Structured Output Learning with Abstention

Application to Accurate Opinion Prediction

Alexandre Garcia
Chloé Clavel

Slim Essid
Florence d'Alché-Buc

LTCI, Télécom ParisTech, Paris, FRANCE
florence.dalche@telecom-paristech.fr

DATAIA-JST International Symposium on Data Science and AI

July 10, 2018

Outline

1. Short Overview of research activities
2. Focus on Structured Output Prediction with Abstention
3. Conclusion

Research activities

Laboratory: LTCI (permanent staff: 120)
Signal, Statistics and Machine Learning group (20 permanent staff members, 40 PhD students)

Albert Bifet - Pascal Bianchi - Thomas Bonald - Chloé Clavel - Stephan Clémençon

Jean-Louis Dessalles - James Eagan - Slim Essid - Olivier Fercoq - Pietro Gori

Robert Gower - Ons Jelassi - Laurence Likforman – François Portier – François Roueff

Mauro Sozio – Anne Sabourin – Joseph Salmon – Umut Şimşekli – Fabian Suchanek – Giovanna Varni

Focus on complex output learning

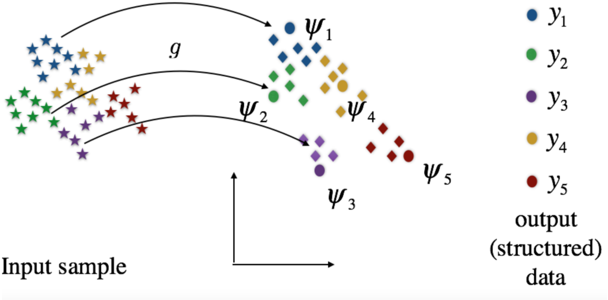
How to learn a function from $\mathcal{X}$ to $\mathcal{Y}$ when $\mathcal{Y}$: set of trees,(labeled) graphs, sequences, functions .... ?

- Multi-task learning: Multiple quantile regression
- Functional-valued Regression: Infinite-task learning
- Structured Output regression: Graph prediction in chemoinformatics
- Zero-shot learning: Predict a class/complex object never seen in the training data

New challenges: make it fast and efficient, make it robust and reliable !

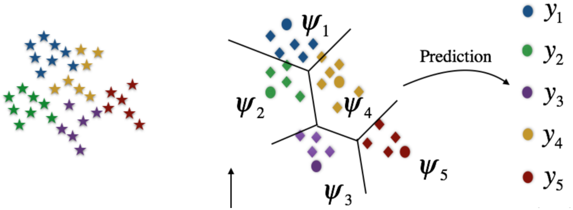
How do we solve these problems? solve an easier surrogate problem

(1) Transform your outputs and solve an easier problem in a well chosen output feature space



How do we solve these problems?

(2) Come back to the original output space by solving a pre-image problem



Outline

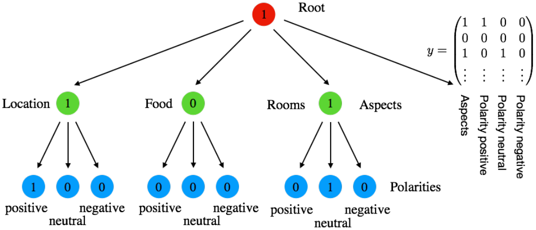
1. Short Overview of research activities
2. Focus on Structured Output Prediction with Abstention
   - Learning framework
   - Empirical results
3. Conclusion

Learning to label a structure with abstention

- Setup : we want to predict the labels of a known target graph structure (encoded by a directed graph).

TripAdvisor review ⇒ sentence level opinion annotations

```
The room was ok, nothing
special, still a perfect
choice to quickly join the
main places.
```

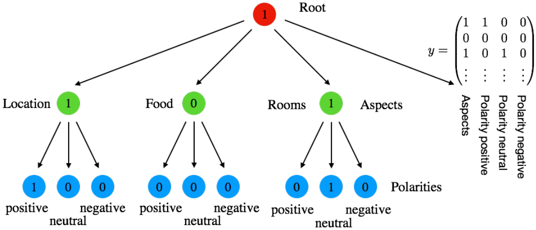


Learning to label a structure with abstention

- Setup : we want to predict the labels of a known target graph structure (encoded by a directed graph).

TripAdvisor review ⇒ sentence level opinion annotations

The room was OK, nothing special, still a perfect choice to quickly join the main places.

