, the pair $\mathbf{x}_{il}$ is within the safety margin of the target pairs $\mathbf{x}_{ij}$. The slack variable $\xi_{ijl} \in [0; 1]$ will have a small penalization effect in the objective function (Eq. 3.5).
- If $D(\mathbf{x}_{il}) > D(\mathbf{x}_{ij}) + 1$, $\xi_{ijl} = 0$ and the slack variable has no effect in the objective function (Eq. 3.5).

By considering a linear combination of the unimodal distance d_h (Chapter 2): $D(\mathbf{x}_i, \mathbf{x}_j) = \sum_h w_h.d_h(\mathbf{x}_i, \mathbf{x}_j)$, optimizing the metric D is equivalent to optimizing the weight vector $\mathbf{w}$. Eqs. 3.5 and 3.6 leads to the TML primal formulation:

$$\underset{\mathbf{w},\xi}{\operatorname{argmin}} \left\{ \underbrace{||\mathbf{X}_{tar}^T\mathbf{w}||}_{pull} + C \underbrace{\sum_{i,j \rightsquigarrow i,l} \frac{1+y_{il}}{2}.\xi_{ijl}}_{push} \right\} \tag{3.8}$$

$$\text{s.t. } \forall j \rightsquigarrow i, y_l \neq y_i,$$

$$\mathbf{w}^T(\mathbf{x}_{il} - \mathbf{x}_{ij}) \geq 1 - \xi_{ijl} \tag{3.9}$$

$$\xi_{ijl} \geq 0 \tag{3.10}$$

where $\mathbf{X}_{tar}$ is a $p \times (k.N)$ matrix containing all targets $\mathbf{x}_{ij}$ and $||\mathbf{X}_{tar}^T\mathbf{w}||$ denotes the norm of the vector $\mathbf{X}_{tar}^T\mathbf{w}$. As in SVM, a L1 or L2 norm can be chosen. L1 norm will priviledge sparse solution of $\mathbf{w}$.
TML can be seen as a large margin problem in the pairwise space and parallels can be done with SVM. The "pull" term acts as a regularizer which aims to minimize the norm of $\mathbf{w}$. Similarly to SVM, mininimizing the norm of $\mathbf{w}$ is equivalent to maximizing the margin $\frac{1}{||\mathbf{w}||}$ between target pairs $\mathbf{x}_{ij}$ and pairs of different class $\mathbf{x}_{il}$.

3.3 Quadratic Programming (QP) formalization

The primal formulation of TML (Eqs. 3.8, 3.9 and 3.10) supposed that the metric D is a linear combination of the metrics d_h. The primal formulation being similar to the one of SVM, it can be derived into its dual form to obtain non-linear solutions for D. For that, we consider in the objective function (Eq. 3.8), the square of the L2-norm on $\mathbf{w}$ as the regularizer term, $\frac{1}{2}||\mathbf{X}_{tar}^T\mathbf{w}||_2^2$:

$$\underset{\mathbf{w},\xi}{\operatorname{argmin}} \left\{ \frac{1}{2}||\mathbf{X}_{tar}^T\mathbf{w}||_2^2 + C \sum_{i,j\rightsquigarrow i,l} \frac{1+y_{il}}{2}.\xi_{ijl} \right\} \tag{3.11}$$

s.t. $\forall j \rightsquigarrow i, y_l \neq y_i$,

$$\mathbf{w}^T(\mathbf{x}_{il} - \mathbf{x}_{ij}) \geq 1 - \xi_{ijl} \tag{3.12}$$

$$\xi_{ijl} \geq 0 \tag{3.13}$$

This formulation can be reduced to the minimization of the following Lagrange function $L(\mathbf{w}, \xi, \boldsymbol{\alpha}, \mathbf{r})$, consisting of the sum of the objective function (Eq. 3.11) and the constraints (Eqs. 3.12 and 3.13) multiplied by their respective Lagrange multipliers $\boldsymbol{\alpha}$ and $\mathbf{r}$:

$$\begin{aligned} L(\mathbf{w}, \xi, \boldsymbol{\alpha}, \mathbf{r}) =& \frac{1}{2}||\mathbf{X}_{tar}^T\mathbf{w}||_2^2 + C\sum_{ijl} \frac{1+y_{il}}{2}\xi_{ijl} - \sum_{ijl} r_{ijl}\xi_{ijl} \\ & - \sum_{ijl} \alpha_{ijl}\left(\mathbf{w}^T(\mathbf{x}_{il} - \mathbf{x}_{ij} - 1 + \xi_{ijl})\right) \end{aligned} \tag{3.14}$$

where $\alpha_{ijl} \geq 0$ and $r_{ijl} \geq 0$ are the Lagrange multipliers. At the minimum value of $L(\mathbf{w}, \xi, \boldsymbol{\alpha}, \mathbf{r})$, we assume the derivatives with respect to $\mathbf{w}$ and ξ_{ijl} are set to zero:

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{X}_{tar}^T\mathbf{X}_{tar}\mathbf{w} - \sum_{ijl} \alpha_{ijl}(\mathbf{x}_{il} - \mathbf{x}_{ij}) = 0$$

$$\frac{\partial L}{\partial \xi_{ijl}} = C - \alpha_{ijl} - r_{ijl} = 0$$

that leads to:

$$\mathbf{w} = (\mathbf{X}_{tar}\mathbf{X}_{tar}^T)^{-1} \sum_{ijl} \alpha_{ijl}(\mathbf{x}_{il} - \mathbf{x}_{ij}) \tag{3.15}$$

$$r_{ijl} = C - \alpha_{ijl} \tag{3.16}$$

Substituting Eq. 3.15 and 3.16 back into $L(\mathbf{w}, \xi, \boldsymbol{\alpha}, \mathbf{r})$ in Eq. 3.14, we get the TML dual formulation[1]:

$$\underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \left\{ \sum_{ijl} \alpha_{ijl} - \frac{1}{2}\sum_{ijl}\sum_{i'j'l'} \alpha_{ijl}\alpha_{i'j'l'}(\mathbf{x}_{il} - \mathbf{x}_{ij})^T(\mathbf{X}_{tar}\mathbf{X}_{tar}^T)^{-1}(\mathbf{x}_{i'l'} - \mathbf{x}_{i'j'}) \right\} \tag{3.17}$$

s.t. $\forall\ i, j \rightsquigarrow i$ and l s.t. $y_{il} = +1$:

$$0 \leq \alpha_{ijl} \leq C \tag{3.18}$$

[1] complete details of the calculations in Appendix D

For any new pair of samples $\mathbf{x}_{i'}$ and $\mathbf{x}_{j'}$, the resulting metric D writes:

$$D(\mathbf{x}_{i'j'}) = \mathbf{w}^T \mathbf{x}_{i'j'} \tag{3.19}$$

$$D(\mathbf{x}_{i'j'}) = \sum_{ijl} \alpha_{ijl} (\mathbf{x}_{il} - \mathbf{x}_{ij})^T (\mathbf{X}_{tar} \mathbf{X}_{tar}^T)^{-1} \mathbf{x}_{i'j'} \tag{3.20}$$

with $\mathbf{w}$ defined in Eq. 3.15. At the optimality, only the triplets $(\mathbf{x}_{il} - \mathbf{x}_{ij})$ with $\alpha_{ijl} > 0$ are considered as the support vectors. The direction $\mathbf{w}$ of the metric D is lead by these triplets. All other triplets have $\alpha_{ijl} = 0$ (non-support vector), and the metric D is independent from this triplets. If we remove some of the non-support vectors, the metric D remains unaffected. From the viewpoint of optimization theory, we can also see this from the Karush-Kuhn-Tucker (KKT) conditions: the complete set of conditions which must be satisfied at the optimum of a constrained optimization problem. At the optimum, the Karush-Kuhn-Tucker (KKT) conditions apply, in particular:

$$\alpha_{ijl} (\mathbf{w}^T (\mathbf{x}_{il} - \mathbf{x}_{ij}) - 1 + \xi_{ijl}) = 0$$

from which we deduce that either $\mathbf{w}^T (\mathbf{x}_{il} - \mathbf{x}_{ij}) > 1$ and $\alpha_{ijl} = 0$ (the triplet $(\mathbf{x}_{il} - \mathbf{x}_{ij})$ is a non-support vector), or $\mathbf{w}^T (\mathbf{x}_{il} - \mathbf{x}_{ij}) = 1 - \xi_{ijl}$ and $\alpha_{ijl} > 0$ (the triplet is a support vector). Therefore, the learned metric D is a combination of scalar products between new pairs $\mathbf{x}_{i'j'}$ and a few number of triplets $\mathbf{x}_{ijl}$ of the training set.

Extension to non-linear function of D

The above formula can extended to non-linear function for the metric D. The dual formulation in Eq. 3.17 only relies on the inner product $(\mathbf{x}_{i'l'} - \mathbf{x}_{i'j'})^T (\mathbf{X}_{tar} \mathbf{X}_{tar}^T)^{-1} (\mathbf{x}_{il} - \mathbf{x}_{ij})$. We can hence apply the kernel trick on Eqs. 3.19 and 3.20 to find non-linear solutions for D:

$$\begin{aligned} D(\mathbf{x}_{i'j'}) &= \mathbf{w}^T \phi(\mathbf{x}_{i'j'}) \\ D(\mathbf{x}_{i'j'}) &= \sum_{ijl} \alpha_{ijl} \phi(\mathbf{x}_{il} - \mathbf{x}_{ij}) \phi(\mathbf{x}_{i'j'}) \\ D(\mathbf{x}_{i'j'}) &= \sum_{ijl} \alpha_{ijl} K(\mathbf{x}_{il} - \mathbf{x}_{ij}; \mathbf{x}_{i'j'}) \end{aligned}$$

These equations suppose that the null vector $\mathbf{0}$ in the original space is transformed through the transformation ϕ into the null vector: $\phi(\mathbf{0}) = \mathbf{0}$ in the feature space. We recall that $D(\mathbf{x}_{ii} = \mathbf{0})$ is expected to be equal to zero (distinguishability property in Section 2.2). However, if the vectors $\mathbf{x}_{ij}$ are projected in a feature space by a transformation ϕ, it doesn't guarantee that $\phi(\mathbf{0}) = \mathbf{0}$. Fig. 3.7 illustrates the idea for a polynomial kernel in which $\phi(\mathbf{0}) = [0, 0, 0, 1]^T$. Thus, the metric measure needs to be computed in the feature space relatively to the projection of $\phi(\mathbf{0})$. This is done by adding a term $\mathbf{w}^T \phi(\mathbf{0})$ to Eqs. 3.19 and 3.20:

$$D(\mathbf{x}_{i'j'}) = \mathbf{w}^T\phi(\mathbf{x}_{i'j'}) - \mathbf{w}^T\phi(\mathbf{0}) \tag{3.21}$$

$$D(\mathbf{x}_{i'j'}) = \sum_{ijl}\alpha_{ijl}\phi(\mathbf{x}_{il}-\mathbf{x}_{ij})\phi(\mathbf{x}_{i'j'}-\mathbf{x}_{ij}) - \sum_{ijl}\alpha_{ijl}\phi(\mathbf{x}_{il}-\mathbf{x}_{ij})\phi(\mathbf{0}-\mathbf{x}_{ij}) \tag{3.22}$$

$$D(\mathbf{x}_{i'j'}) = \sum_{ijl}\alpha_{ijl}K\left(\mathbf{x}_{il}-\mathbf{x}_{ij};\mathbf{x}_{i'j'}-\mathbf{x}_{ij}\right) - \sum_{ijl}\alpha_{ijl}K\left(\mathbf{x}_{il}-\mathbf{x}_{ij};\mathbf{0}-\mathbf{x}_{ij}\right) \tag{3.23}$$

where $\mathbf{0}$ denotes the null vector. The resulting metric D is made of two terms. The first one, $\mathbf{w}^T\phi(\mathbf{x}_{i'j'})$, is the metric measure for a new pair $\mathbf{x}_{i'j'}$. The second term, $\mathbf{w}^T\phi(\mathbf{0})$, adapts the metric measure relatively to the origin point.

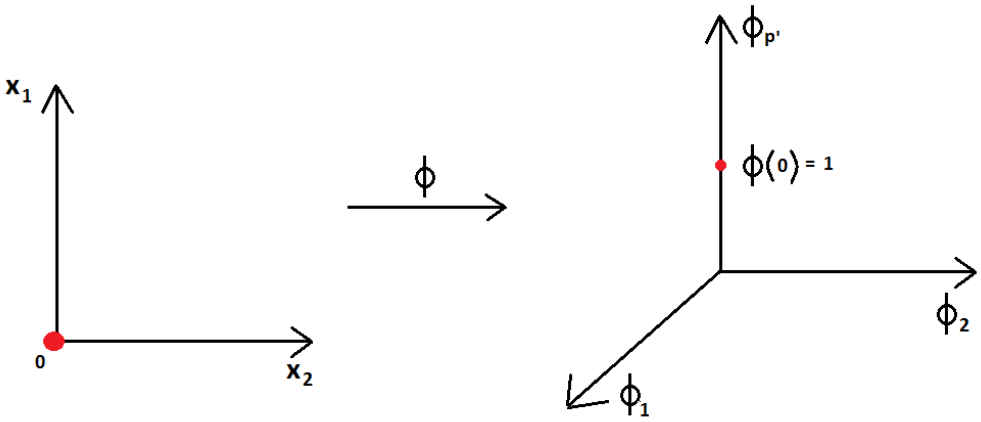


Figure 3.7: Illustration of samples in $\mathbb{R}^2$. The transformation ϕ for a polynomial kernel $K(\mathbf{x}_i,\mathbf{x}_j) = (\mathbf{x}_i^T\mathbf{x}_j + c)^d$ with $c = 1$ and $d = 2$ can be written explicitly: $\phi(\mathbf{x}_i) = [x_{i1}^2, x_{i2}^2, \sqrt{2}x_{i1}x_{i2}, 1]^T$. The origin point $\mathbf{x}_i = [0,0]^T$ is projected in the Hilbert space as $\phi(\mathbf{x}_i = \mathbf{0}) = [0,0,0,1]^T$.

