

MACHINE LEARNING

Overview



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Exam Format

The exam lasts a total of 3 hours:

- Upon entering the room, you must leave you bag, cell phone, etc, in a corner of the room; you are allowed to keep a couple of pen/pencil/ eraser and a few blank sheets of paper.

- Exam will be graded anonymously; make sure to have your camipro card with you to write your sciper number on your exam sheet

Exam is closed book but you can bring one A4 page with personal notes written recto-verso.

What to know for the exam

Formalism:

- Be capable of giving formal definitions of a pdf, marginal, likelihood
- Be capable of giving principle of basic ML algorithms such as maximum likelihood and E-M, of explaining which algorithm uses these (e.g. GMM, HMM)

Taxonomy:

- Know the difference between supervised / unsupervised learning and be able to give examples of algorithms in each case
- Be able to discuss concepts such as generative vs. discriminative methods (see comparison GMR vs SVR)

Principles of evaluation:

- Know the basic principles of evaluation of ML techniques: training vs. testing sets, crossvalidation, ground truth
- Know the principle of each method of evaluation seen in class and know which method of evaluation to apply where (F-measure in clustering vs. classification, BIC, etc)

What to know for the exam

- For each algorithm, be able to explain:
 - what it can do: classification, regression, structure discovery / reduction of dimensionality
 - what one should be careful about (limitations of the algorithm, choice of hyperparameters) and how does this choice influence the results.
 - the key steps of the algorithm, its hyperparameters, the variables it takes as input and the variables it outputs

What to know for the exam

- For each algorithm, be able to explain:

SVM

- what it can do: classification, regression, structure discovery / reduction of dimensionality

Performs *binary* classification; can be extended to multi-class classification; can be extended to regression (SVR)

- what one should be careful about (limitations of the algorithm, choice of hyperparameters)

e.g. choice of kernel; too small kernel width in Gaussian kernels may lead to over-fitting;

- the key steps of the algorithm, its hyperparameters, the variables it takes as input and the variables it outputs

Class Overview

This overview is meant to highlight similarities and differences across the different methods presented in class.

To be well prepared to the exam, read carefully the lecture notes and the slides.

Class Overview

This class has presented groups of methods for doing the classification, regression, structure discovery, estimation of time series.

Note that several algorithms do more than one of these types of computation.

Structure Discovery

PCA, ICA, LDA

K-Means, GMM

HMM

Classification

LDA, SVM, GMM + Bayes,
Boosting/Bagging

Regression

SVR
GMR

Time Series

HMM

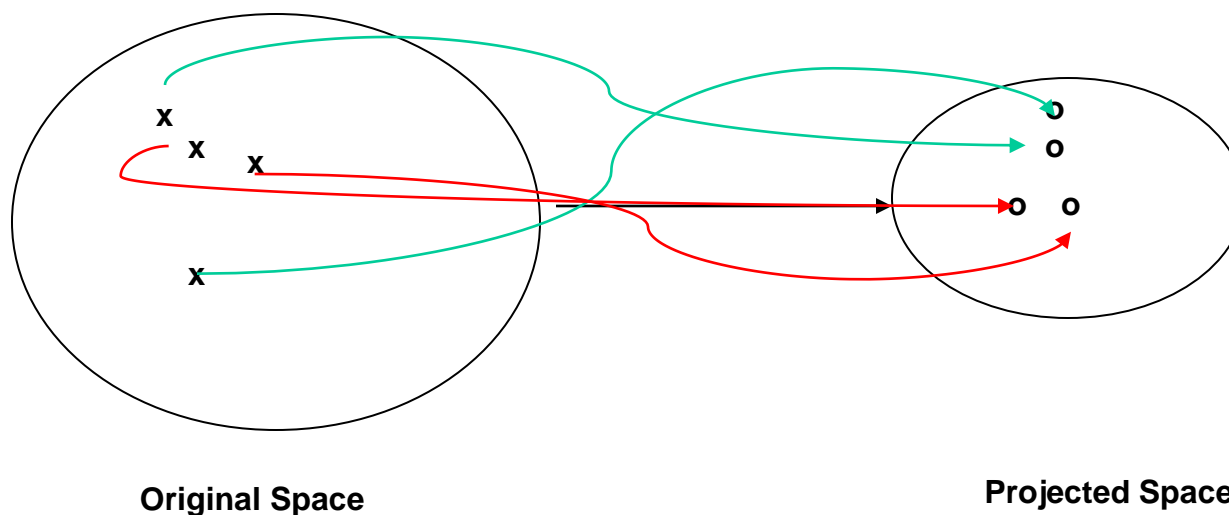
SVR, GMR with time as
input dimension

Overview: Finding Structure in Data

Techniques for finding structure in data proceed by projecting the data from the original space into another space of either lower dimension or higher dimension.

The projected space is chosen so as to highlight particular features common to subsets of datapoints.

The found structure may be exploited in a second stage by another algorithm for regression, classification, etc.