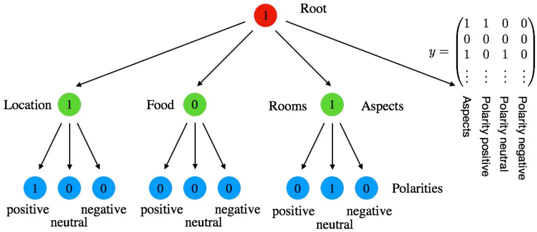


$$y = \begin{array}{cccc} \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \\ \text{Aspects} \quad \text{Polarity positive} \quad \text{Polarity neutral} \quad \text{Polarity negative} \end{array}$$

Problem

- Problem: Error at a node penalizes the prediction of descendants.

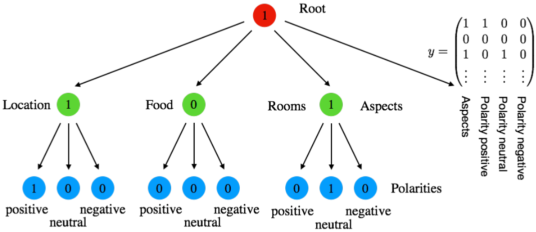
`The room was ok, nothing special, still a perfect choice to quickly join the main places.`



Problem

- Problem: Error at a node penalizes the prediction of descendants.

The room was ok, nothing special, still a perfect choice to $\underbrace{\texttt{quickly}}_{\texttt{Service, Checkin}}$ join the $\underbrace{\texttt{main places}}_{\texttt{Location}}$.



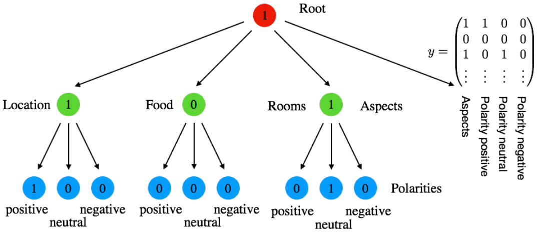
Problem

- Problem: Error at a node penalizes the prediction of descendants.

The room was ok, nothing special, still a perfect choice to quickly join the main places.

Service, Checkin

Location



$$y = \begin{array}{c} \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \\ \begin{array}{cccc} \text{Aspects} & \text{Polarity positive} & \text{Polarity neutral} & \text{Polarity negative} \end{array} \end{array}$$

Can we build a mechanism allowing to abstain on difficult nodes?

Problem

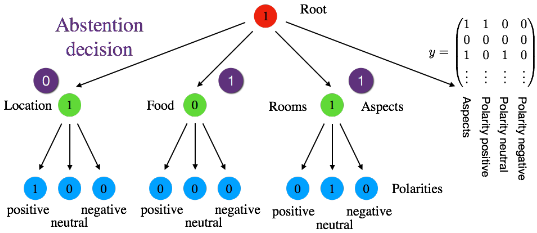
- Problem: Error at a node penalizes the prediction of descendants.

The room was ok, nothing special, still a perfect choice to quickly join

Service, Checkin

the main places.

Location



Can we build a mechanism allowing to abstain on difficult nodes?

Mathematical setup

- $\mathcal{X}$ an input sample space.
- $\mathcal{Y}$ the subset of $\{0,1\}^d$ that contains all possible legal labelings of an output structure $\mathcal{G}$.

Goal of Structured Output Learning with Abstention: learn a pair of functions (h, r) from $\mathcal{X}$ to $\mathcal{Y}^{H,R} \subset \{0,1\}^d \times \{0,1\}^d$ where a **predictor** h predicts the labels of $\mathcal{G}$ and an **abstention function** r chooses on which components of $\mathcal{G}$ to abstain from predicting a label.

- $\mathcal{Y}^\star \subset \{0,1,a\}^d$ is the set of legal labelings with abstention where a denotes the abstention label,
- Abstention-aware predictive model $f^{h,r} : \mathcal{X} \rightarrow \mathcal{Y}^\star$ defined by :

$$\begin{cases} f^{h,r}(x)^T & = [f_1^{h,r}(x), \ldots, f_d^{h,r}(x)], \\ f_i^{h,r}(x) & = 1_{h(x)_i=1} 1_{r(x)_i=1} + a 1_{r(x)_i=0}. \end{cases}$$

Learning setup

- $(x_i, y_i)_{i=1,\dots,n} \sim \mathcal{D}$ are n i.i.d. samples from a distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$.
- Suppose that we have access to an abstention aware loss $\Delta_a : \mathcal{Y}^{H,R} \times \mathcal{Y} \to \mathbb{R}^+$ then the risk of an abstention aware predictor is:

$$\mathcal{R}(h, r) = \mathbb{E}_{x,y\sim\mathcal{D}}\ \Delta_a(h(x), r(x), y).$$

Where Δ_a can be rewritten under the general form :

$$\Delta_a(h(x), r(x), y) = \langle \psi_{wa}(y), C\psi_a(h(x), r(x)) \rangle,$$

With $C : \mathbb{R}^p \to \mathbb{R}^q$ a bounded linear operator and $\psi_a : \mathcal{Y}^{H,R} \to \mathbb{R}^p$, $\psi_{wa} : \mathcal{Y} \to \mathbb{R}^q$ output embeddings.