However, to define proper metrics that respects the properties of metrics (Section 2.2), specific kernels must be used. Our work don't propose any solutions to this problem but open the field for new research on this topic.

3.4 Support Vector Machine (SVM) approximation

3.4.1 Motivations

Many parallels have been studied between Large Margin Nearest Neighbors (LMNN) and SVM (Section 2.6.3). Similarly, the proposed TML approach can be linked to SVM: both are convex optimization problem based on a regularized and a loss term. SVM is a well known framework: its has been well implemented in many libraries (e.g., LIBLINEAR [FCH08] and LIBSVM [HCL08]), well studied for its generalization properties and extension to non-linear solutions.

Motivated by these advantages, we propose to solve the TML problem by solving a similar

SVM problem. Then, we can naturally extend TML approach to find non-linear solutions for the metric D thanks to the 'kernel trick'. In the following, we show the similarities and the differences between LP/QP and SVM formulation.

For a time series $\mathbf{x}_i$, we define the set of pairs $\mathbf{X}_{pi} = \{(\mathbf{x}_{ij}, y_{ij}) \text{ s.t. } j \rightsquigarrow i \text{ or } y_{ij} = +1\}$. It corresponds for a time series $\mathbf{x}_i$ to the set of pairs with target samples $\mathbf{x}_j$ (k nearest samples of same labels $j \rightsquigarrow i$) or samples $\mathbf{x}_l$ that has a different label from $\mathbf{x}_i$ ($y_l \neq y_i$) . Identity pairs $\mathbf{x}_{ii}$ are not considered. We refer to $\mathbf{X}_p = \bigcup_i \mathbf{X}_{pi}$ and consider the following standard soft-margin weighted SVM problem on $\mathbf{X}_p$ [2]:

$$\begin{aligned} &\underset{\mathbf{w},b,\xi}{\operatorname{argmin}} \left\{ \frac{1}{2}||\mathbf{w}||_2^2 + C \sum_{i,j,y_{ij}=-1} p_i^- \xi_{ij} + C \sum_{i,j,y_{ij}=+1} p_i^+ \xi_{ij} \right\} \\ &\text{s.t. } y_{ij}(\mathbf{w}^T \mathbf{x}_{ij} + b) \geq 1 - \xi_{ij} \end{aligned} \tag{3.24}$$

where p_i^- and p_i^+ are the weight factors for target pairs and pairs of different class.

We show in the following that solving the SVM problem in Eq. 3.24 for $\mathbf{w}$ and b solves a similar TML problem in Eq. 3.11 for $D(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{2}(\mathbf{w}^T \mathbf{x}_{ij} + b)$. If we set p_i^+ being the half of the number of targets of $\mathbf{x}_i$ and p_i^-, the half of the number of time series L of a different class than $\mathbf{x}_i$:

$$p_i^+ = \frac{k}{2} = \sum_{j \rightsquigarrow i} \frac{1}{2} \tag{3.25}$$

$$p_i^- = \frac{L}{2} = \frac{1}{2} \sum_l \frac{1 + y_{il}}{2} \tag{3.26}$$

3.4.2 Similarities and differences in the constraints

First, we recall the SVM constraints in Eq. 3.24:

$$y_{ij}(\mathbf{w}^T \mathbf{x}_{ij} + b) \geq 1 - \xi_{ij}$$

These constraints can be split into two sets of constraints:

$$\begin{aligned} -(\mathbf{w}^T \mathbf{x}_{ij} + b) &\geq 1 - \xi_{ij} \quad \text{(same class: } y_{ij} = -1) \\ (\mathbf{w}^T \mathbf{x}_{il} + b) &\geq 1 - \xi_{il} \quad \text{(different classes: } y_{ij} = +1) \end{aligned}$$

By defining $D(\mathbf{x}_{ij}) = \frac{1}{2}(\mathbf{w}^T \mathbf{x}_{ij} + b)$, this leads to:

$$\begin{aligned} -D(\mathbf{x}_{ij}) &\geq \frac{1}{2} - \frac{\xi_{ij}}{2} \\ D(\mathbf{x}_{il}) &\geq \frac{1}{2} - \frac{\xi_{il}}{2} \end{aligned}$$

[2] the SVM formulation below divides the loss part into two terms similarly to asymetric SVM

By summing each constraint two by two, this set of constraints implies the following set of constraints:

$$\begin{cases} \bullet \forall i,j,k,l \text{ such that } y_{ij} = -1, \text{ and } y_{kl} = +1, i \neq j \text{ and } i \neq k : \\ D(\mathbf{x}_k, \mathbf{x}_l) - D(\mathbf{x}_i, \mathbf{x}_j) \geq 1 - \frac{\xi_{kl}+\xi_{ij}}{2} \\ \bullet \forall i,j,l \text{ such that } y_{ij} = -1, \text{ and } y_{il} = +1, i \neq j : \\ D(\mathbf{x}_i, \mathbf{x}_l) - D(\mathbf{x}_i, \mathbf{x}_j) \geq 1 - \frac{\xi_{il}+\xi_{ij}}{2} \end{cases} \tag{3.27}$$

By defining $\xi_{ijl} = \frac{\xi_{ij}+\xi_{il}}{2}$, the second constraint in Eq. 3.27 from the SVM formulation is the same as the constraints in the TML formulation in Eq. 3.12.

However, an additional set of constraints is present in the SVM formulation (first set of constraints in Eq. 3.27) and not in the proposed TML. Geometrically, this can be interpreted as superposing the neighborhoods of all samples $\mathbf{x}_i$, making the union of all of their target sets $\mathbf{X}_{pi}$, and then pushing away all imposters $\mathbf{x}_{il}$ from this resulting target set. This is therefore creating "artificial imposters" $\mathbf{x}_{kl}$ that don't violate the local target space of sample $\mathbf{x}_k$, but are still considered as imposters because they invade the target of sample $\mathbf{x}_i$ (because of the neighborhoods superposition) (Figure 3.8). This is more constraining in the SVM resolution for the resulting metric D especially if the neighborhoods have different spread.

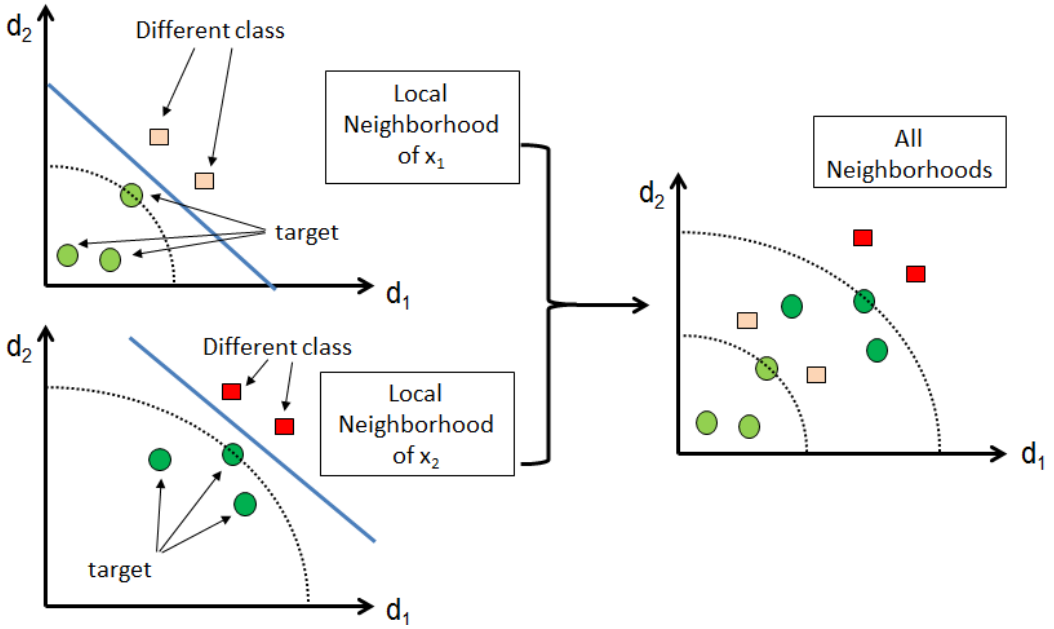


Figure 3.8: Geometric representation of the neighborhood of $k = 3$ for two time series $\mathbf{x}_1$ and $\mathbf{x}_2$ (left). For each neighborhood, time series of different class are represented by a square and the margin by a blue line. Taking each neighborhood separately, the problem is linearly separable (LP/QP formulation). By combining the two neighborhoods (SVM formulation), the problem is no more linearly separable and in this example, the time series of different class of $\mathbf{x}_1$ (orange square) are "artificial imposters" of $\mathbf{x}_2$.

3.4.3 Similarities and differences in the objective function

Mathematically, from Eq. 3.25, we write:

$$\begin{aligned}\sum_{i,l,y_{il}=+1} p_i^+ \xi_{il} &= \sum_{il} p_i^+ \frac{1+y_{il}}{2} \xi_{il} \\ &= \sum_{il} \left(\sum_{j \rightsquigarrow i} \frac{1}{2} \right) \frac{1+y_{il}}{2} \xi_{il} \\ &= \frac{1}{2} \sum_{i,j \rightsquigarrow i,l} \frac{1+y_{il}}{2} \xi_{il} \end{aligned} \tag{3.28}$$

And from Eq. 3.26, we write:

$$\begin{aligned}\sum_{i,j,y_{ij}=-1} p_i^- \xi_{ij} &= \sum_{i,j \rightsquigarrow i} p_i^- \xi_{ij} \\ &= \sum_{i,j \rightsquigarrow i} \left(\frac{1}{2} \sum_{l} \frac{1+y_{il}}{2} \right) \xi_{ij} \\ &= \frac{1}{2} \sum_{i,j \rightsquigarrow i,l} \frac{1+y_{il}}{2} \xi_{ij} \end{aligned} \tag{3.29}$$

By replacing Eqs. 3.28 and 3.29 back into Eq. 3.24, the objective function becomes:

$$\begin{aligned}&\min_{\mathbf{w},\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C \sum_{i,j \rightsquigarrow i,l} \frac{1+y_{il}}{2} \frac{\xi_{ij}+\xi_{il}}{2} \\ &\min_{\mathbf{w},\xi} \underbrace{\frac{1}{2}\mathbf{w}^T\mathbf{w}}_{Regularization} + \underbrace{C \sum_{i,j \rightsquigarrow i,l} \frac{1+y_{il}}{2} \xi_{ijl}}_{Loss} \end{aligned} \tag{3.30}$$

Even if the loss part (push cost) is the same for both objective functions, the regularization part (pull cost) is different. In the SVM formulation (Eq. 3.30), the regularization part tends to minimize the norm of $\mathbf{w}$ whereas in TML (Eq. 3.11), it tends to minimize the norm of $\mathbf{w}$ after a linear transformation through $\mathbf{X}_{tar}$. This transformation can be interpreted as a Mahalanobis norm in the pairwise space with $\mathbf{M} = \mathbf{X}_{tar}\mathbf{X}_{tar}^T$. Nevertheless, both have the same objective: improve the conditioning of the problem by enforcing solutions with small norms. In practice, even with these differences, the SVM provides suitable solutions for our time series metric learning problem.

3.4.4 Geometric interpretation

Michèle pense que l'état, cette section est dure à comprendre. D'après Michèle, il faut 1) soit prendre + de place pour expliquer la signification géométrique 2) ou soit ne pas mettre cette partie car étant compliquée, cela pourrait nuire au lecteur. Qu'en penses tu Ahlame?

In this section, we give a geometric understanding of the differences between LP/QP resolution (left) and SVM-based resolution (right). Fig. 3.10 shows the Linear Programming (LP) and SVM resolutions of a k-NN problem with $k = 2$ neighborhoods.

For LP, the problem is solved for each neighborhood (blue and red) independently as shown in Fig. 3.9. We recall that LP/QP resolutions, support vectors are triplets of time series made of a target pair $\mathbf{x}_{ij}$ and a pair of different classes $\mathbf{x}_{il}$ (black arrows). Support vectors represent triplet which resulting distance $D(\mathbf{x}_{ij}, \mathbf{x}_{il})$ are the lowest. The optimization problem tends to maximize the margin between these triplets. The global solution (Fig. 3.10 (left)) is a compromise of all of the considered margins. In this case, the global margin is equal to one of the local margin. Note that the global LP solution is not always the same as the best local solution. For SVM-based resolution (Fig. 3.10 (right)), the problem involves all pairs and the margin is optimized so that pairs $\mathbf{x}_{ij}$ and $\mathbf{x}_{il}$ are globally separated.

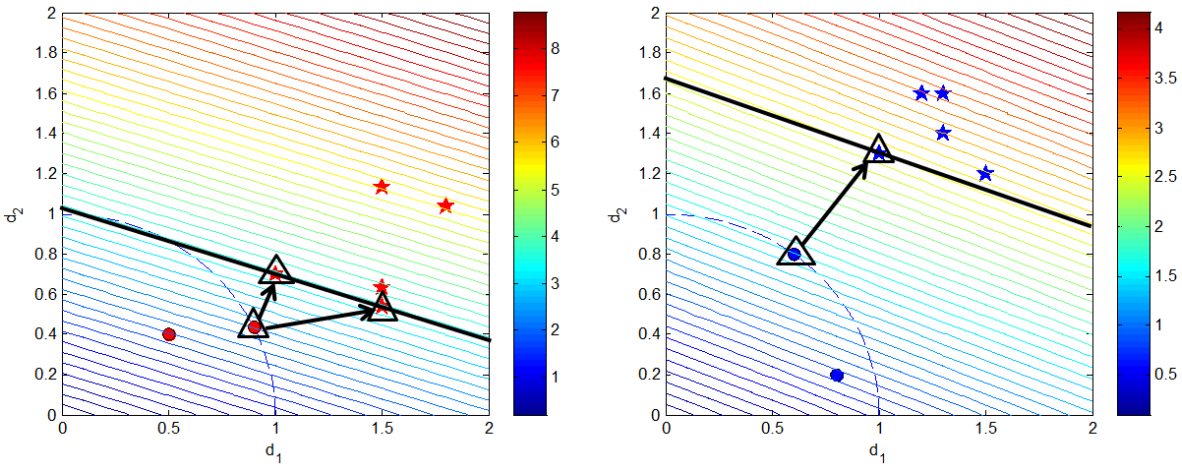


Figure 3.9: Solutions found by solving the LP problem for $k = 2$ neighborhood. Positive pairs (different classes) are indicated in stars and negative pairs (target pairs) are indicated in circle. Red and blue lines shows the margin when solving the problem for each neighborhood (red and blue points) separately. Support vector are indicated in black triangles: in the red neighborhood (left), 2 support vectors are retained and in the blue neighborhood (right), only one support vector is necessary.

