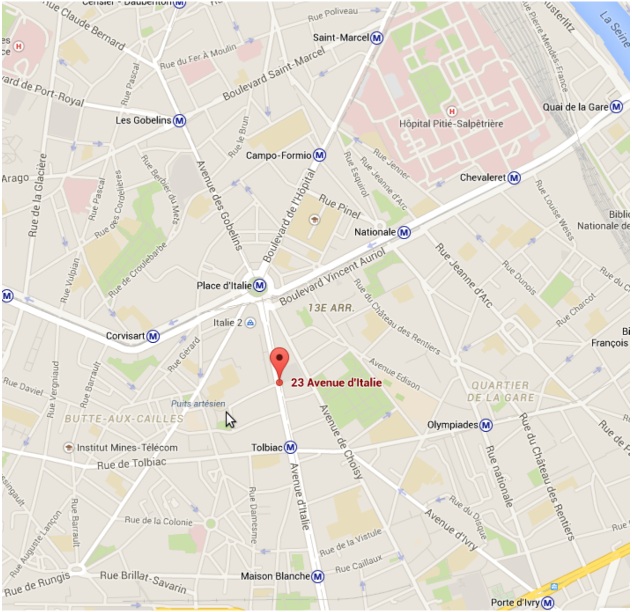
Machine Learning With Structured Outputs: a Glimpse Over the Topic

Rémi Lajugie

Inria Junior Seminar

April 21, 2015


23 Avenue d'Italie
Place d'Italie
Les Gobelins
Campo-Formio
Saint-Marcel
Quai de la Gare
Hôpital Pitié-Salpêtrière
Chevaleret
Nationale
Corvisart
Italie 2
Tolbiac
Olympiades
Maison Blanche
Porte d'Ivry
13E ARR.
BUTTE-AUX-CAILLES
QUARTIER DE LA GARE
Institut Mines-Télécom
Puits artésien
Boulevard Saint-Marcel
Boulevard de l'Hôpital
Boulevard Vincent Auriol
Avenue des Gobelins
Avenue d'Italie
Avenue de Choisy
Avenue d'Ivry
Rue de Tolbiac
Rue Brillat-Savarin
Rue de Rungis
Rue de la Colonie
Rue Pinel
Rue Jeanne d'Arc
Rue du Château des Rentiers
Rue Nationale
Rue Claude Bernard
Rue de la Glacière
Rue Poliveau
Rue du Fer À Moulin
Avenue Edison
Rue de la Vistule
Rue Caillaux
La Seine

Introduction

Some topics of interest among SIERRA people:

- Machine learning in a broad sense.

Introduction

Some topics of interest among SIERRA people:

- Machine learning in a broad sense.
- Signal processing (image, videos, audio...).

Introduction

Some topics of interest among SIERRA people:

- Machine learning in a broad sense.
- Signal processing (image, videos, audio...).
- Optimization (CONVEX !!!).

Introduction

Some topics of interest among SIERRA people:

- Machine learning in a broad sense.
- Signal processing (image, videos, audio...).
- Optimization (CONVEX !!!).
- Statistics (change-point detection problem).

Introduction

Some topics of interest among SIERRA people:

- Machine learning in a broad sense.
- Signal processing (image, videos, audio...).
- Optimization (CONVEX !!!).
- Statistics (change-point detection problem).
- Structured prediction.

This presentation

SIERRA Team

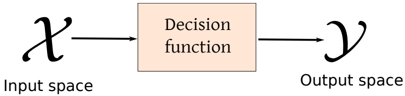
Machine Learning in a Nutshell

Structured Outputs in Machine Learning

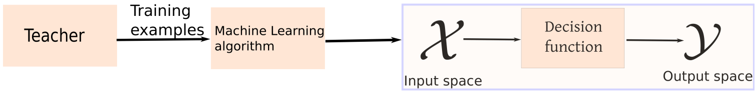
Dealing with Partial Information: Application to Computer Vision

Part I : Machine Learning in a nutshell

Supervised machine learning



VS



Standard Binary Classification Problem

- Ubiquitous in many real life applications (spam classification)
- The goal is to build a prediction function from annotated data.
- This is the supervised setting.