Abstention-aware H-loss (Ha-loss)

$$\Delta_a(h(x), r(x), y) = \sum_{i=1}^{d} \underbrace{c_{Ai} 1_{\{f_i^{h,r}=a, f_{p(i)}^{h,r}=y_{p(i)}\}}}_{\text{abstention cost}} + \underbrace{c_{A_c i} 1_{\{f_i^{h,r}\neq y_i, f_{p(i)}^{h,r}=a\}}}_{\text{abstention regret}} + \underbrace{c_i 1_{\{f_i^{h,r}\neq y_i, f_{p(i)}^{h,r}=y_{p(i)}, a\neq f_i^{h,r}\}}}_{\text{misclassification cost}}$$

This loss writes as a inner product $\langle \psi_{wa}(y), C\psi_a(h(x), r(x))\rangle$.

Square surrogate framework

True risk:

$$\mathcal{R}(h, r) = \mathbb{E}_x \left\langle \mathbb{E}_{y|x} \psi_{wa}(y), C\psi_a(h(x), r(x)) \right\rangle.$$

Procedure :

- Solve a surrogate risk minimization problem :

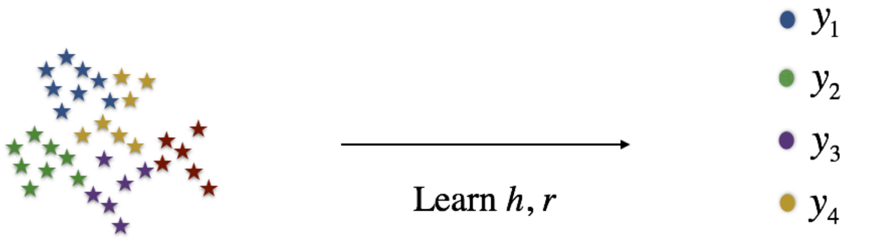
$$\min_{g \in \mathcal{H}} \underbrace{\mathbb{E}_{x,y} \|\psi_{wa}(y) - g(x)\|^2}_{\text{surrogate risk}}.$$

- Solve a *pre-image* problem

$$(\hat{h}(x), \hat{r}(x)) = \underset{(y_h, y_r) \in \mathcal{Y}^{H,R}}{\arg\min} \langle \hat{g}(x), C\psi_a(y_h, y_r) \rangle,$$