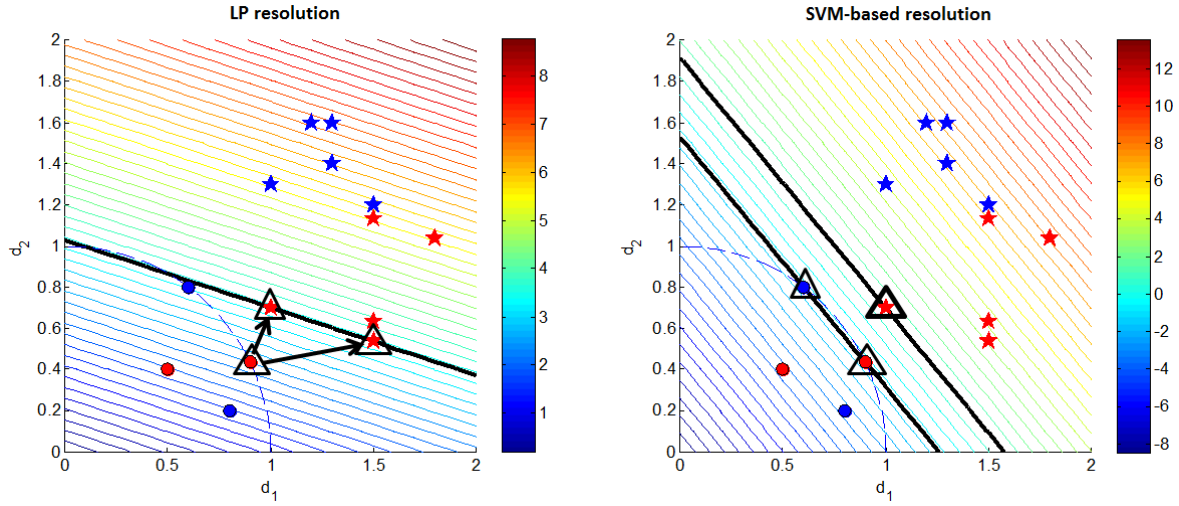


Figure 3.10: Solutions found by solving the LP problem (left) and the SVM problem (right). The global margin is indicated in black and the metric is represented in color levels. Support vectors made of triplets are indicated in black triangles. For the SVM, the black lines indicates the SVM canonical hyperplan where the support vector lies (black triangles).

3.5 Conclusion of the chapter

To learn a combined metric D from several unimodal metrics d_h that optimizes the k-NN performances, we first proposed a new space representation, the pairwise space where each pair of time series is projected as a vector described the unimodal metrics. Then, we propose three formalizations of our metric learning problem: Linear Programming, Quadratic Programming, SVM-based approximation.

In the following, we consider the SVM-based approximation because SVM framework is well known and well implementated. In the next chapter, we give the details of the steps of our proposed algorithm: Multi-modal and Multi-scale Time series Metric Learning (M^2TML).

Chapter 4

Multi-modal and Multi-scale Time series Metric Learning (M^2TML) implementation

Sommaire

In this chapter, we present the steps of our proposed algorithm referred as Multi-modal and Multi-scale Time series Metric Learning (M^2TML).
First, we introduce the multi-scale comparison concept for time series. Then, we present the steps to learn a Multi-modal and Multi-scale Time series metric for a robust k-Nearest Neighbor classifier: projection in the pairwise space, neighborhood construction and scaling, SVM-based metric learning resolution, and definition of the dissimilarity measure. We conclude by extending the algorithm M^2TML for multivariate and regression problems.

4.1 Multi-scale approach

In some applications, time series may exhibit similarities among the classes based on local patterns in the signal. Fig. 4.1 illustrates a toy example from the BME dataset in which time series of different classes seems to be similar on a global scale. However, at a more locally scale, a characteristic upward bell at the beginning or at the end of the time series allows to differentiate the classe B (upward bell at the beginning) from the class E (upward bell at the end). Also, in massive time series datasets, computing the metric on all time series elements

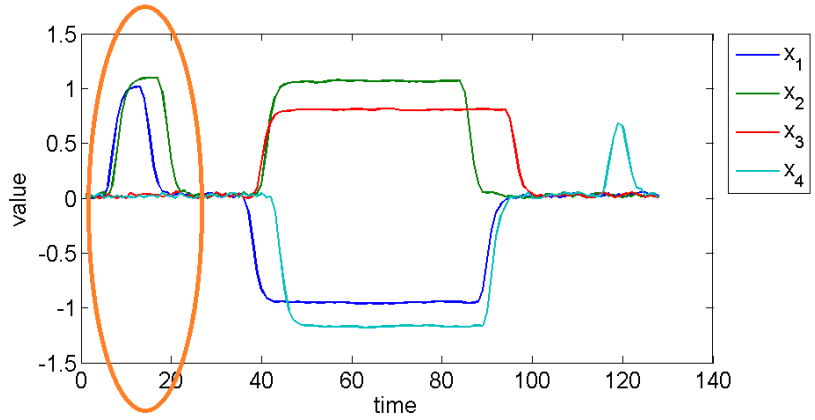


Figure 4.1: Example of 4 time series from the BME dataset, made of 3 classes : Begin, Middle and End. The 'Up' class has a characteristic bell at the beginning of the time series. The 'End' class has a characteristic bell at the end of the time series. The 'Middle' class has no characteristic bell. Orange circle show the region of interest of these bells for the class 'Begin'. This region is local and standard global metric fails to show these characteristics.

x_{it} might become time consuming. Computing the metric on a smaller part of the signal and not all the time series elements x_{it} makes the metric computation faster.

Localizing patterns of interest in huge time series datasets has become an active area of search in many applications including diagnosis and monitoring of complex systems, biomedical data analysis, and data analysis in scientific and business time series . A large number of methods have been proposed covering the extraction of local features from temporal windows [BC94] or the matching of queries according to a reference sequence [FRM94]. We focus on the computation of "local metrics".

ref

It can be noted that the distance measures (amplitude-based d_A[1], frequential-based d_F, behavior-based d_B) in Eqs. 2.1, 2.4 and 2.6 implies systematically the total time series elements x_{it} and thus, restricts the distance measures to capture local temporal differences. In our work, we provide a multi-scale framework for time series comparison using a hierarchical structure. Many methods exist in the literature such as the sliding window or the dichotomy . We detailed here the latter one.

ref

A multi-scale description can be obtained by repeatedly segmenting a time series expressed at a given temporal scale to induce its description at a more locally level. Many approaches have been proposed assuming fixed either the number of the segments or their lengths. In our work, we consider a binary segmentation at each level. Let $I = [a; b]$ be a temporal interval of size $(b - a)$. The interval I is decomposed into two equal overlaped intervals I_L (left interval) and I_R (right interval). A parameter α that allows to overlap the two intervals I_L and I_R, covering discriminating subsequences in the central region of I (around $\frac{b-a}{2}$):

[1]We recall that d_A is the Euclidean distance d_E in our work.

$$I = [a; b] \tag{4.1}$$
$$I_L = [a; a + \alpha(b - a)] \tag{4.2}$$
$$I_R = [a - \alpha(b - a); b] \tag{4.3}$$

For $\alpha = 0.6$, the overlap covers 10% of the size of the interval I. Then, the process is repeated on the intervals I_L and I_R. We obtain a set of intervals I_s illustrated in Fig. 4.2.

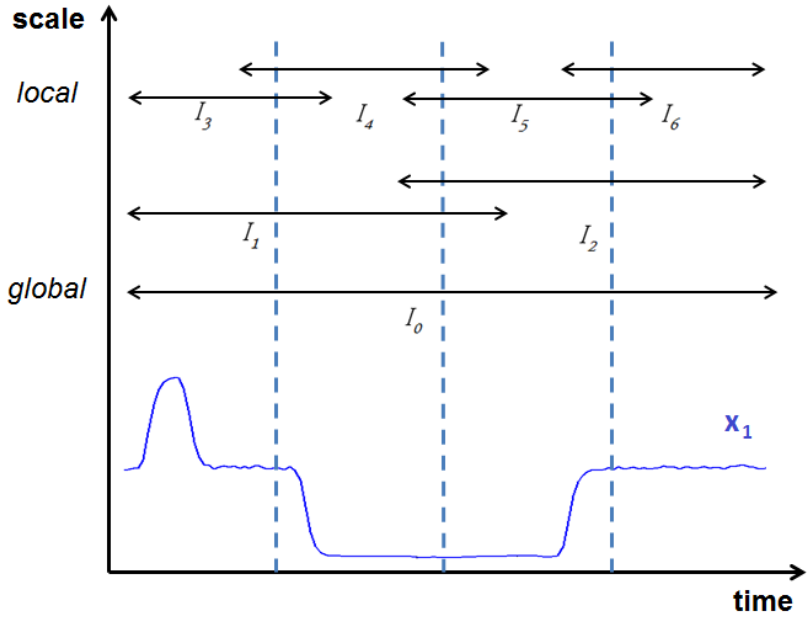


Figure 4.2: Multi-scale amplitude-based measures d_A^{Is}

A multi-scale description is obtained on computing the usual time series metrics (d_A, d_B, d_F) on the resulting segments I_s. For a multi-scale amplitude-based comparison based on binary segmentation, Fig. 4.2 shows the set of involved amplitude-based measures d_A^{Is}:

$$d_A^{Is}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{t \in Is} (x_{it} - x_{jt})^2} \tag{4.4}$$

The local behaviors- and frequential- based measures d_B^{Is} and d_F^{Is} are obtained similarly.

4.2 Projection in the pairwise space

Let $\{\mathbf{x}_i; y_i\}_{i=1}^n$ be n time series $\mathbf{x}_i \in \mathbb{R}^Q$ of length Q and of class label y_i. Let $d_1, \ldots, d_p$ be the set of multi-modal and multi-scale dissimilarity measures $d_A^{I_s}$, $d_B^{I_s}$ and $d_F^{I_s}$ related to

segments I_s of several temporal scales, as described in Section 4.1.

Projection in the pairwise space
We note ψ an embedding function that maps each pair of time series $(\mathbf{x}_i; \mathbf{x}_j)$ to a vector $\mathbf{x}_{ij}$ in a dissimilarity space $\mathbb{R}^p$ whose dimensions are the dissimilarities $d_1, \ldots, d_p$ as explained in Chapter 3:

$$\begin{aligned} \psi : \mathbb{R}^Q \times \mathbb{R}^Q &\to \mathscr{E} \\ (\mathbf{x}_i; \mathbf{x}_j) &\to \mathbf{x}_{ij} = [d_1(\mathbf{x}_i; \mathbf{x}_j), \ldots, d_p(\mathbf{x}_i; \mathbf{x}_j)]^T \end{aligned} \tag{4.5}$$

We cast the problem of learning a multi-modal and multiscale temporal metric as learning the metric D a combination function of $d_1, \ldots, d_p$ in the pairwise space $\mathscr{E}$:

$$D = f(d_1, \ldots, d_p) \tag{4.6}$$

The learning process is guided by local constraints to ensure dissimilarities between neighbors of a same class (i.e. $y_{ij} = -1$) lower than the dissimilarity between neighbors of different classes ($y_{ij} = +1$). The learned metric D should satisfy, in addition, the properties of a dissimilarity measure, i.e. positivity ($D(\mathbf{x}_{ij}) \geq 0$), distinguishability ($D(\mathbf{x}_{ij}) = 0, \mathbf{x}_i = \mathbf{x}_j$) and symmetry ($D(\mathbf{x}_{ij}) = D(\mathbf{x}_{ji})$).

Pairwise space normalization
This operation is performed to scale the data within the pairwise space and ensure comparable ranges for the p basic metrics d_h. In our experiment, we use dissimilarity measures with values in $[0; +\infty[$. Therefore, we propose to Z-normalize their log distributions.

In the following we detail the three main steps of the proposed solution. First, the neighborhood for each time series $\mathbf{x}_i$ is built to construct the pairwise training set $\{\mathbf{x}_{ij}, y_{ij}\}$. Secondly, an SVM is operated in the pairwise space $\mathscr{E}$ to learn a direction (i.e. weights) that discriminates positive ($y_{ij} = +1$) from negative ($y_{ij} = -1$) pairs. Thirdly, an exponential transformation of the projected norm on the learned discriminative direction allows to induce a dissimilarity measure that satisfy all required conditions, as well as homogeneous neighborhoods for a robust k-NN classification.