Toy example: Day/night classifier

- Sensor to measure light intensity let say $x \in \mathbb{R}$, you want to predict whether it is night (0) or day (1).

Toy example: Day/night classifier

- Sensor to measure light intensity let say $x \in \mathbb{R}$, you want to predict whether it is night (0) or day (1).
- The goal is to build a prediction function from annotated data.
- Modelize the classifier as being of the following form: $f(x) = 1_{F(x)>1.}$ where $F(x) = wx$ for some real $w \in \mathbb{R}$.

Toy example: Day/night classifier

- Sensor to measure light intensity let say $x \in \mathbb{R}$, you want to predict whether it is night (0) or day (1).
- The goal is to build a prediction function from annotated data.
- Modelize the classifier as being of the following form: $f(x) = 1_{F(x)>1.}$ where $F(x) = wx$ for some real $w \in \mathbb{R}$.
- Teacher gives you: (x_i, y_i).

Toy example: Day/night classifier

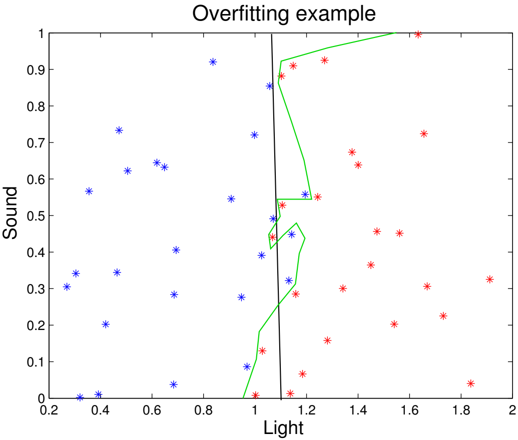
- Sensor to measure light intensity let say $x \in \mathbb{R}$, you want to predict whether it is night (0) or day (1).
- The goal is to build a prediction function from annotated data.
- Modelize the classifier as being of the following form: $f(x) = 1_{F(x)>1.}$ where $F(x) = wx$ for some real $w \in \mathbb{R}$.
- Teacher gives you: (x_i, y_i).
- Learn f (or F) by $\min_{w\in\mathbb{R}} \sum_i 1_{f(x)\neq y_i}$ (Empirical risk minimization).

In higher dimension (2) : the overfitting problem

- For now, the model is linear in the feature x, but what would have happen if we have let (assuming the underlying optimization problem is tractable) F be any function ?
- Now let us consider to be in dimension 2 (imagine that we have light intensity and volume of noise).

In higher dimension (2) : the overfitting problem

- For now, the model is linear in the feature x, but what would have happen if we have let (assuming the underlying optimization problem is tractable) F be any function ?
- Now let us consider to be in dimension 2 (imagine that we have light intensity and volume of noise).



-

Overfitting

- Fundamental tradeoff in machine learning

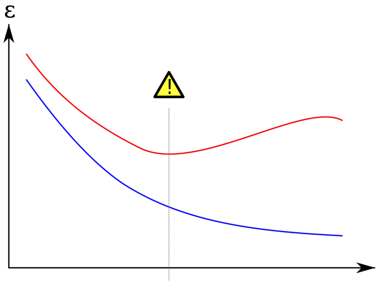


Figure: From wikipedia.

Overfitting

- Fundamental tradeoff in machine learning

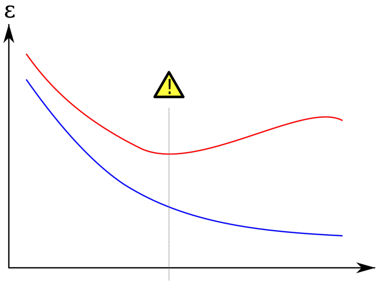


Figure: From wikipedia.