Overview: Finding Structure in Data

Classical techniques for finding structure in data:

- PCA
- ICA
- Clustering techniques (K-means, GMM)

But also implicit extraction of structure in more complex techniques such as:

- SVM, HMM

Overview: Finding Structure in Data

PCA, ICA and LDA: 3 examples of projection pursuit techniques

The term *projection pursuit* refers to a variety of algorithms that search for the most “interesting” projections.

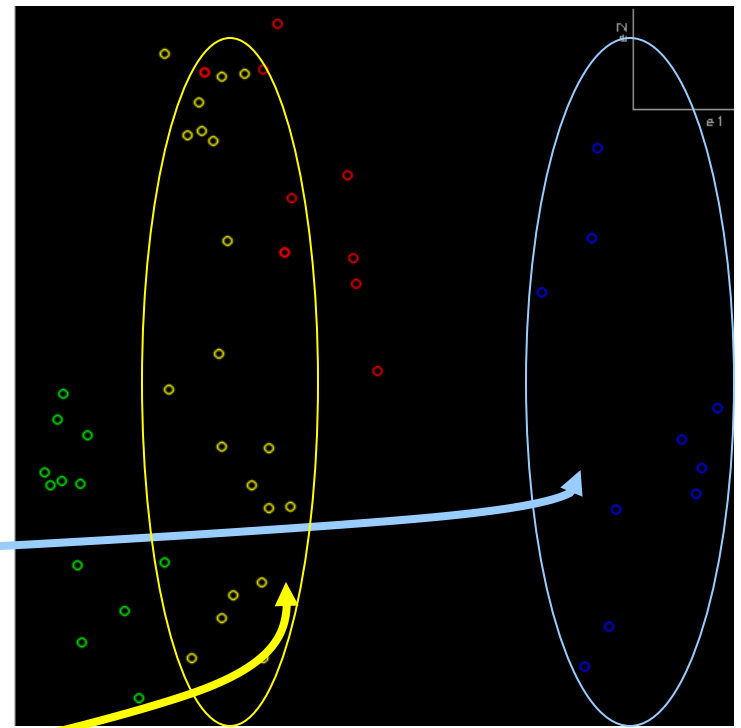
How interesting a projection depends on the task at hand.

Given a dataset $X \subset \mathbb{R}^{N \times M}$ and a unit vector $a \in \mathbb{R}^P$, $p \leq N$ one defines an index I_a that measures the *interest* of the associated projection $P_a(X)$.

Projection Pursuit finds the vector a that maximize I_a .

Overview: Finding Structure in Data

Principal Component Analysis (PCA)