4.3 Neighborhood construction and scaling

The metric learning problem aims to learn a metric D that pulls the k nearest neighbors (targets) while pushing the time series of different classes. Thus, the preliminary step defines the target pairs. For that, an initial distance is necessary to build the neighborhood.
Let $\mathbf{x}_{ij} \in \mathbb{R}^p$ $i, j \in \{1, \ldots, n\}$ be a set of samples into the pairwise space $\mathscr{E}$ as described in Eq. 4.5. For each time series $\mathbf{x}_i$, we denote X_i^+ the set of **positive pairs** $\mathbf{x}_{ij}$ such that $y_{ij} = +1$ (i.e. the time series $\mathbf{x}_i$ and $\mathbf{x}_j$ has different class label $y_j \neq y_i$). Similarly, we denote X_i^- the

set of **negative pairs** $\mathbf{x}_{ij}$ such that $y_{ij} = -1$ (i.e. the time series $\mathbf{x}_i$ and $\mathbf{x}_j$ has the same class label $y_j = y_i$):

$$X_i^- = \{\mathbf{x}_{ij}, y_{ij} = -1\} \quad \text{(same class)} \tag{4.7}$$
$$X_i^+ = \{\mathbf{x}_{ij}, y_{ij} = +1\} \quad \text{(different classes)} \tag{4.8}$$

A L2 norm is used as an initial metric to define the positive and negative sets:

$$||\mathbf{x}_{ij}||_2 = \sqrt{\sum_{h=1}^{p} \left(d_h(\mathbf{x}_i, \mathbf{x}_j)\right)^2} \tag{4.9}$$

The **target set** X_i^{-*} is a subset of the negative set X_i^- of pairs $\mathbf{x}_{ij}$ such that the time series $\mathbf{x}_j$ are the k-nearest neighbors of $\mathbf{x}_i$, denoted $j \rightsquigarrow i$:

$$X_i^{-*} = \{\mathbf{x}_{ij}, y_{ij} = -1\} \quad \text{s.t. } j \rightsquigarrow i \tag{4.10}$$

The k nearest neighbors of a sample $\mathbf{x}_i$, denoted $\mathbf{x}_j$ ($j \rightsquigarrow i$), are defined in the pairwise space $\mathscr{E}$ by the k-th lowest norm $||\mathbf{x}_{ij}||_2$ negative pairs. Similarly, the **imposter set** X_i^{+*} is a subset of the positive set X_i^+ of pairs $\mathbf{x}_{il}$ such that the time series $\mathbf{x}_l$ is an imposter of $\mathbf{x}_i$, denoted $l \nrightarrow i$. It corresponds the pairs $\mathbf{x}_{il}$ that have a L2 norm lower that the L2 norm of the k-th nearest neighbor:

$$X_i^{+*} = \{\mathbf{x}_{il}, y_{il} = +1\} \quad \text{s.t. } l \nrightarrow i \tag{4.11}$$

To build the pairwise training set X_p, three solutions are proposed, illustrated in Fig 4.3:

1. **k-NN vs impostors**: it corresponds to the union for all $\mathbf{x}_i$ of the set of target set and impostor set:
$$X_p = X_i^{-*} \cup X_i^{+*} \tag{4.12}$$

2. **k-NN vs all**: it corresponds to the union for all $\mathbf{x}_i$ of the set of target set and positive set. It ensures that no pairs $\mathbf{x}_{il}$ of different classes will invade the target neighborhood during the learning process:
$$X_p = X_i^{-*} \cup X_i^{+} \tag{4.13}$$

3. **m-NN$^+$ vs m-NN$^-$**: it corresponds to the union for all $\mathbf{x}_i$ of the set of the m-nearest neighbors of the same class, denoted m-NN$^+$, and the m-nearest neighbor of $\mathbf{x}_i$ of a different class ($y_j \neq y_i$), denoted m-NN$^-$. For a k-NN classifier, by considering larger neighborhoods with $m = \alpha k (\alpha > 1)$ one includes more variability to generalize better the obtained solution:
$$X_p = m\text{-NN}^+ \cup m\text{-NN}^- \tag{4.14}$$

In our experiment, we use the m-NN$^+$ vs m-NN$^-$ strategy for better generalization of the solution compared to k-NN vs impostors strategy and for faster solutions compared to k-NN vs all strategy. Note that in m-NN$^+$ vs m-NN$^-$ strategy, the set of positive and negative pairs

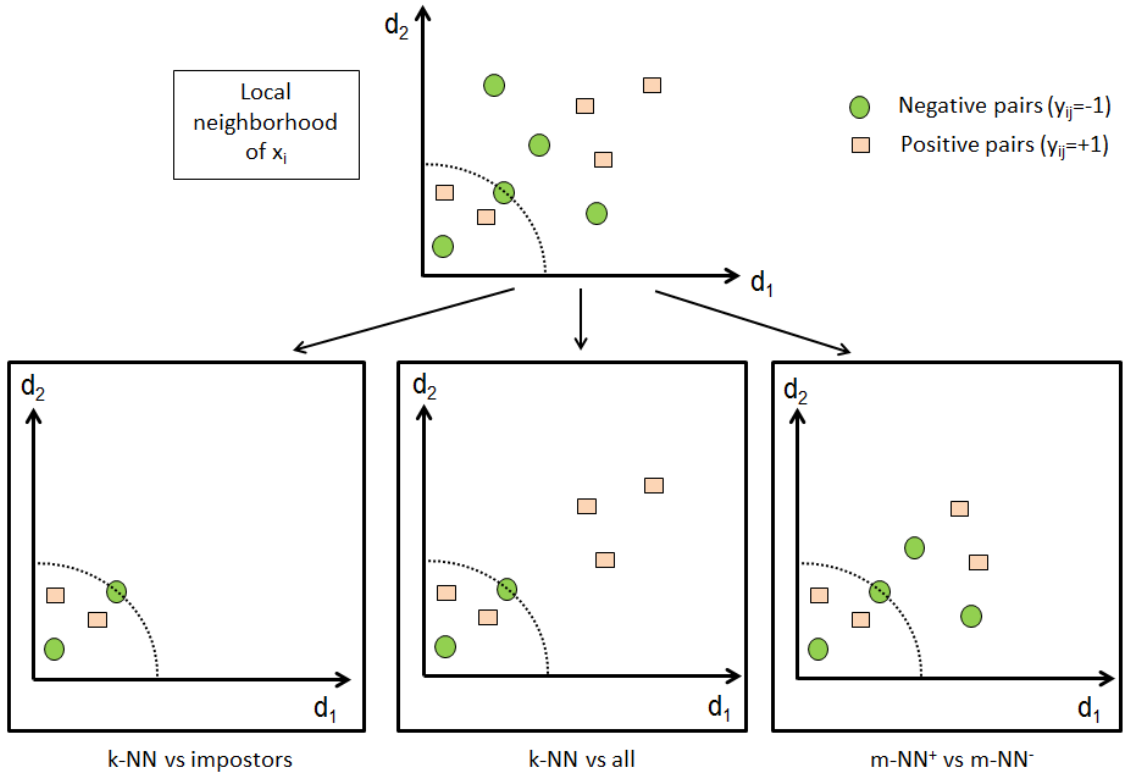


Figure 4.3: Example of a k-NN problem with $k = 2$. 3 different strategies (bottom) for pairwise training set X_p construction from the embedding of time series $\mathbf{x}_i$ in the pairwise space (top): k-NN vs impostor strategy (left), k-NN vs all strategy (middle) and m-NN$^+$ vs m-NN$^-$ (right) with $m = 4$.

is balanced.

Neighborhood scaling

Let r_i be the radius associated to $\mathbf{x}_i$ corresponding to the maximum norm of its m-th nearest neighbor of same class in m-NN$^-$:

$$r_i = \max_{\mathbf{x}_{ij} \in m\text{-NN}^-} ||\mathbf{x}_{ij}||_2 \tag{4.15}$$

As explained in Chapter 3, Section 3.4.3, there exists an heterogeneity in the neighborhood. In real datasets, local neighborhoods can have very different scales as illustrated in Fig. 3.8. To make the target neighborhood spreads comparable, we propose for each $\mathbf{x}_i$ to scale its neighborhood vectors $\mathbf{x}_{ij}$ such that the L2 norm (radius) of the farthest m-th nearest neighbor is 1:

$$\mathbf{x}_{ij}^{norm} = \left[\frac{d_1(\mathbf{x}_{ij})}{r_i}, \ldots, \frac{d_p(\mathbf{x}_{ij})}{r_i}\right]^T \tag{4.16}$$

For simplification purpose, we denote in the following $\mathbf{x}_{ij}$ as $\mathbf{x}_{ij}^{norm}$. Fig. 4.4 illustrates the effect of neighborhood scaling in the pairwise space.

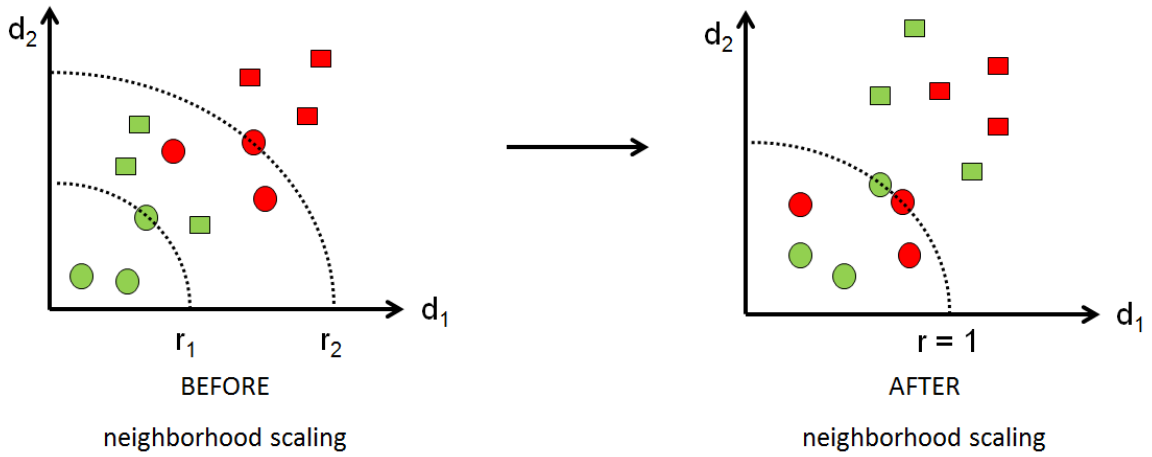


Figure 4.4: Effect of neighborhood scaling before (left) and after (right) on the neighborhood of two time series $\mathbf{x}_1$ (green) and $\mathbf{x}_2$ (red). Circle represent negative pairs (m-NN) and square represents positive pairs (m-diff) for m=2 neighbors. Before scaling, the problem is not linearly separable. The spread of each neighborhood are not comparable. After scaling, the target neighborhood becomes comparable and in this example, the problem becomes linearly separable between the circles and the squares.

4.4 Solving the Support Vector Machine (SVM) problem

Let $\{\mathbf{x}_{ij}; y_{ij} = \pm 1\}$, $\mathbf{x}_{ij} \in m\text{-NN}^+ \cup m\text{-NN}^-$ be the training set, with $y_{ij} = +1$ for $\mathbf{x}_{ij} \in m\text{-NN}^+$ (same label) and -1 for $\mathbf{x}_{ij} \in m\text{-NN}^-$ (different labels). For a maximum margin between positive and negative pairs, the problem is formalized in an SVM framework as follows in the pairwise space $\mathscr{E}$:

$$\underset{\mathbf{w},b,\boldsymbol{\xi}}{\operatorname{argmin}} \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i,j}\xi_{ij} \tag{4.17}$$

$$\text{s.t. } y_{ij}(\mathbf{w}^T\mathbf{x}_{ij} + b) \geq 1 - \xi_{ij} \tag{4.18}$$

$$\xi_{ij} \geq 0 \tag{4.19}$$

Thanks to the unit radii normalization $\mathbf{x}_{ij}/r_i$, the SVM ensures a global large margin solution involving equally local neighborhood constraints (i.e. local margins). An L1 regularization in Eq. 4.17 leads to a sparse and interpretable $\mathbf{w}$ that uncovers the modalities, periods and scales that differentiate best positive from negative pairs for a robust nearest neighbors classification:

$$\underset{\mathbf{w},b,\boldsymbol{\xi}}{\operatorname{argmin}} ||\mathbf{w}||_1 + C\sum_{i,j}\xi_{ij} \tag{4.20}$$

4.5 Definition of the dissimilarity measure

The proposed M^{2}TML approach differs from the one of Time series Metric Learning (TML) by Linear/Quadratic programming (LP/QP) in which a SVM pairwise is used to learn the best weight vector $\mathbf{w}$ such that positive pairs are widely separated from negative pairs. Thus, defining the learned metric D from the vector $\mathbf{w}$ needs to be careful.

Let $\mathbf{x}_{test}$ be a new sample, $\mathbf{x}_{i,test} \in \mathscr{E}$ gives the proximity between $\mathbf{x}_i$ and $\mathbf{x}_{test}$ based on the p multi-modal and multi-scale metrics d_h. We denote $\mathbf{P}_{\mathbf{w}}(\mathbf{x}_{i,test})$ the orthogonal projection of $\mathbf{x}_{i,test}$ on the axis of direction $\mathbf{w}$ and $||\mathbf{P}_{\mathbf{w}}(\mathbf{x}_{i,test})||$ its norm that allows to measure the closeness between $\mathbf{x}_{test}$ and $\mathbf{x}_i$ while considering the discriminative features between positive and negative pairs. We review in this section different proposition to define the learned metric D: Scalar product, Projection Norm, Exponential transformation.

Scalar product and norm

First, the learned metric D can be defined as the decision function obtained by solving the SVM problem:

$$D(\mathbf{x}_{i,test}) = \mathbf{w}^T \mathbf{x}_{i,test} + b \tag{4.21}$$

The obtained metric D doesn't necessarily satisfy the distinguishability ($D(\mathbf{x}_{ii} = 0)$) and positivity ($D(\mathbf{x}_{ij} \geq 0)$) property, especially when positive pairs (different classes) are situated nearer to the origin point than negative pairs (same class) (Fig. 4.5).

