Audition Prof. Assistant en Sciences des Données et Intelligence Artificielle

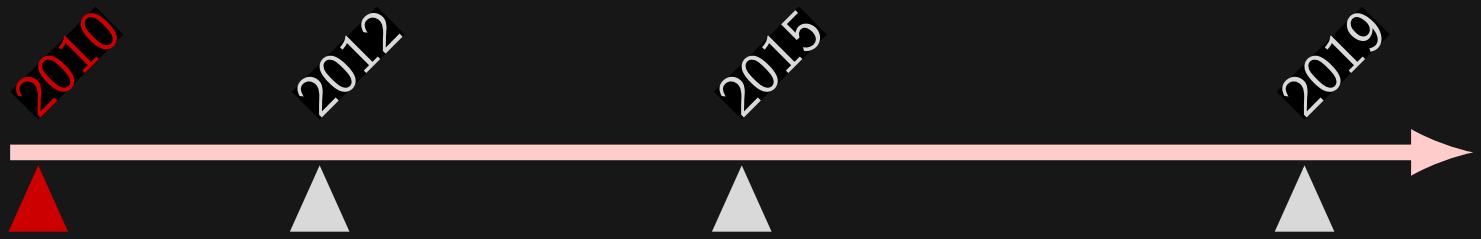
Adrien Ehrhardt

14/05/2019

Table of Contents

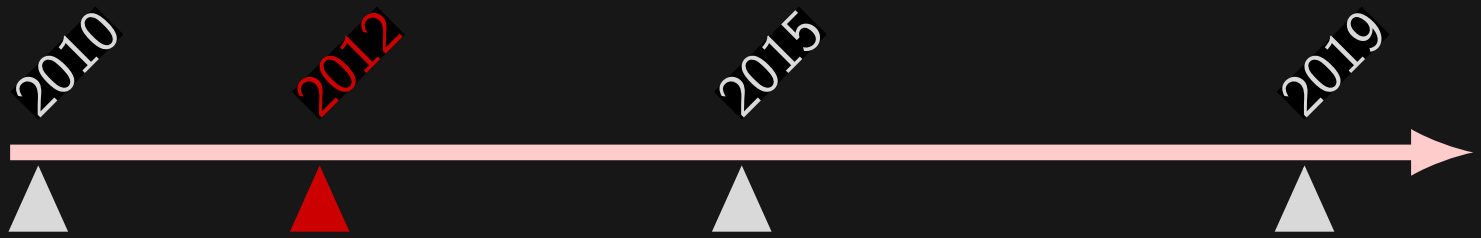
Short Bio

Short Bio



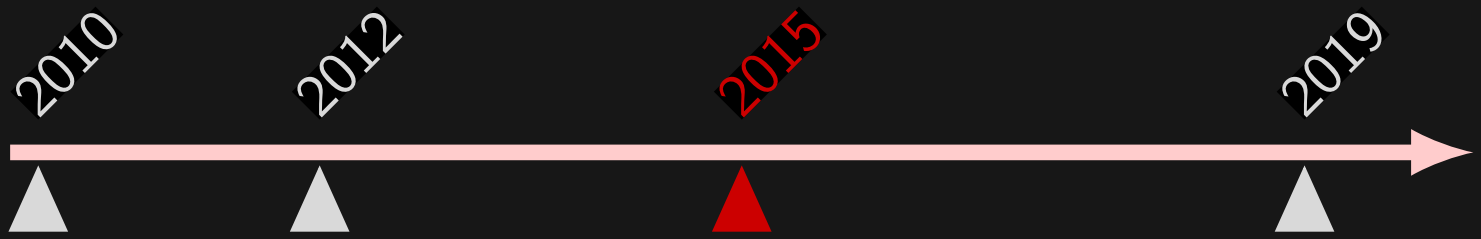
- 2010: BAC S & start of MPSI / MPI preparatory classes in Strasbourg;