$$x \in \mathbb{R}^N$$

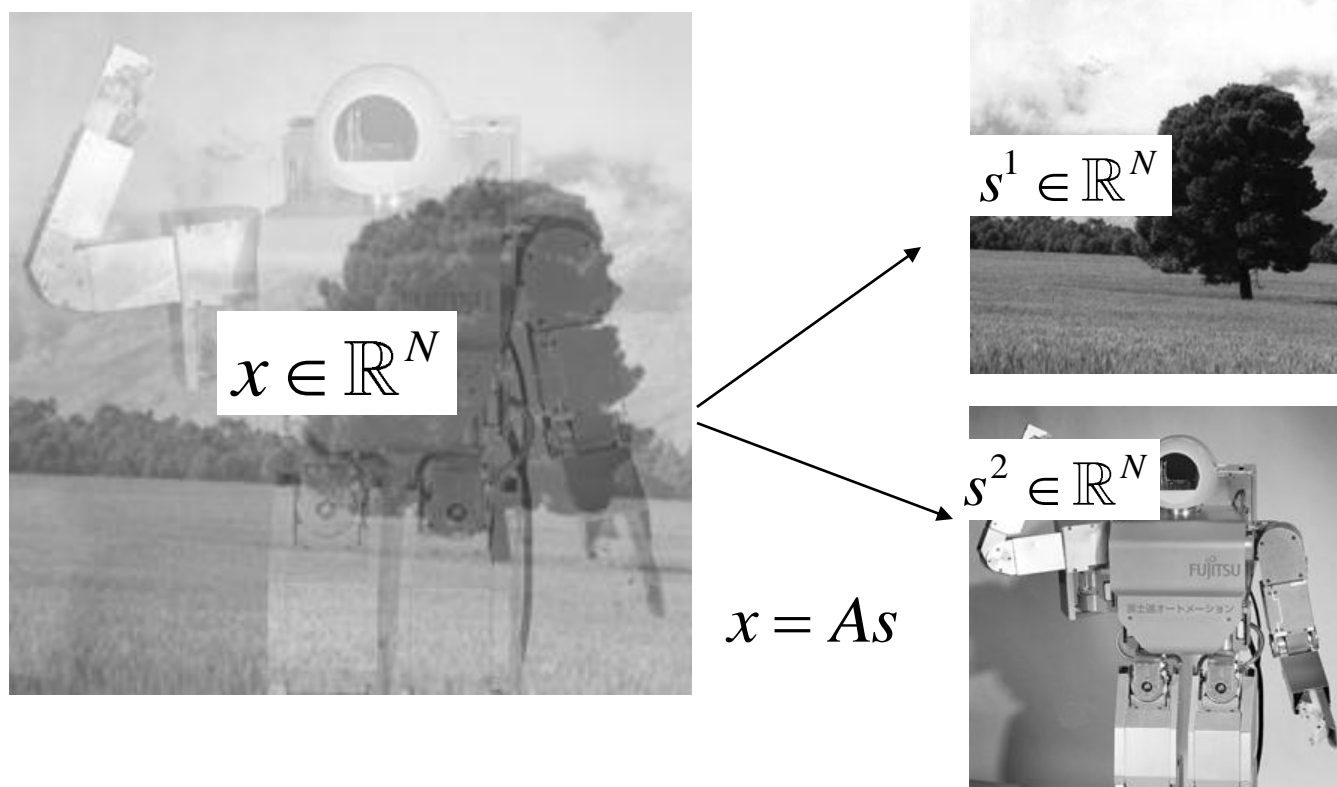
$$X' = WX$$

$$x' \in \mathbb{R}^q, \quad q \leq N$$

- Determines what is most common across datapoints.
- Projects onto axes that maximize correlation (eigenvectors of covariance matrix).
- Discard dimensions with the smallest eigenvalues.

Overview: Finding Structure in Data

Independent Component Analysis (ICA)



- Extract from the data the set of statistically independent sources:
- Determine the number of sources (e.g. with PCA first).
- Maximize statistical independence across projections: $p(s_1, s_2) = p(s_1)p(s_2)$
- Find each source iteratively (minimization of negentropy and orthog. of source)

Overview: Finding Structure in Data

Linear Discriminant Analysis (LDA)

$$A: x^i \in \mathbb{R}^N \rightarrow y^i \in \mathbb{R}^p \quad p \leq N$$

Given the scatter matrices S_w and S_b , find the linear projection through a map A that maximizes:

$$J(A) = \frac{|A^T S_b A|}{|A^T S_w A|}$$

which corresponds to:

increasing the between-class dissimilarity S_b (maximizes numerator)

increasing the within-class similarity S_w (minimizes denominator)

Clustering versus Classification

Fundamental difference between clustering and classification:

- Clustering is *unsupervised* classification
- Classification is *supervised* classification

Both use the F-measure but not in the same way.

The clustering F-measure assume a semi-supervised model, in which only a subset of the points are labelled

Clustering Methods

Except for hierarchical clustering, all three methods for clustering we have seen in class (K-means, soft K-means, GMM) are all solved through E-M (expectation-maximization).

You should be able to spell out the similarities and differences across K-means, soft K-means and GMM.

- They are similar in their representation of the problem and optimization method, etc.
- They differ in the number of parameters to estimate and number of hyper-parameters, etc.

Clustering Methods and Metric of Similarity

All clustering methods depend on choosing well a metric of similarity to measure how similar subgroup of data-points are.

You should be able to list which metric of similarity can be used in each case and how this choice may impact the clustering.

Kernel Methods

We have seen a single example of kernel method with SVM/SVR.

Kernel Methods implicit search for structure in the data prior to performing another computation (classification or regression)

- The kernel allows to extract *non-linear* types of correlations.

- These methods exploit the **Kernel Trick**:

The kernel trick exploits the observation that all *linear* methods for finding structure in data are based on computing an *inner product* across variables.

This inner product can be replaced by the kernel function if known. The problem becomes then *linear in feature space*.

$$k : X \times X \rightarrow \mathbb{R}$$

$$k(x^i, x^j) = \langle \phi(x^i), \phi(x^j) \rangle$$

Metric of similarity across datapoints

Overview: Regression Techniques

SVR and GMR lead to the following regressive model:

For a query point x^* , predict the associated output y^*

$$y^* = \sum_{i=1}^M \alpha_i k(x^*, x^i)$$

In SVR, the computation is reduced to summing only over the support vectors (a subset of datapoints)

In GMR, the sum is over the set of Gaussians. The centers of the Gaussians are usually not located on any particular datapoint.

Overview: Regression Techniques

SVR and GMR are based on the same probabilistic regressive model, but do not optimize the same objective function.

- SVR:
 - minimizes reconstruction error through convex optimization
 - finds a Nbr. of models \leq Nbr. of datapoints (support vectors)
- GMR:
 - learns $p(x,y)$ through likelihood maximization and then compute $p(y|x)$;
 - starts with a low Nbr. of models \ll Nbr. of datapoints

Overview of Topics Covered

This course covered a variety of topics that are core to Machine Learning. It gives you the basis to go and read recent advances in each of these topics.

We hope that you will find this material useful and that you will use some of these algorithms in the future.

If you do so, drop us a note and we would be glad to include your application in future lectures as examples!