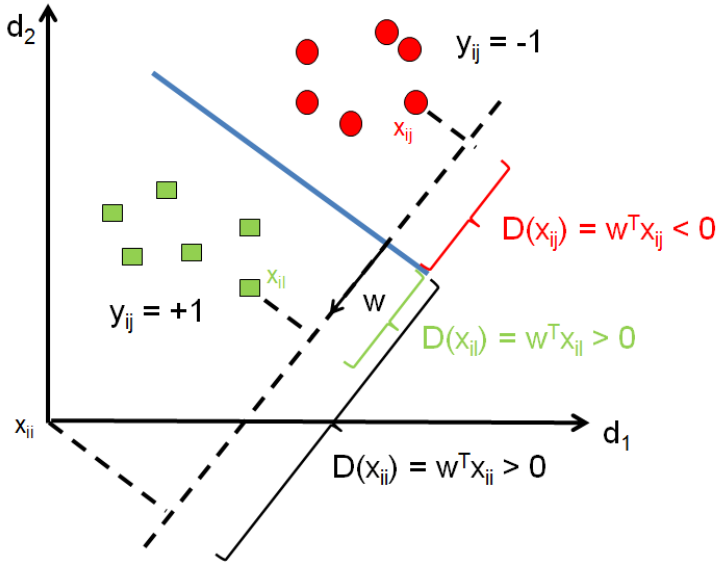


Figure 4.5: Example of SVM solutions and of the resulting metric D defined by a scalar product. The vector $\mathbf{w} = [-1 - 1]$ indicates that positive pairs (y_{ij}) are on the side of the origin point. Two problems arises: 1) For negative pairs, $D(\mathbf{x}_{ij}) \leq 0$. 2) For the origin point $\mathbf{x}_{ii}$, we obtain $D(\mathbf{x}_{ii}) \neq 0$.

The norm of the projection $\mathbf{P_w}(\mathbf{x}_{i,test})$ can be used to define the learned metric D as it measures the distance of the pair $\mathbf{x}_{i,test}$ from the origin point $\mathbf{x}_{ii}$ along to the direction $\mathbf{w}$:

$$D(\mathbf{x}_{i,test}) = ||\mathbf{P_w}(\mathbf{x}_{i,test})|| = ||\mathbf{w}^T\mathbf{x}_{i,test}|| \tag{4.22}$$

Although the norm $||\mathbf{P_w}(\mathbf{x}_{i,test})||$ satisfies metric properties, it doesn't always guarantee lower distances between negative pairs (same class) than positive pairs (different classes) as illustrated in Fig 4.6.

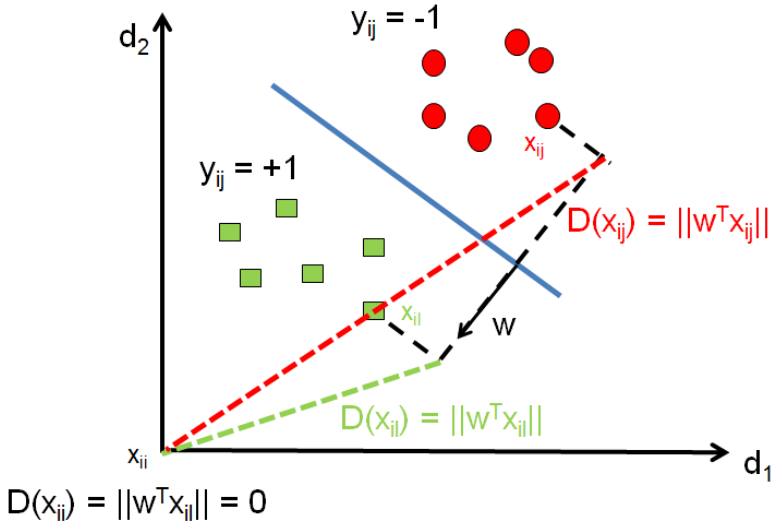


Figure 4.6: Example of SVM solutions and of the resulting metric D defined by the norm of the projection on $\mathbf{w}$. The vector $\mathbf{w} = [-1 - 1]$ indicates that positive pairs (y_{ij}) are on the side of the origin point $\mathbf{x}_{ii} = 0$. One problem: distance of positive pairs $D(\mathbf{x}_{il})$ is lower than the distance of negative pairs $D(\mathbf{x}_{ij})$.

Exponential transformation

Secondly, we propose to add an exponential term to operate a "push" on negative pairs based on their distances to the separator hyperplan, that leads to the dissimilarity measure D of required properties:

$$D(\mathbf{x}_{i,test}) = ||\mathbf{P_w}(\mathbf{x}_{i,test})|| \exp(\lambda[\mathbf{wP_w}(\mathbf{x}_{i,test}) + b]_+) \quad \lambda > 0 \tag{4.23}$$

where λ controls the "push" term and $\mathbf{wP_w}(\mathbf{x}_{i,test}) + b$ defines the distance between the orthogonal projected vector and the separator hyperplane; $[t]_+ = \max(0; t)$ being the positive operator. Note that, for a pair lying into the negative side ($y_{ij} = -1$), $[\mathbf{wP_w}(\mathbf{x}_{i,test}) + b]_+ = 0$, the exponential term is vanished (i.e. no "pull" action) and the dissimilarity leads to the norm term. For a pair situated in the positive side ($y_{ij} = +1$), the norm is expanded by the push term, all the more the distance to the hyperplane is high.

Fig. 4.7, illustrates for $p = 2$ the behavior of the learned dissimilarity according to two

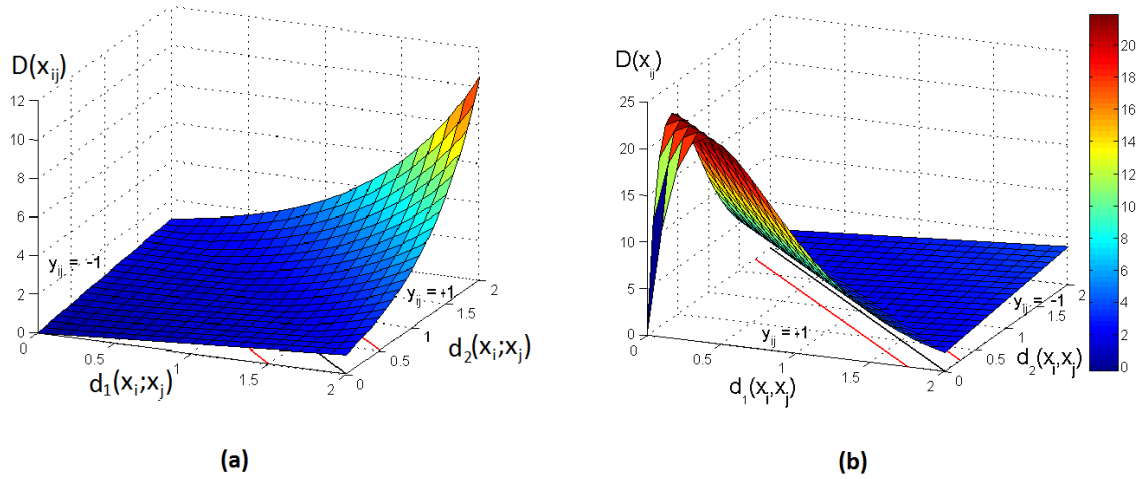


Figure 4.7: The behavior of the learned metric D ($p = 2$; $\lambda = 2.5$) with respect to common (a) and challenging (b) configurations of positive and negatives pairs.

extreme cases. The first one (Fig. 4.7-a), represents common expected configuration where negative pairs ($\mathbf{x}_i$ and $\mathbf{x}_j$ are of same class) are situated in the same side as the origin. The dissimilarity increases proportionally to the norm in the negative side, then exponentially on the positive side. Although the expansion operated in the positive side is dispensable in that case, it doesn't affect nearest neighbors classification. Fig. 4.7-b, shows a challenging configuration where positive pairs ($\mathbf{x}_i$ and $\mathbf{x}_j$ are of different classes) are situated in the same side as the origin. That means that time series $\mathbf{x}_j$ that are of different classes from $\mathbf{x}_i$ are closer to $\mathbf{x}_i$ than its nearest neighbors. They are thus impostors. The dissimilarity behaves proportionally to the norm on the negative side, and increases exponentially from the hyperplane until an abrupt decrease induced by a norm near 0. Note that the region under the abrupt decrease mainly uncovers false positive pairs, i.e., pairs of norm zero labeled differently.

The above solution holds true for any kernel K and allows to extend the dissimilarity D given in Eq. 4.23 to non linearly separable positive and negative pairs. Let K be a kernel defined in the pairwise space $\mathscr{E}$ and the related Hilbert space (feature space) $\mathscr{H}$. For a non linear combination function of the metrics $d_h, h = 1, \ldots, p$ in $\mathscr{E}$, we define the dissimilarity measure $D_{\mathscr{H}}$ in the feature space $\mathscr{H}$ as:

$$\begin{aligned} D_{\mathscr{H}}(\mathbf{x}_{i,test}) =& (||\mathbf{P}_{\mathbf{w}}(\phi(\mathbf{x}_{i,test}))|| - ||\mathbf{P}_{\mathbf{w}}(\phi(\mathbf{0}))||). \\ & \exp\left(\lambda \left[\sum_{ij} y_{ij}\alpha_{ij}K(\mathbf{x}_{ij}, \mathbf{x}_{i,test}) + b\right]_+\right) \qquad \lambda > 0 \end{aligned} \tag{4.24}$$

with $\phi(\mathbf{w})$ the image of $\mathbf{w}$ into the feature space $\mathscr{H}$ and the norm of the orthogonal projection

of $\phi(\mathbf{x}_{i,test})$ on $\phi(\mathbf{w})$ as:

$$||\mathbf{P}_{\mathbf{w}}(\mathbf{x}_{i,test})|| = \frac{\sum_{ij} y_{ij}\alpha_{ij}K(\mathbf{x}_{ij}, \mathbf{x}_{i,test})}{\sqrt{\sum_{ijkl} \alpha_{ij}\alpha_{kl}y_{ij}y_{kl}K(\mathbf{x}_{ij}, \mathbf{x}_{kl})}} \tag{4.25}$$

Note that as $\phi(\mathbf{0})$ doesn't meet the origin in the feature space $\mathscr{H}$, the norms in Eq. 4.24 are centered with respect to $\phi(\mathbf{0})$.

Limitations

à compléter avec la figure de Sylvain

4.6 Algorithms and extensions

Algorithm 1 summarizes the main steps to learn a multi-modal and multi-scale metric D for a robust nearest neighbors classification. Algorithm 2 details the steps to classify a new sample $\mathbf{x}_{test}$ using the learned metric D.

Algorithm 1 Multi-modal and Multi-scale Temporal Metric Learning (M^{2}TML) for k-NN classification

1: Input: $X = \{\mathbf{x}_i, y_i\}_{i=1}^N$ N labeled time series
$d_1, ..., d_p$ metrics as described in Eqs. 2.1, 2.4, 2.6, 4.4
a kernel K
2: Output: the learned dissimilarity D or $D_{\mathscr{H}}$ depending of K
3: *Pairwise embedding*
Embed pairs $(\mathbf{x}_i, \mathbf{x}_j)$ $i, j \in 1, ..., N$ into $\mathscr{E}$ as described in Eq. 4.5 and normalize d_hs
4: *Build positive and negative pairs*
Build the sets positive m-NN$^+$ and negative m-NN$^-$ pairs and scale the radii to 1 as described in 4.3
5: Train an SVM for a large margin classifier between m-NN$^+$ and m-NN$^-$ (Eq. 4.17)
6: *Dissimilarity definition*
Consider Eq. 4.23 (resp. Eq. 4.24) to define D (resp. $D_{\mathscr{H}}$) a linear (resp. non linear) combination function of the metrics d_hs.

Algorithm 1 can be easily extended for multivariate and regression problem. First, for multivariate problem, each unimodal metric d_h can be computed for each variable. Then, the above framework can be applied. For regression problem, the label y_i for each time series $\mathbf{x}_i$ is a continuous value. The only modification is at the neighborhood steps, when defining the positive and negative pairs labeled y_{ij}. For that, In Chapter 3, Section 3.1.1, we propose two different strategies to define the pairwise labels y_{ij}.

Algorithm 2 k-NN classification using the learned metric D or $D_{\mathscr{H}}$

Input: $X = \{\mathbf{x}_i, y_i\}_{i=1}^N$ N labeled time series
$\{\mathbf{x}_{test}, y_{test}\}$ a labeled time series to test
$d_1, ..., d_p$ metrics as described in Eqs. 2.1, 2.4, 2.6, 4.4
the learned dissimilarity D or $D_{\mathscr{H}}$ depending of the kernel K
Output: Predicted label $\hat{y}_{test}$
Pairwise embedding
Embed pairs $(\mathbf{x}_i, \mathbf{x}_{test})$ $i \in 1, ..., N$ into $\mathscr{E}$ as described in Eq. 4.5 and normalize d_hs using the same normalization parameters in Algorithm 1
Dissimilarity computation
Consider Eq. 4.23 (resp. Eq. 4.24) to compute $D(\mathbf{x}_i, \mathbf{x}_{test})$ (resp. $D_{\mathscr{H}}(\mathbf{x}_i, \mathbf{x}_{test})$) a linear (resp. non linear) combination function of the metrics $d_h(\mathbf{x}_i, \mathbf{x}_{test})$.
Classification
Consider the k lowest dissimilarities $D(\mathbf{x}_i, \mathbf{x}_{test})$ (resp. $D_{\mathscr{H}}(\mathbf{x}_i, \mathbf{x}_{test})$). Extract the labels y_i of the considered $\mathbf{x}_i$ and make a vote scheme to predict the label $\hat{y}_{test}$ of $\mathbf{x}_{test}$

4.7 Conclusion of the chapter

To learn a multi-modal and multi-scale time series metric for a robust k-NN classifier, we detailed in this chapter the different steps of our proposed framework.
First, time series are embedded in the pairwise space using a multi-scale description. Secondly, for each time series, we construct its m-nearest neighborhood of the same class and of different classes, forming our pairwise training set. A pairwise SVM is learnt on the pairwise training set. Finally, the dissimilarity measure is defined to satisfy the required conditions of a metric.