Short Bio



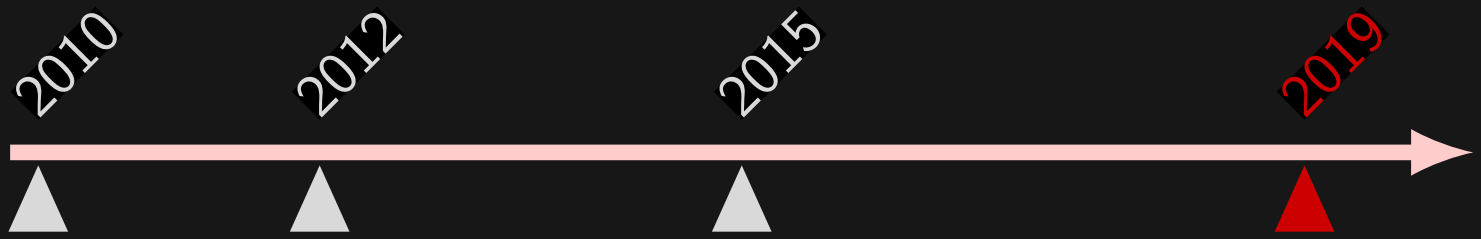
- 2010: BAC S & start of MPSI / MPI preparatory classes in Strasbourg;
- 2012: École Centrale de Lille;
  - 2014: B.Sc. (Licence 3) of Mathematics at Université de Lille;
  - 2014: Research project in integer programming;
  - 2014: Internship at Eiffage Group $\approx$ penalized regression;
  - 2015: "Décision et Analyse de Données" specialization + apprenticeship at BNP Paribas as a Quant. Analyst.

Short Bio



- 2010: BAC S & start of MPSI / MPI preparatory classes in Strasbourg;
- 2012: École Centrale de Lille;
  - 2014: B.Sc. (Licence 3) of Mathematics at Université de Lille;
  - 2014: Research project in integer programming;
  - 2014: Internship at Eiffage Group $\approx$ penalized regression;
  - 2015: "Décision et Analyse de Données" specialization + apprenticeship at BNP Paribas as a Quant. Analyst.
- 2015: Crédit Agricole Consumer Finance
  - 11/2015: Data Scientist position (permanent contract);
  - 04/2016: Official start of the CIFRE funding;

Short Bio



- 2010: BAC S & start of MPSI / MPI preparatory classes in Strasbourg;
- 2012: École Centrale de Lille;
  - 2014: B.Sc. (Licence 3) of Mathematics at Université de Lille;
  - 2014: Research project in integer programming;
  - 2014: Internship at Eiffage Group $\approx$ penalized regression;
  - 2015: "Décision et Analyse de Données" specialization + apprenticeship at BNP Paribas as a Quant. Analyst.
- 2015: Crédit Agricole Consumer Finance
  - 11/2015: Data Scientist position (permanent contract);
  - 04/2016: Official start of the CIFRE funding;
- 2019:
  - 04/2019: Official end of the CIFRE funding;
  - Mid-07/2019: Defense;
  - 09/2019: Part-time Ass. Prof. + (current position + telework | position in the "Groupe de Recherches Opérationnelles").

Research

Research: the industrial setting

Job	Home	Time in job	Family status	Wages		Repay-ment
Craftsman	Owner	20	Widower	2000		0
?	Renter	10	Common-law	1700		1
Licensed profes-sional	Starter	5	Divorced	4000		0
Executive	By work	8	Single	2700		1
Office employee	Renter	12	Married	1400		NA
Worker	By family	2	?	1200		NA

Table: Dataset with outliers and missing values.

Research: the industrial setting

Job	Home	Time in job	Family status	Wages		Repay-ment
Craftsman	Owner	20	Widower	2000		0
?	Renter	10	Common-law	1700		1
Licensed profes-sional	Starter	5	Divorced	4000		0
Executive	By work	8	Single	2700		1
Office employee	Renter	12	Married	1400		NA
Worker	By family	2	?	1200		NA

Table: Dataset with outliers and missing values.