Square surrogate framework: intuition

$$\mathbb{E}_{x,y\sim\mathcal{D}}\Delta_a(h(x), r(x), y)$$



Learn h, r

● y_1

● y_2

● y_3

● y_4

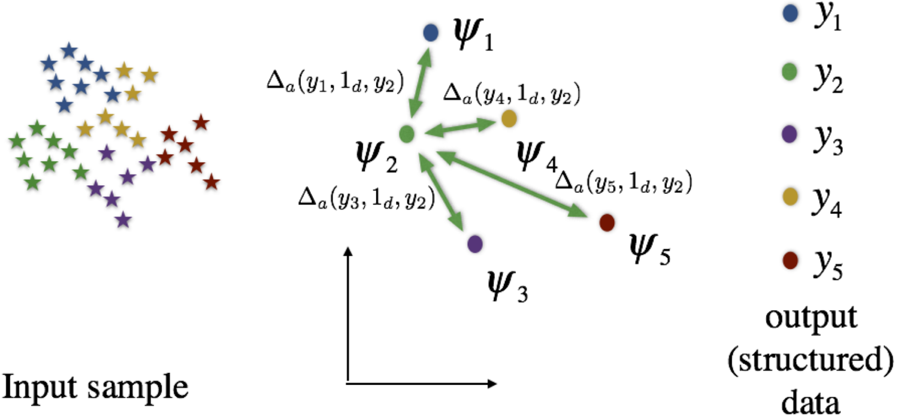
● y_5

Input sample

output (structured) data

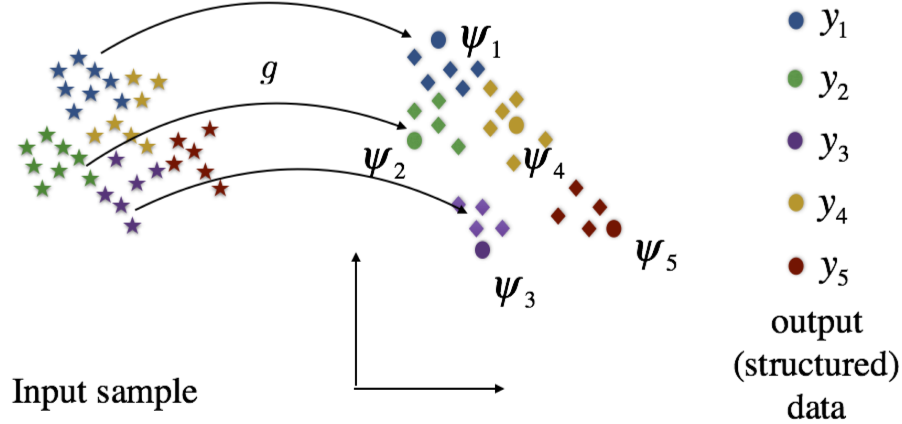
Square surrogate framework: intuition

$$\mathbb{E}_{x,y\sim\mathcal{D}}\Delta_a(h(x),r(x),y)$$



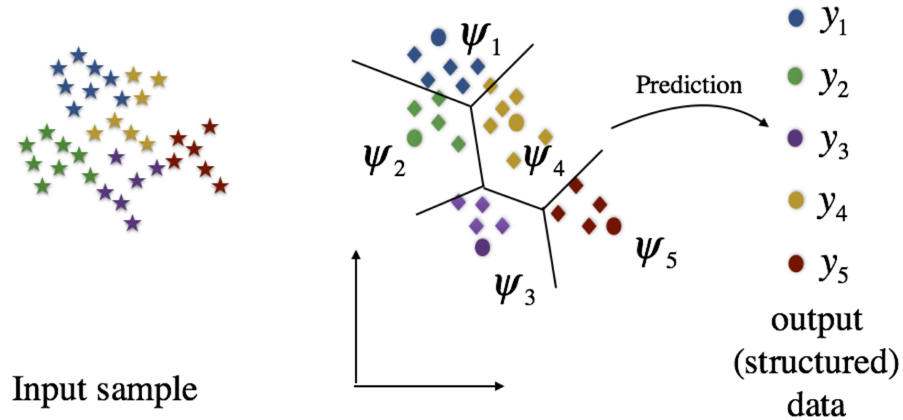
Square surrogate framework: intuition

$$\mathbb{E}_{x,y\sim\mathcal{D}}\Delta_a(h(x),r(x),y) \xrightarrow{\text{Replaced by}} \mathbb{E}_{x,y\sim\mathcal{D}}\|\psi_{wa}(y)-g(x)\|^2$$



Square surrogate framework: intuition

$$\mathbb{E}_{x,y\sim\mathcal{D}}\Delta_a(h(x),r(x),y) \xrightarrow{\text{Replaced by}} \mathbb{E}_{x,y\sim\mathcal{D}}\|\psi_{wa}(y)-g(x)\|^2$$



Surrogate risk minimization

Goal :

$$g^* = \min_{g \in \mathcal{H}} \underbrace{\mathbb{E}_{x,y} \|\psi_{wa}(y) - g(x)\|^2}_{\text{surrogate risk}}.$$

Based on an empirical sample $(x_i, y_i)_{i \in 1,\dots,n}$:

$$\hat{g} = \min_{g} \frac{1}{n} \sum_{i=1}^{n} \|\psi_{wa}(y_i) - g(x_i)\|^2 + \lambda \Omega(g),$$

Multivariate regression problem

$\mathcal{H}$: operator valued kernel, vector random forest, kNN, ...

Learning guarantees

Theorem

Based on the previous notations, the optimal predictor (h^*, r^*) is defined as:

$$(h^*(x), r^*(x)) = \underset{(y_h, y_r) \in \mathcal{Y}^{H,R}}{\arg\min} \langle C\psi_a(y_h, y_r), \mathbb{E}_{y|x}\psi_{wa}(y)\rangle.$$

The excess risk of an abstention aware predictor $(\hat{h}, \hat{r})$ defined from $\hat{g}$: $\mathcal{R}(\hat{h}, \hat{r}) - \mathcal{R}(h^\star, r^\star)$ is linked to the estimation error of the regression step.

$$\mathcal{R}(\hat{h}, \hat{r}) - \mathcal{R}(h^\star, r^\star) \leq 2c_l \sqrt{\mathcal{L}(\hat{g}) - \mathcal{L}(\mathbb{E}_{y|x}\psi_{wa}(y))},$$

where $\mathcal{L}(g) = \mathbb{E}_{x,y}\|\psi_{wa}(y) - g(x)\|^2$*, and* $c_l = \|C\| \max_{y_h, y_r \in \mathcal{Y}^{H,R}} \|\psi_a(y_h, y_r)\|_{\mathbb{R}^p}$.