Conclusion of Part II

Part III

Experiments

Chapter 5

Experiments

Sommaire

Chapeau introductif

- Application sur des bases de séries temporelles univariés de la littérature (Keogh)
- Données Schneider? ou Expliquer les problématiques de Schneider

5.1 Description

The efficiency of the learned multi-modal and multi-scale dissimilarities D and $D_{\mathscr{H}}$ is evaluated through a 1−NN classification on 30 public datasets [1]. The considered data encompass time series that involve global or local temporal differences, require or not time warping, with linearly or non linearly separable neighborhoods. D and $D_{\mathscr{H}}$ are compared to five alternative uni-modal metrics covering: i) the standard Euclidean distance and dynamic time warping referenced as d_A (Eq. 2.1) and DTW,
ii) the behavior-based measures d_B (Eq. 2.6) and $d_{B-\mathrm{DTW}}$ its counterpart for asynchronous time series, that is d_B is evaluated once time series synchronized by dynamic programing, and iii) frequential-based metric d_F (Eq. 2.4). The alternative metrics are evaluated as usual by involving the all time series elements (*i.e.* at the global scale). For D and $D_{\mathscr{H}}$, we consider a 21-dimensional embedding space $\mathscr{E}$ that relies, for synchronous (resp. asynchronous) data, on 3 log-normalized dissimilarities d_A^s, d_B^s (resp. DTW^s, $d_{B-\mathrm{DTW}}^s$), and d_F^s, at 7 temporal granularities $s \in \{0, ..., 6\}$ obtained by binary segmentation as described in Figure **??**.

[1] PowerCons: `https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption`,
BME and UMD: `http://ama.liglab.fr/~douzal/tools.html`, the rest are provided in **UCRArchive**

5.2 Experimental protocol

The combined metrics D and $D_{\mathscr{H}}$ (κ as the gaussian kernel) are learned under L_1 and L_2 regularization, respectively. The parameters are estimated on a validation set by line/grid search. A cross-validation and stratified sampling for unbalanced datasets are used. Particularly, for each couple (r, λ) $r \in \{1, 4, 10\}$ and $\lambda \in \{0, 10, 30\}$, the pairwise SVM parameters (C, α, γ) are learned by grid search as indicated in Table 5.1. The temporal order r is noise-dependent, typically 1 is retained for noise-free data. The parameter λ corresponds to the strength of the '*push*' term; precisely, if no, moderate or strong '*push*' is required during the training process, a λ value of 0, 10 and 30 is learned, respectively. The parameters retained are those that minimize the average classification error on the validation set. In the case of multiple solutions (C, α, γ) leading to equal performances, the most discriminant one is retained (*i.e.* making closer positive pairs and far a way negative pairs).

Parameter		Ranges
d_B	r	$\{1, 2, 3, , ..., T-1\}$
D, $D_{\mathscr{H}}$	λ	$\{0, 10, 30\}$
D, $D_{\mathscr{H}}$	C	$\{10^{-3}, 0.5, 1, 5, 10, 20, 30, ..., 150\}$
D, $D_{\mathscr{H}}$	α	$\{1, 2, 3\}$
$D_{\mathscr{H}}$	γ	$\{10^{-3}, 10^{-2}, ..., 10^{3}\}$

Table 5.1: Parameter ranges

5.3 Results

Table 5.2 reports the 1-NN classification test errors for uni-modal and M^{2}TML metrics; the results that are statistically and significantly better than the rest are indicated in bold (z-test at 5% risk).

5.4 Discussion

Several results are analyzed to evaluate the effectiveness of the learned D and $D_{\mathscr{H}}$ for time series nearest neighbors classification. Table 5.2 compares the 1-NN error rates when based on uni-modal metrics (first 5 columns) and on D and $D_{\mathscr{H}}$. The last column 'WARP' indicates the synchronous (✓) or asynchronous (×) data type.
First, from Table 5.2 we can see that the 1-NN classification reaches the best results in: i) less than one-third of the data when based on d_A, d_B or d_F, ii) slightly more than one-third for DTW and $d_{B-\mathrm{DTW}}$ and iii) more than two-thirds (23 times on 30) when based on D or $D_{\mathscr{H}}$. Particularly, note that for nearly all datasets for which an uni-modal metric succeeds, the M^{2}TML metrics succeed similarly or lead to equivalent results. However, for several challenging datasets (*e.g.* FaceFour, Beef, FaceUCR, SonyAIBO, BME) M^{2}TML realizes

Dataset	Alternative uni-modal metrics					M²TML		WARP
	d_A	d_B	d_F	DTW	$d_{B-\mathrm{DTW}}$	$D(\lambda^*)$	$D_{\mathscr{H}}(\lambda^*)$	WARP
1 CC	0.120	0.113	0.383	**0.007**	0.027	**0.003** (0)	**0.007** (0)	✓
2 GunPoint	0.087	0.113	**0.027**	0.093	**0.027**	**0.020** (10)	**0.040** (10)	✓
3 CBF	0.148	0.140	0.382	**0.003**	**0.000**	0.031 (30)	**0.003** (0)	✓
4 OSULeaf	0.484	0.475	0.426	0.409	**0.265**	0.380 (0)	0.376 (0)	✓
5 SwedishLeaf	0.211	0.186	0.146	0.208	**0.109**	**0.110** (0)	**0.114** (0)	✓
6 Trace	0.240	0.240	0.140	**0.000**	**0.000**	**0.000** (0)	**0.010** (0)	✓
7 FaceFour	0.216	0.216	0.239	0.170	0.136	**0.000** (0)	0.034 (0)	✓
8 Lighting2	**0.246**	**0.246**	**0.148**	**0.131**	**0.213**	**0.148** (0)	**0.131** (0)	✓
9 Lighting7	0.425	0.411	**0.316**	**0.274**	**0.288**	0.397 (0)	**0.233** (0)	✓
10 ECG200	**0.120**	**0.070**	0.160	0.230	0.190	**0.080** (0)	**0.080** (0)	×
11 Adiac	0.389	**0.297**	**0.261**	0.396	0.338	0.358 (0)	0.361 (0)	×
12 FISH	0.217	**0.149**	0.229	**0.166**	**0.137**	**0.149** (0)	0.240 (0)	✓
13 Beef	0.467	0.300	0.500	0.500	0.500	**0.033** (0)	0.267 (0)	×
14 Coffee	0.250	**0.000**	0.357	0.179	0.143	**0.000** (0)	**0.000** (10)	×
15 OliveOil	**0.133**	**0.133**	**0.167**	**0.200**	**0.100**	**0.167** (0)	**0.100** (10)	✓
16 CinCECGtorso	0.103	0.367	0.167	0.349	0.367	**0.092** (0)	**0.079** (0)	×
17 DiatomSizeR	0.065	0.076	0.069	**0.033**	**0.029**	**0.026** (0)	**0.029** (0)	✓
18 ECG5Days	0.203	0.153	**0.006**	0.232	0.236	**0.007** (10)	0.024 (0)	×
19 FacesUCR	0.231	0.227	0.175	0.095	0.102	**0.068** (10)	**0.059** (0)	✓
20 InlineSkate	**0.658**	**0.658**	0.675	**0.616**	**0.623**	0.733 (10)	**0.625** (0)	✓
21 ItalyPowerD	0.045	**0.028**	0.078	0.050	0.055	**0.028** (30)	**0.037** (10)	×
22 MedicalImages	0.316	0.313	0.345	**0.263**	0.290	**0.237** (0)	**0.236** (10)	✓
23 MoteStrain	**0.121**	0.263	0.278	0.165	0.171	0.185 (0)	0.153 (10)	✓
24 SonyAIBOII	**0.141**	**0.142**	**0.128**	0.169	0.194	**0.155** (0)	**0.131** (0)	×
25 SonyAIBO	0.305	0.308	0.258	0.275	0.343	**0.188** (0)	**0.165** (30)	×
26 Symbols	0.101	0.111	0.080	**0.050**	**0.043**	**0.034** (30)	**0.046** (30)	✓
27 TwoLeadECG	0.253	0.153	0.103	0.096	**0.008**	**0.006** (0)	**0.016** (10)	✓
28 PowerCons	**0.366**	0.445	**0.315**	0.397	0.401	**0.318** (0)	**0.308** (0)	✓
29 BME	0.173	0.180	0.373	0.107	0.120	**0.040** (30)	**0.000** (10)	✓
30 UMD	0.194	0.222	0.299	0.118	**0.090**	0.104 (0)	**0.042** (0)	✓

Table 5.2: 1-NN error rates for standard and M²TML measures.

drastic improvements, to the best of our knowledge never achieved before for these challenging public data. For instance, the impressive scores of 3% obtained for Beef against an error rate varying from 30% to 50% for alternative metrics, and of 0% obtained for FaceFour v.s. 13% to 23% for alternative metrics. Finally, D and $D_{\mathscr{H}}$ are all the more outperforming if only compared to the standard metrics d_A (the Euclidean distance) and DTW.

Thanks to the L_1 regularization, the learned SVM reveals the features that most differentiate positive from negative pairs.

Table 5.3 shows the sparse, muti-modal and multi-scale potential of M²TML approach. It gives for each dataset, the weights of the top five 'discriminative' features that contribute to the definition of D. For instance, for FaceFour D reaches the 0% by combining, in the order of importance, the behavior, frequential and amplitude modalities, at the global (I^0) and local (I^4, I^5, I^2) scales. For Beef, besides the impressive error rate of 3%, the learned model is very sparse as D involves only the behavior modality based on the segment I^3 (d_B^3). Similarly

Dataset	Feature weights %				
CC	DTW^2 (56.3%)	d_F^0 (18.8%)	d_F^4 (5.9%)	d_F^1 (4%)	$d_{B-\text{DTW}}^0$ (3.2%)
GunPoint	$d_{B-\text{DTW}}^0$ (42.1%)	DTW^5 (10.9%)	DTW^1 (10.2%)	$d_{B-\text{DTW}}^6$ (8.4%)	d_F^4 (8.1%)
CBF	DTW^4 (56.5%)	d_F^3 (43.4%)	$d_{B-\text{DTW}}^1$ (0.2%)	DTW^0 (0%)	$d_{B-\text{DTW}}^0$ (0%)
OSULeaf	DTW^2 (23.7%)	d_F^0 (19.5%)	$d_{B-\text{DTW}}^0$ (14.6%)	$d_{B-\text{DTW}}^2$ (9.4%)	DTW^1 (9%)
SwedishLeaf	d_F^0 (21.5%)	DTW^0 (15.9%)	$d_{B-\text{DTW}}^0$ (15.2%)	DTW^6 (11.5%)	$d_{B-\text{DTW}}^1$ (6.1%)
Trace	DTW^0 (58.3%)	DTW^6 (6.9%)	$d_{B-\text{DTW}}^0$ (5.8%)	DTW^2 (5.6%)	DTW^5 (5.5%)
FaceFour	$d_{B-\text{DTW}}^4$ (44.5%)	d_F^4 (12.7%)	DTW^5 (11.1%)	DTW^0 (8.3%)	DTW^2 (6.4%)
Lighting2	$d_{B-\text{DTW}}^0$ (30.4%)	DTW^6 (18.7%)	d_F^1 (16.5%)	$d_{B-\text{DTW}}^6$ (13.4%)	DTW^0 (10.7%)
Lighting7	$d_{B-\text{DTW}}^6$ (87.4%)	d_F^6 (8.6%)	$d_{B-\text{DTW}}^5$ (4%)	-	-
ECG200	d_B^0 (89.6%)	d_B^6 (2.4%)	d_A^3 (2.3%)	d_B^1 (2.2%)	d_B^4 (2%)
Adiac	d_F^0 (79.2%)	d_B^4 (13.8%)	d_A^4 (3.5%)	d_F^5 (1.7%)	d_B^5 (1.2%)
FISH	$d_{B-\text{DTW}}^5$ (17.9%)	d_F^0 (10.5%)	$d_{B-\text{DTW}}^6$ (9.9%)	$d_{B-\text{DTW}}^4$ (8.3%)	$d_{B-\text{DTW}}^3$ (7.8%)
Beef	d_B^3 (100%)	-	-	-	-
Coffee	d_B^2 (22.4%)	d_F^4 (20.1%)	d_B^6 (14.6%)	d_B^0 (8.1%)	d_F^5 (7%)
OliveOil	d_F^5 (97%)	$d_{B-\text{DTW}}^2$ (3%)	-	-	-
CinCECGtorso	d_F^0 (38.4%)	d_A^5 (13.1%)	d_B^4 (11.5%)	d_F^1 (11.2%)	d_A^2 (9.8%)
DiatomSizeR	d_F^5 (39.1%)	d_F^0 (36%)	$d_{B-\text{DTW}}^4$ (24.9%)	-	-
ECG5Days	d_B^5 (59.5%)	d_B^6 (32.3%)	d_A^4 (3.9%)	d_B^2 (3.1%)	d_B^4 (1.2%)
FacesUCR	d_F^2 (21.5%)	$d_{B-\text{DTW}}^0$ (19.5%)	d_F^4 (16.7%)	DTW^0 (12.6%)	$d_{B-\text{DTW}}^2$ (8.6%)
InlineSkate	d_F^4 (42.5%)	DTW^5 (22.8%)	DTW^4 (17.6%)	DTW^2 (6.7%)	$d_{B-\text{DTW}}^6$ (5.9%)
ItalyPowerD	d_B^6 (68.7%)	d_B^0 (25.9%)	d_B^3 (5.2%)	d_B^4 (0.2%)	-
MedicalImages	$d_{B-\text{DTW}}^1$ (53.3%)	d_F^3 (12.9%)	$d_{B-\text{DTW}}^2$ (10.7%)	$d_{B-\text{DTW}}^3$ (10.1%)	$d_{B-\text{DTW}}^0$ (3.8%)
MoteStrain	$d_{B-\text{DTW}}^5$ (93.2%)	$d_{B-\text{DTW}}^6$ (6.8%)	-	-	-
SonyAIBOII	d_B^3 (100%)	-	-	-	-
SonyAIBO	d_F^3 (30.8%)	d_B^6 (27.3%)	d_B^5 (5%)	d_A^1 (4.1%)	d_B^0 (3.9%)
Symbols	$d_{B-\text{DTW}}^0$ (45.6%)	$d_{B-\text{DTW}}^6$ (35.3%)	$d_{B-\text{DTW}}^5$ (19%)	DTW^0 (0.1%)	-
TwoLeadECG	$d_{B-\text{DTW}}^4$ (60%)	d_F^1 (12%)	DTW^4 (11.4%)	$d_{B-\text{DTW}}^6$ (7.6%)	$d_{B-\text{DTW}}^1$ (4.2%)
PowerCons	d_F^0 (26.1%)	DTW^0 (20.3%)	d_F^1 (19.3%)	$d_{B-\text{DTW}}^0$ (6.1%)	d_F^2 (5.1%)
BME	$d_{B-\text{DTW}}^0$ (75.2%)	d_F^4 (15.5%)	$d_{B-\text{DTW}}^2$ (5.8%)	$d_{B-\text{DTW}}^1$ (1.9%)	d_F^1 (0.7%)
UMD	$d_{B-\text{DTW}}^0$ (99.8%)	$d_{B-\text{DTW}}^5$ (0.2%)	-	-	-

Table 5.3: Top 5 multi-modal and multi-scale features involved in D

for Coffee, the obtained 0% involves only the behavior and frequential modalities at several scales.