1. Discarding rejected applicants
2. Feature selection
3. Discretization / grouping
4. Interaction screening
5. Segmentation
6. Logistic regression fitting

Research: the industrial setting

Job	Home	Time in job	Family status	Wages		Repay-ment
Craftsman	Owner	20	Widower	2000		0
?	Renter	10	Common-law	1700		1
Licensed profes-sional	Starter	5	Divorced	4000		0
Executive	By work	8	Single	2700		1
~~Office employee~~	~~Renter~~	~~12~~	~~Married~~	~~1400~~		NA
~~Worker~~	~~By family~~	~~2~~	~~?~~	~~1200~~		NA

Table: Dataset with outliers and missing values.

1. **Discarding rejected applicants**
2. Feature selection
3. Discretization / grouping
4. Interaction screening
5. Segmentation
6. Logistic regression fitting

Research: the industrial setting

Job			Family status	Wages		Repay-ment
Craftsman			Widower	2000		0
?			Common-law	1700		1
Licensed profes-sional			Divorced	4000		0
Executive			Single	2700		1
~~Office employee~~	~~Renter~~	~~12~~	~~Married~~	~~1400~~		NA
~~Worker~~	~~By family~~	~~2~~	~~?~~	~~1200~~		NA

Table: Dataset with outliers and missing values.

1. Discarding rejected applicants
2. **Feature selection**
3. Discretization / grouping
4. Interaction screening
5. Segmentation
6. Logistic regression fitting

Research: the industrial setting

Job			Family status	**Wages**		Repay-ment
Craftsman			Widower	**]1500;2000]**		0
?			Common-law	]1500;2000]		1
Licensed profes-sional			Divorced	]2000;∞[		0
Executive			Single	]2000;∞[		1
~~Office employee~~	~~Renter~~	~~12~~	~~Married~~	~~1400~~		NA
~~Worker~~	~~By family~~	~~2~~	~~?~~	~~1200~~		NA

Table: Dataset with outliers and missing values.

1. Discarding rejected applicants
2. Feature selection
3. **Discretization** / grouping
4. Interaction screening
5. Segmentation
6. Logistic regression fitting

Research: the industrial setting

Job			Family status	Wages		Repay-ment
?+Low-qualified			?+Alone	]1500;2000]		0
?+Low-qualified			Union	]1500;2000]		1
High-qualified			?+Alone	]2000;∞[		0
High-qualified			?+Alone	]2000;∞[		1
~~Office employee~~	~~Renter~~	~~12~~	~~Married~~	~~1400~~		NA
~~Worker~~	~~By family~~	~~2~~	~~?~~	~~1200~~		NA

Table: Dataset with outliers and missing values.

1. Discarding rejected applicants
2. Feature selection
3. Discretization / **grouping**
4. Interaction screening
5. Segmentation
6. Logistic regression fitting

Research: the industrial setting

Job			Family status x Wages		Repay-ment
?+Low-qualified			?+Alone x]1500;2000]		0
?+Low-qualified			**Union x]1500;2000]**		1
High-qualified			**?+Alone x]2000;∞[**		0
High-qualified			**?+Alone x]2000;∞[**		1
~~Office employee~~	~~Renter~~	~~12~~	~~Married~~ ~~1400~~		NA
~~Worker~~	~~By family~~	~~2~~	~~?~~ ~~1200~~		NA

Table: Dataset with outliers and missing values.

1. Discarding rejected applicants
2. Feature selection
3. Discretization / grouping
4. **Interaction screening**
5. Segmentation
6. Logistic regression fitting

Research: the industrial setting

Job			Family status x Wages		Repay-ment
?+Low-qualified			?+Alone x]1500;2000]		0
?+Low-qualified			Union x]1500;2000]		1
High-qualified			?+Alone x]2000;∞[		0
High-qualified			?+Alone x]2000;∞[		1
~~Office employee~~	~~Renter~~	~~12~~	~~Married~~ ~~1400~~		NA
~~Worker~~	~~By family~~	~~2~~	~~?~~ ~~1200~~		NA

Table: Dataset with outliers and missing values.