Pre-image for hierarchical structures with abstention

Step 2 : Solve a *pre-image* problem

$$(\hat{h}(x), \hat{r}(x)) = \underset{(y_h, y_r) \in \mathcal{Y}^{H,R}}{\arg\min} \langle \hat{g}(x), C\psi_a(y_h, y_r) \rangle,$$

Problem: search over the set $\mathcal{Y}^{H,R}$ which is a subset of $\{0,1\}^d \times \{0,1\}^d$ under the constraint $A \begin{pmatrix} y_h \\ y_r \\ c \end{pmatrix} \leq b$.

Canonical form :

$$(\hat{h}(x), \hat{r}(x)) = \underset{(y_h, y_r)}{\arg\min} [y_h^T y_r^T c^T] M^T \psi_x$$

$$\text{s.t. } A_{\text{canonical}} \begin{pmatrix} y_h \\ y_r \\ c \end{pmatrix} \leq b_{\text{canonical}},$$

$$(y_h, y_r) \in \{0,1\}^d \times \{0,1\}^d,$$

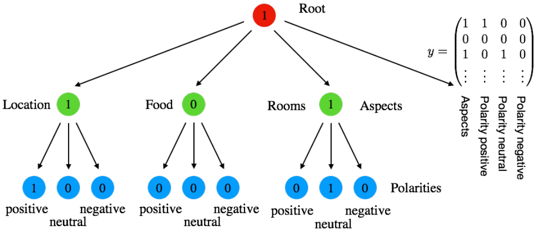
Where $A_{\text{canonical}}, b_{\text{canonical}}$ encode the constraints of A, b and the one of $\mathcal{Y}^{H,R}$

In general → NP-Hard

There exists polynomial time good initialization techniques.

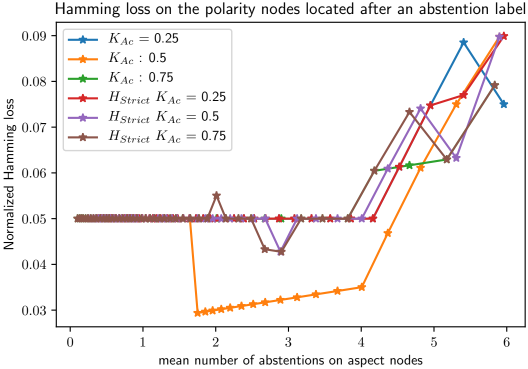
Experimental Setting

Dataset :
Input: TripAdvisor reviews annotated at the sentence level from [Marcheggiani et al. 2014]
We use the dense InferSent [Conneau et al. 2017] (feature representation for handling input data).



Experiments (subset): Joint Aspect and polarity prediction with abstention

Parameterization: $c_i = \frac{c_{p(i)}}{|\text{siblings(i)}|}$, $c_{Ai} = K_A c_i$, $c_{A_c i} = K_{A_c} c_i$.



Outline

1 Short Overview of research activities

2 Focus on Structured Output Prediction with Abstention

3 Conclusion

Conclusion and future works

- SOLA extends two families of approaches: learning with abstention and least-squares surrogate structured prediction.
- Beyond ridge regression, any vector-valued regression is eligible (including deep learning).
- Allows to build a robust representation for star rating in a pipeline framework.
- Beyond the target problem: develop general approaches to efficiently provide **robust** and **reliable** structured output prediction, whatever the underlying predictor architecture.
- Other ways to define $r(x)$: Bayesian approaches

References

- Hierarchical Multi-label Conditional Random Fields for Aspect-Oriented Opinion Mining, Marcheggiani, Diego and Täckström, Oscar and Esuli, Andrea and Sebastiani, Fabrizio, ECIR 2014.
- Supervised Learning of Universal Sentence Representations from Natural Language Inference Data, A. Conneau and D. Kiela and H. Schwenk and L. Barrault and A.Bordes, arxiv, 2017.
- Structured Output Learning with Abstention, A. Garcia, C. Clavel, S. Essid, F. d'Alché-Buc, ICML 2018.
- Output Fisher Embedding Regression, M. Djerrab, A. Garcia, M. Sangnier, F. d'Alché-Buc, ECML/PKDD 2018 and MLJ, May 2018