In Figure 5.1 we plot the the weights of all features for SonyAIBO case as an example, that illustrates the sparsity of the M^2TML approach. In summary, we can emphasize that for almost all datasets, the definition of D involves no more than five features (the most contributive ones), that assesses not only the model's sparsity but also the representativeness of the revealed features.

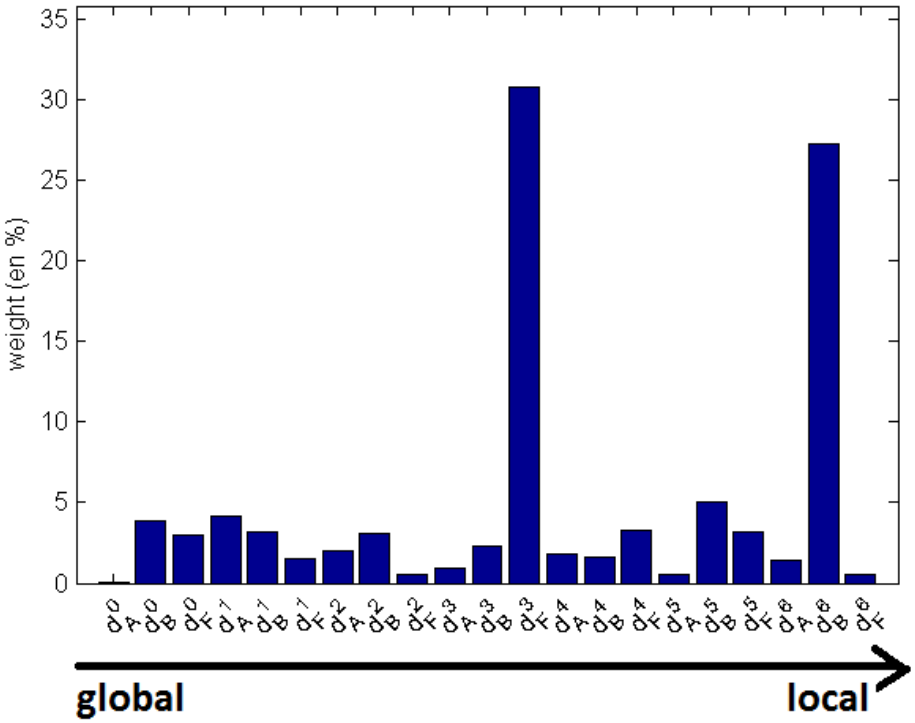


Figure 5.1: SonyAIBO: M^{2}TML feature weights

In the second part, we perform a graphical analysis for a global comparison on the whole datasets. In Figure 5.2, each dataset is projected according to, on the x-axis its best error rate obtained for D and $D_{\mathscr{H}}$, and on y-axis its best performance w.r.t the standard metrics d_A and DTW. In Figure 5.3, the y-axis is related to the best error rate w.r.t DTW and $d_{B-\mathrm{DTW}}$, the two most performant uni-modal metrics. For both plots we can note that the datasets are principally projected above the first bisector, indicating higher error rates mostly obtained for alternative metrics than for M^{2}TML. For the less challenging datasets, although almost projected near the bisector denoting equal performances for the compared metrics, M^{2}TML still bring improvements with projections clearly positioned above the bisector. Finally, from Figure 5.3 we can see that M^{2}TML metrics perform significantly lower than $d_{B-\mathrm{DTW}}$ on OSUleaf, while InlineSkate dataset remains challenging for all studied metrics.

In the last part we compare the global effect of the alternative and M^{2}TML metrics on the 1-NN neighborhood distribution and class discrimination. For that, an MDS[2] is used to vizualize the distribution of samples according to their pairwise dissimilarities. For instance, for FaceFour, Figure 5.4 shows the first obtained plans and their corresponding stresses, the classes being indicated in different symbols and colors. We can see distinctly the effect of the learned D that leads to more compact and more isolated classes with robust neighborhoods for 1-NN classification (*i.e.* closer positive pairs and far away negative pairs) than the best alter-

[2]matlab function: mdscale for metrics and non metrics

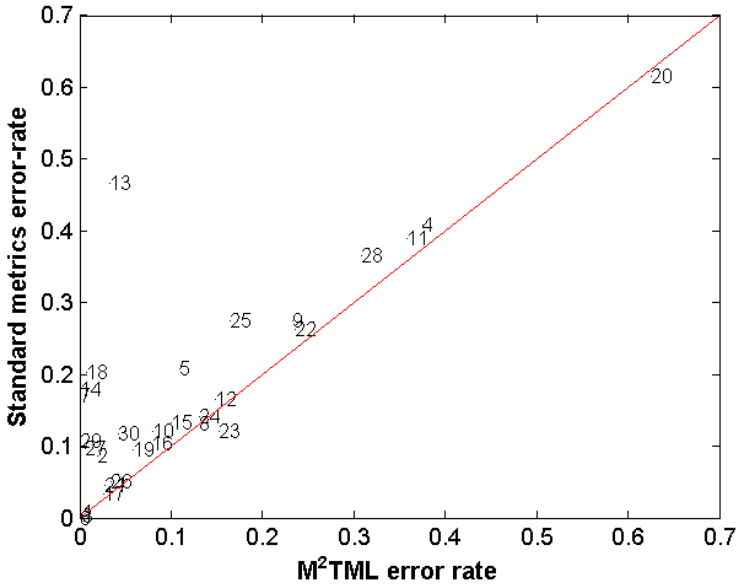


Figure 5.2: Standard (Euclidean distance d_A and DTW) *vs.* M^{2}TML (D and $D_{\mathscr{H}}$) metrics

native metric $d_{B-\mathrm{DTW}}$ that shows more overlapping classes and heterogeneous neighborhoods.

5.5 Conclusion of the chapter

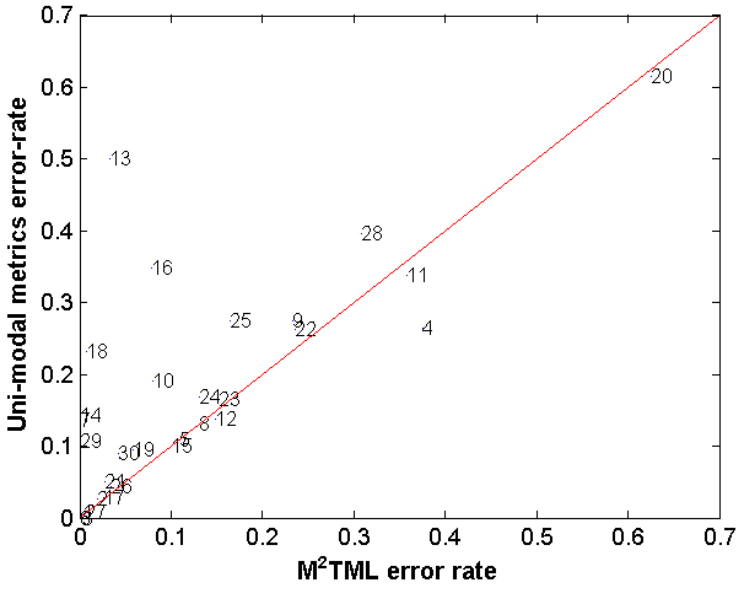


Figure 5.3: Best Uni-modal (DTW and $d_{B-\mathrm{DTW}}$) *vs.* M^{2}TML (D and $D_{\mathscr{H}}$) metrics

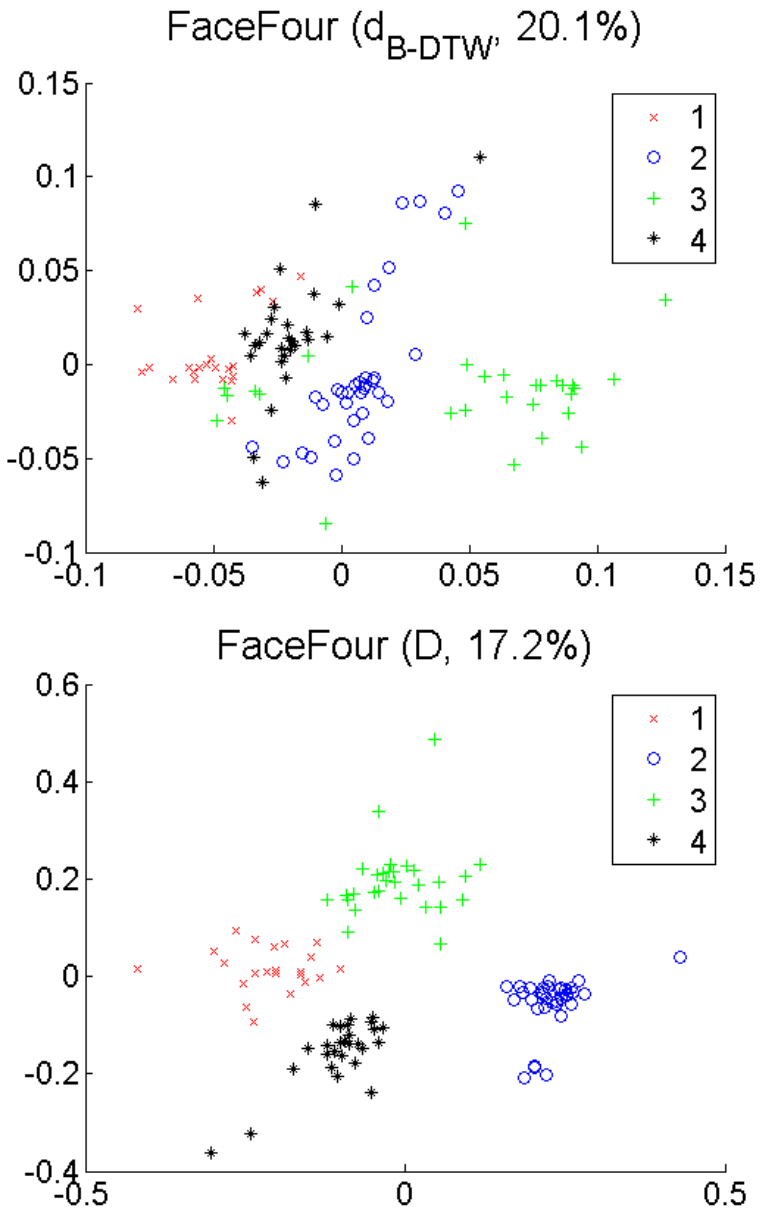


Figure 5.4: MDS visualization of the $d_{B-\mathrm{DTW}}$ (top) and D (bottom) dissimilarities for FaceFour data

Conclusion of Part III

Conclusion and perspectives

- Bilan des apports de la thèse
- Perspectives
  - Multi-pass learning
  - Kernel pour la résolution du problème QP
  - Utilisation de la distance apprise dans d'autres algorithmes de machine learning (Arbre de décision) pour obtenir une interprétabilité?
  - Utilisation d'autres distances (wavelets, etc.)
  - Apprentissage locale de la métrique

Appendix A

Detailed presentation of the datasets

Appendix B

Solver library

Appendix C

SVM library

Appendix D

QP resolution

Bibliography

[ABR64] M. Aizerman, E. Braverman, and L. Rozonoer. "Theoretical foundations of the potential function method in pattern recognition learning." In: *Automation and Remote Control* 25 (1964), pp. 821–837 (cit. on p. 21).

[Alt92] Ns Altman. "An introduction to kernel and nearest-neighbor nonparametric regression." In: *The American Statistician* 46.3 (1992), pp. 175–185 (cit. on p. 16).

[AT10] Z. Abraham and P.N. Tan. "An Integrated Framework for Simultaneous Classification and Regression of Time-Series Data." In: *ACM SIGKDD*. 2010 (cit. on p. 35).

[BC94] Donald Berndt and James Clifford. "Using dynamic time warping to find patterns in time series." In: *Workshop on Knowledge Knowledge Discovery in Databases* 398 (1994), pp. 359–370 (cit. on pp. 37, 39, 68).

[Ben+09] J. Benesty et al. "Pearson correlation coefficient." In: *Noise Reduction in Speech Processing* (2009) (cit. on p. 35).

[BGV92] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. "A Training Algorithm for Optimal Margin Classifiers." In: *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*. 1992, pp. 144–152 (cit. on pp. 17, 22).

[Bis06] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Vol. 4. 4. 2006, p. 738. arXiv: `0-387-31073-8` (cit. on pp. 8, 28).

[BM67] E. O. Brigham and R. E. Morrow. "The fast Fourier transform." In: *Spectrum, IEEE* 4.12 (1967), pp. 63 –70 (cit. on p. 34).

[BMP02] Serge Belongie, Jitendra Malik, and Jan Puzicha. "Shape Matching and Object Recognition Using Shape Contexts." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (2002), pp. 509–522 (cit. on p. 16).

[CH67] T. Cover and P. Hart. "Nearest neighbor pattern classification." In: *IEEE Transactions on Information Theory* 13.1 (1967), pp. 21–27 (cit. on p. 15).