- Two ways to handle overfitting: either by restricting the class of function f you learn: $\min_{f \in \mathcal{F}} \sum_i 1_{f(x) \neq y_i}$
- Or: $\min_f \sum_i 1_{f(x) \neq y_i} + \lambda \Omega(f)$

Overfitting

- Fundamental tradeoff in machine learning

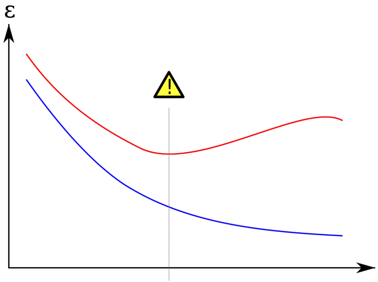


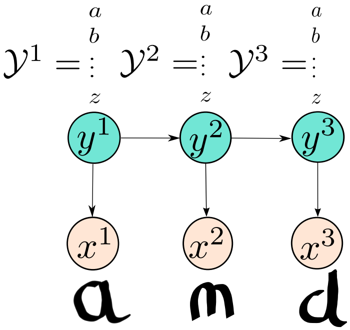
Figure: From wikipedia.

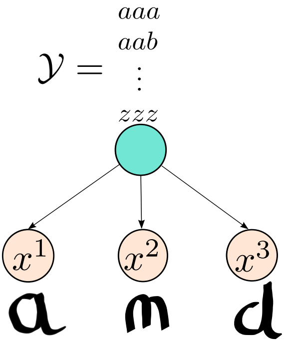
- Two ways to handle overfitting: either by restricting the class of function f you learn: $\min_{f \in \mathcal{F}} \sum_i 1_{f(x) \neq y_i}$
- Or: $\min_f \sum_i 1_{f(x) \neq y_i + \lambda \Omega(f)}$
- We need to adjust λ or $\mathcal{F}$ carefully.

Part II : What I care about: Structured outputs

Structured outputs

- Beyond binary classification.
- Structured outputs arises everywhere: genomics, finance, images, videos, audio signals,
- Historical example: The Optical character recognition problem.
- The idea was not to treat OCR as a sequence of binary classification problems.
- Structure occurs naturally. If two words differs from only one letter they should be closer.


$\mathcal{Y}^1 = \begin{matrix} a \\ b \\ \vdots \\ z \end{matrix}$
$\mathcal{Y}^2 = \begin{matrix} a \\ b \\ \vdots \\ z \end{matrix}$
$\mathcal{Y}^3 = \begin{matrix} a \\ b \\ \vdots \\ z \end{matrix}$
y^1
y^2
y^3
x^1
x^2
x^3


$\mathcal{Y} = \begin{matrix} aaa \\ aab \\ \vdots \\ zzz \end{matrix}$
x^1
x^2
x^3

What is different from binary classification ?

- Many classes... $\mathcal{Y}$ may be huge

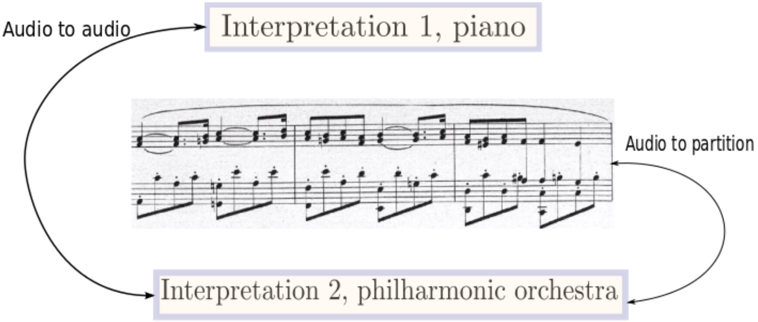
What is different from binary classification ?

- Many classes... $\mathcal{Y}$ may be huge
- Not all errors are equivalent ("sheep" closer of "ship" than "rotor"), need for a good loss ℓ.

What is different from binary classification ?

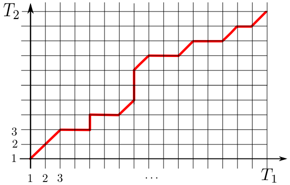
- Many classes... $\mathcal{Y}$ may be huge
- Not all errors are equivalent ("sheep" closer of "ship" than "rotor"), need for a good loss ℓ.
- Overall optimization program is: $\min_f \sum_i \ell(f(x) \neq y_i) + \lambda\Omega(f)$.