[Cha04] Christopher Chatfield. *The analysis of time series : an introduction*. 2004, xiii, 333 p. (Cit. on p. 32).

[CHL05] Sumit Chopra, Raia Hadsell, and Yann LeCun. "Learning a similarity metric discriminatively, with application to face verification." In: *CVPR*. Vol. 1. 2005, pp. 539–546 (cit. on p. 42).

[CHY96] Ming Syan Chen, Jiawei Han, and Philip S. Yu. *Data mining: An Overview from a Database Perspective*. 1996 (cit. on p. 8).

[Coc77] William C Cochran. "Snedecor G W & Cochran W G. Statistical methods applied to experiments in agriculture and biology. 5th ed. Ames, Iowa: Iowa State University Press, 1956." In: *Citation Classics* 19 (1977), p. 1 (cit. on p. 12).

[CS01] Koby Crammer and Yoram Singer. "On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines." In: *Journal of Machine Learning Research* 2 (2001), pp. 265–292 (cit. on p. 27).

[CT01] Lijuan Cao and Francis E H Tay. "Financial Forecasting Using Support Vector Machines." In: *Neural Computing & Applications* (2001), pp. 184–192 (cit. on p. 33).

[CV95] Corinna Cortes and Vladimir Vapnik. "Support-vector networks." In: *Machine Learning* 20.3 (1995), pp. 273–297. arXiv: `arXiv:1011.1669v3` (cit. on p. 17).

[CY11] Colin Campbell and Yiming Ying. *Learning with Support Vector Machines*. Vol. 5. 1. 2011, pp. 1–95 (cit. on pp. 17, 21).

[DCA11] A. Douzal-Chouakria and C. Amblard. "Classification trees for time series." In: *Pattern Recognition journal* (2011) (cit. on pp. 36, 40).

[DCN07] A. Douzal-Chouakria and P. Nagabhushan. "Adaptive dissimilarity index for measuring time series proximity." In: *Advances in Data Analysis and Classification* (2007) (cit. on p. 37).

[Den95] T. Denoeux. "A k-nearest neighbor classification rule based on Dempster-Shafer theory." In: *IEEE Transactions on Systems, Man, and Cybernetics* 25.5 (1995), pp. 804–813 (cit. on p. 16).

[DHB95] Thomas G. Dietterich, Hermann Hild, and Ghulum Bakiri. "A comparison of ID3 and backpropagation for English text-to-speech mapping." In: *Machine Learning* 18.1 (1995), pp. 51–80 (cit. on p. 12).

[Die97] T. Dietterich. "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms." In: (1997) (cit. on p. 12).

[Din+08] Hui Ding et al. "Querying and Mining of Time Series Data : Experimental Comparison of Representations and Distance Measures." In: *Proceedings of the VLDB Endowment* 1.2 (2008), pp. 1542–1552. arXiv: `1012.2789v1` (cit. on pp. 16, 33, 34).

[Do+12] Huyen Do et al. "A metric learning perspective of SVM: on the relation of LMNN and SVM." In: *Proceedings of the 15th International Con- ference on Artificial Intelligence and Statistics (AISTAS '12)* (2012), pp. 308–317. arXiv: `arXiv:1201.4714v1` (cit. on pp. 44, 45).

[Dud76] Sahibsingh a. Dudani. "DISTANCE-WEIGHTED k-NEAREST-NEIGHBOR RULE." In: *IEEE Transactions on Systems, Man and Cybernetics* SMC-6.4 (1976), pp. 325–327 (cit. on p. 16).

[FCH08] RE Fan, KW Chang, and CJ Hsieh. "LIBLINEAR: A library for large linear classification." In: *The Journal of Machine Learning* (2008) (cit. on pp. 23, 60).

[FRM94] Christos Faloutsos, M. Ranganathan, and Yannis Manolopoulos. "Fast subsequence matching in time-series databases." In: *ACM SIGMOD Record* 23.2 (1994), pp. 419–429 (cit. on p. 68).

[Gol+04] Jacob Goldberger et al. "Neighbourhood Components Analysis." In: *Advances in Neural Information Processing Systems* (2004), pp. 513–520 (cit. on p. 42).

[HCL08] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. "A Practical Guide to Support Vector Classification." In: *BJU international* 101.1 (2008), pp. 1396–400. arXiv: `0-387-31073-8` (cit. on pp. 14, 60).

[HHK12] Seok Hwan Hwang, Dae Heon Ham, and Joong Hoon Kim. "Forecasting performance of LS-SVM for nonlinear hydrological time series." In: *KSCE Journal of Civil Engineering* 16.5 (2012), pp. 870–882 (cit. on p. 33).

[HHP01] B Heisele, P Ho, and T Poggio. "Face recognition with support vector machines: global versus component-based approach." In: *IEEE International Conference on Computer Vision, ICCV*. Vol. 2. July. 2001, pp. 688–694 (cit. on p. 17).

[HWZ13] Jianming Hu, Jianzhou Wang, and Guowei Zeng. "A hybrid forecasting approach applied to wind speed time series." In: *Renewable Energy* 60 (2013), pp. 185–194 (cit. on p. 33).

[JMF99] a. K. Jain, M. N. Murty, and P. J. Flynn. "Data clustering: a review." In: *ACM Computing Surveys* 31.3 (1999), pp. 264–323. arXiv: `arXiv:1101.1881v2` (cit. on p. 8).

[Kal60] R E Kalman. "A New Approach to Linear Filtering and Prediction Problems." In: *Transactions of the ASME Journal of Basic Engineering* 82.Series D (1960), pp. 35–45 (cit. on p. 36).

[KGG85] James M. Keller, Michael R. Gray, and James a. Givens. *A fuzzy K-nearest neighbor algorithm*. 1985 (cit. on p. 16).

[KR04] Eamonn Keogh and Chotirat Ann Ratanamahatana. "Exact indexing of dynamic time warping." In: *Knowledge and Information Systems* 7.3 (2004), pp. 358–386 (cit. on p. 38).

[KU02] B Kijsirikul and N Ussivakul. "Multiclass Support Vector Machines using Adaptive Directed Acyclic Graph." In: *Neural Networks, 2002. IJCNN '02. Proceedings of the 2002 International Joint Conference on* 1 (2002), pp. 980–985 (cit. on p. 27).

[Lhe+11] S. Lhermitte et al. "A comparison of time series similarity measures for classification and change detection of ecosystem dynamics." In: *Remote Sensing of Environment* 115.12 (2011), pp. 3129–3152 (cit. on p. 34).

[Lia+12] Chunquan Liang et al. "Learning very fast decision tree from uncertain data streams with positive and unlabeled samples." In: *Information Sciences* 213 (2012), pp. 50–67 (cit. on p. 33).

[MV14] Pablo Montero and José Vilar. "TSclust : An R Package for Time Series Clustering." In: *Journal of Statistical Software November* 62.1 (2014) (cit. on p. 33).

[Naj+12] H. Najmeddine et al. "Mesures de similarité pour l'aide à l'analyse des données énergétiques de bâtiments." In: *RFIA*. 2012 (cit. on pp. 31, 33, 37).

[Ngu+12] L. Nguyen et al. "Predicting collective sentiment dynamics from time-series social media." In: *WISDOM*. 2012 (cit. on p. 31).

[OE73] Richard O Duda and Peter E Hart. *Pattern Classification and Scene Analysis*. Vol. 7. 1973, p. 482 (cit. on pp. 8, 11, 16).

[PAN+08] COSTAS PANAGIOTAKIS et al. "SHAPE-BASED INDIVIDUAL/GROUP DETECTION FOR SPORT VIDEOS CATEGORIZATION." In: *International Journal of Pattern Recognition and Artificial Intelligence* 22.06 (2008), pp. 1187–1213 (cit. on p. 31).

[PL12] Zoltán Prekopcsák and Daniel Lemire. "Time series classification by class-specific Mahalanobis distance measures." In: *Advances in Data Analysis and Classification* 6.3 (2012), pp. 185–200. arXiv: `1010.1526` (cit. on p. 34).

[Ram+08] E. Ramasso et al. "Human action recognition in videos based on the transferable belief model : AAAApplication to athletics jumps." In: *Pattern Analysis and Applications* 11.1 (2008), pp. 1–19 (cit. on p. 31).

[RJ93] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Vol. 103. 1993 (cit. on p. 39).

[SC] Stan Salvador and Philip Chan. "FastDTW : Toward Accurate Dynamic Time Warping in Linear Time and Space." In: () (cit. on p. 38).

[SC78] H. Sakoe and S. Chiba. "Dynamic Programming Algorithm Optimization for Spoken Word Recognition." In: *IEEE transactions on acoustics, speech, and signal processing* (1978) (cit. on p. 39).

[She+02] Noam Shental et al. "Adjustment Learning and Relevant Component Analysis." In: *European Conference on Computer Vision (ECCV)* 2353 (2002), pp. 776–790 (cit. on p. 42).

[SJ89] B W Silverman and M C Jones. "E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951)." In: *International Statistical Review / Revue Internationale de Statistique* 57.3 (1989), pp. 233–238 (cit. on p. 15).

[SS12] Md Sahidullah and Goutam Saha. "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition." In: *Speech Communication* 54.4 (2012), pp. 543–565 (cit. on p. 34).

[SS13] Bernhard Schlkopf and Alexander J. Smola. *Learning with Kernels*. Vol. 53. 2013, pp. 1689–1699. arXiv: `arXiv:1011.1669v3` (cit. on pp. 17, 23).

[SSB03] Javad Sadri, Ching Y Suen, and Tien D. Bui. "Application of Support Vector Machines for recognition of handwritten Arabic/Persian digits." In: *Second Conference on Machine Vision and Image Processing & Applications (MVIP 2003)* 1 (2003), pp. 300–307 (cit. on p. 17).

[TC98] Christopher Torrence and Gilbert P. Compo. "A Practical Guide to Wavelet Analysis." In: *Bulletin of the American Meteorological Society* 79.1 (1998), pp. 61–78 (cit. on p. 34).

[Wan02] Jung-Ying Wang. "Support Vector Machines (SVM) in bioinformatics Bioinformatics applications." In: *Bioinformatics* (2002), pp. 1–56 (cit. on p. 17).

[WS09] K. Weinberger and L. Saul. "Distance Metric Learning for Large Margin Nearest Neighbor Classification." In: *Journal of Machine Learning Research* 10 (2009), pp. 207–244 (cit. on pp. 42, 43).

[Xi+06] Xiaopeng Xi et al. "Fast time series classification using numerosity reduction." In: *Proceedings of the 23rd international conference on Machine learning (ICML)*. 2006, pp. 1033–1040 (cit. on p. 16).

[YG08] J. Yin and M. Gaber. "Clustering distributed time series in sensor networks." In: *ICDM*. 2008 (cit. on p. 31).

[YL99] Yiming Yang and Xin Liu. "A re-examination of text categorization methods." In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 99*. 1999, pp. 42–49 (cit. on p. 17).

[G. 06] S. Thiria G. Dreyfus, J.-M. Martinez, M. Samuelides M. B. Gordon, F. Badran. *Apprentissage Apprentissage statistique*. Eyrolles. 2006, p. 471 (cit. on pp. 8, 11).

[Wie42] Wiener N. *Extrapolation, Interpolation & Smoothing of Stationary Time Series - With Engineering Applications*. Tech. rep. Report of the Services 19, Research Project DIC-6037 MIT, 1942 (cit. on p. 36).

Résumé — Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed non risus. Suspendisse lectus tortor, dignissim sit amet, adipiscing nec, ultricies sed, dolor. Cras elementum ultrices diam. Maecenas ligula massa, varius a, semper congue, euismod non, mi. Proin porttitor, orci nec nonummy molestie, enim est eleifend mi, non fermentum diam nisl sit amet erat. Duis semper. Duis arcu massa, scelerisque vitae, consequat in, pretium a, enim. Pellentesque congue. Ut in risus volutpat libero pharetra tempor. Cras vestibulum bibendum augue. Praesent egestas leo in pede. Praesent blandit odio eu enim. Pellentesque sed dui ut augue blandit sodales. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam nibh. Mauris ac mauris sed pede pellentesque fermentum. Maecenas adipiscing ante non diam sodales hendrerit. Ut velit mauris, egestas sed, gravida nec, ornare ut, mi. Aenean ut orci vel massa suscipit pulvinar. Nulla sollicitudin. Fusce varius, ligula non tempus aliquam, nunc turpis ullamcorper nibh, in tempus sapien eros vitae ligula. Pellentesque rhoncus nunc et augue. Integer id felis.

Mots clés : Série temporelle, Apprentissage de métrique, k-NN, SVM, classification, régression.

Abstract — Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed non risus. Suspendisse lectus tortor, dignissim sit amet, adipiscing nec, ultricies sed, dolor. Cras elementum ultrices diam. Maecenas ligula massa, varius a, semper congue, euismod non, mi. Proin porttitor, orci nec nonummy molestie, enim est eleifend mi, non fermentum diam nisl sit amet erat. Duis semper. Duis arcu massa, scelerisque vitae, consequat in, pretium a, enim. Pellentesque congue. Ut in risus volutpat libero pharetra tempor. Cras vestibulum bibendum augue. Praesent egestas leo in pede. Praesent blandit odio eu enim. Pellentesque sed dui ut augue blandit sodales. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aliquam nibh. Mauris ac mauris sed pede pellentesque fermentum. Maecenas adipiscing ante non diam sodales hendrerit. Ut velit mauris, egestas sed, gravida nec, ornare ut, mi. Aenean ut orci vel massa suscipit pulvinar. Nulla sollicitudin. Fusce varius, ligula non tempus aliquam, nunc turpis ullamcorper nibh, in tempus sapien eros vitae ligula. Pellentesque rhoncus nunc et augue. Integer id felis.

Keywords: Time series, Metric Learning, k-NN, SVM, classification, regression.

Schneider Electric
Université Grenoble Alpes, LIG
Université Grenoble Alpes, GIPSA-Lab