Introduction to my work



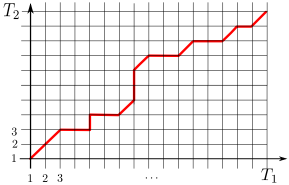
A more complex setting: Learning a Metric for Audio to Audio Alignment (Lajugie, Garreau et al., 2014)

- Inputs are pair of signals $X = (X_1, X_2) \in \mathbb{R}^{T_1 \times p} \times \mathbb{R}^{T_2 \times p}$.
- We denote by a_i the i-th row of X_1, and b_j the j-th row of X_2.
- The time warping problem consists in finding a path while respecting some constraints. The set of paths respecting these constraints is $\mathcal{Y}$.
- We assume to be given a similarity measure $s(a_i, a_j)$



A more complex setting: Learning a Metric for Audio to Audio Alignment (Lajugie, Garreau et al., 2014)

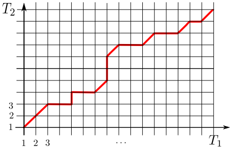
- Inputs are pair of signals $X = (X_1, X_2) \in \mathbb{R}^{T_1 \times p} \times \mathbb{R}^{T_2 \times p}$.
- We denote by a_i the i-th row of X_1, and b_j the j-th row of X_2.
- The time warping problem consists in finding a path while respecting some constraints. The set of paths respecting these constraints is $\mathcal{Y}$.
- We assume to be given a similarity measure $s(a_i, a_j)$



- We consider the alignment as the maximization of a certain criterion $S(X_1^i, X_2^i) = \max_{Y \in \mathcal{Y}} \mathrm{Tr}(CY)$ where $C_{i,j} = s(i,j)$ is some affinity matrix.
- $Y \in \mathcal{Y} \subset \{0,1\}^{T_1, T_2}$ is a binary matrix respecting alignment constraints.

Learning the Metric for Audio to Audio Alignment

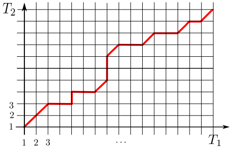
- Problem: How to set the similarity measure S ?



Learning the Metric for Audio to Audio Alignment

- Problem: How to set the similarity measure S ?
- Learn it from data!
- In some contexts we have audio representation in some high dimensional space (whole spectrogram) with a groundtruth alignment.
- Given N such annotated pairs of signals (X_1^i, X_2^i) with their optimal warping Y^i, we want to use the empirical risk minimization framework as in the binary case.
- Namely we want to

$$\min_{S \in \mathcal{S}} \sum_{i=1}^{N} \ell(S(X_1^i, X_2^i), Y^i) + \lambda \Omega(S).$$

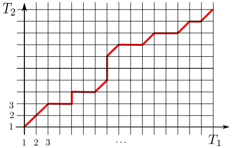


Learning the Metric for Audio to Audio Alignment

- Problem: How to set the similarity measure S ?
- Learn it from data!
- In some contexts we have audio representation in some high dimensional space (whole spectrogram) with a groundtruth alignment.
- Given N such annotated pairs of signals (X_1^i, X_2^i) with their optimal warping Y^i, we want to use the empirical risk minimization framework as in the binary case.
- Namely we want to

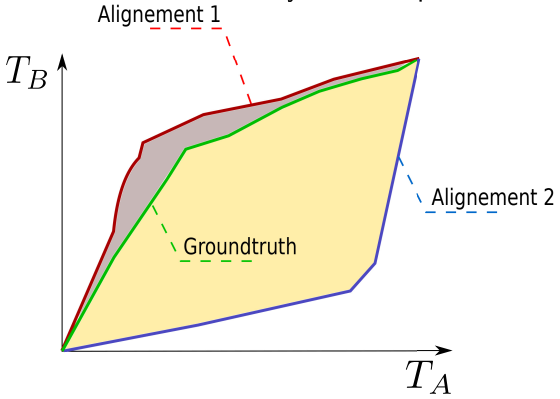
$$\min_{S \in \mathcal{S}} \sum_{i=1}^{N} \ell(S(X_1^i, X_2^i), Y^i) + \lambda \Omega(S).$$

- We need to find a good loss between alignments.

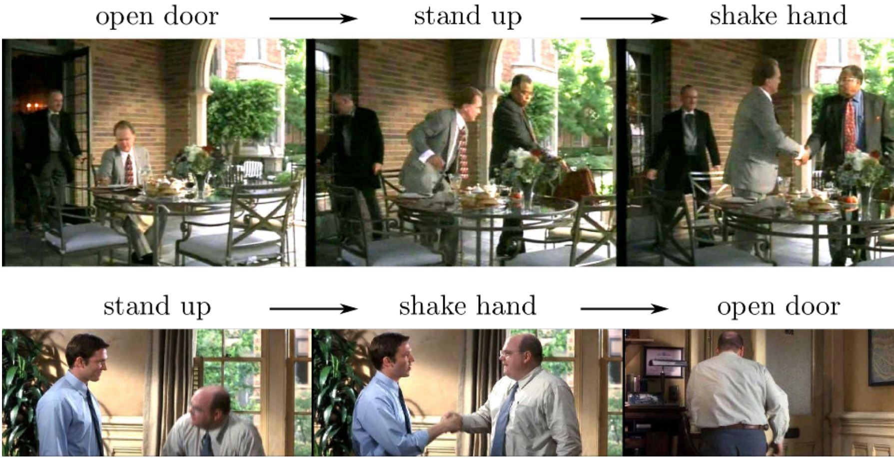


Good loss for the learning task.

- Simplest loss: Hamming (counting disagreements)
- Loss we are interested in: area.
- Have an interpretation in terms of delays with respect to onset of notes.

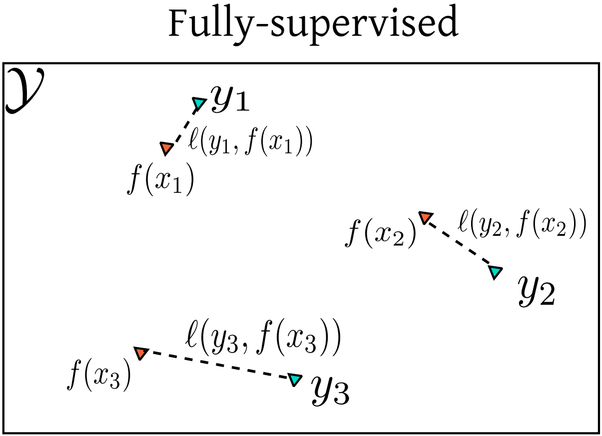


A practical problem: alignment of video scripts with video (Bojanowski, Lajugie et al., 2014)

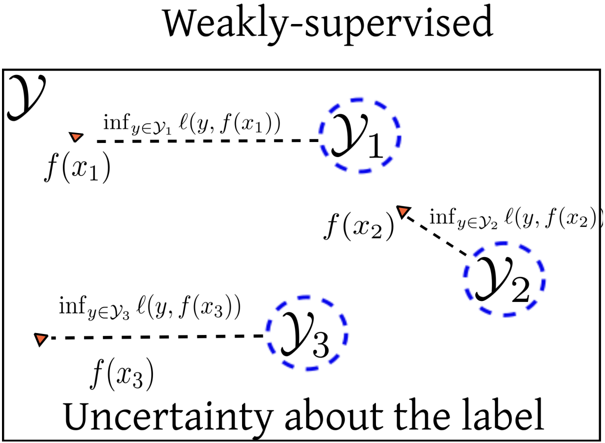


- We only know the temporal order of actions.
- We want to *localize* them.

Modelization of weak supervision (1)



Modelization of weak supervision (2)



Conclusion and perspectives

- We are working on the problem of audio to partition.
- Weak supervision is probably a major topic for the next few years.

Conclusion and perspectives

- We are working on the problem of audio to partition.
- Weak supervision is probably a major topic for the next few years.
- Thanks for your attention!