1. Discarding rejected applicants
2. Feature selection
3. Discretization / grouping
4. Interaction screening
5. **Segmentation**
6. Logistic regression fitting

Research: the industrial setting

Job			Family status x Wages	Score	Repay-ment
?+Low-qualified			?+Alone x]1500;2000]	225	0
?+Low-qualified			Union x]1500;2000]	190	1
High-qualified			?+Alone x]2000;∞[	218	0
High-qualified			?+Alone x]2000;∞[	202	1
~~Office employee~~	~~Renter~~	~~12~~	~~Married~~ ~~1400~~	NA	NA
~~Worker~~	~~By family~~	~~2~~	~~?~~ ~~1200~~	NA	NA

Table: Dataset with outliers and missing values.

1. Discarding rejected applicants
2. Feature selection
3. Discretization / grouping
4. Interaction screening
5. Segmentation
6. **Logistic regression fitting**

Research: four main topics tackled in the PhD

1. "Reject Inference"
   2 conferences; 1 ongoing paper; 1 R package.

Research: four main topics tackled in the PhD

1. "Reject Inference"
   2 conferences; 1 ongoing paper; 1 R package.

2. "Quantization": discretization + grouping.
   1 conference; 1 submitted preprint; 2 packages.

Research: four main topics tackled in the PhD

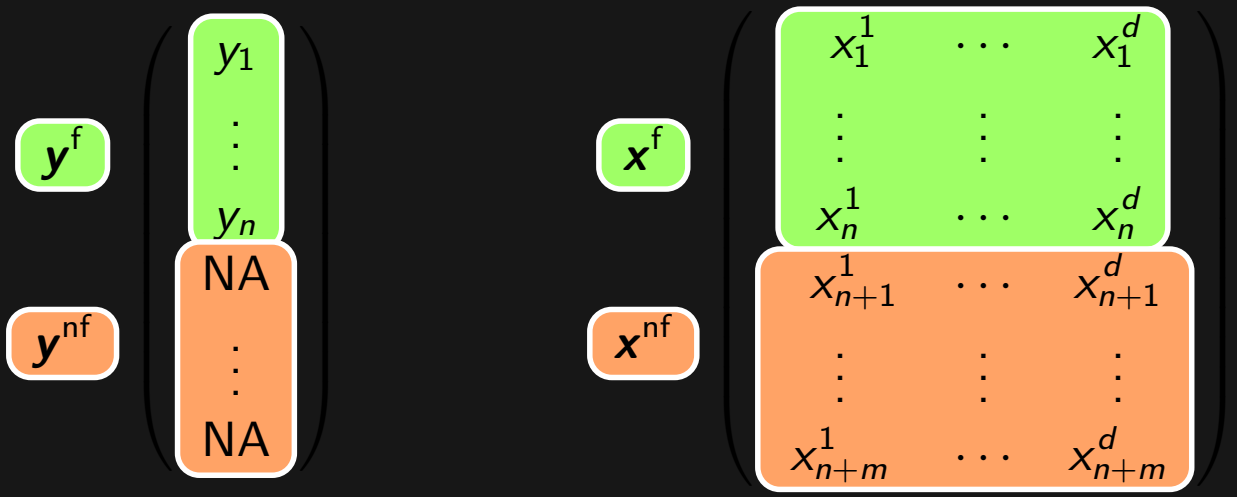
1. "Reject Inference"
   2 conferences; 1 ongoing paper; 1 R package.
2. "Quantization": discretization + grouping.
   1 conference; 1 submitted preprint; 2 packages.
3. Pairwise interactions.
   1 ongoing paper (w. another estimation technique for quantization).

Research: four main topics tackled in the PhD

1. "Reject Inference"
   2 conferences; 1 ongoing paper; 1 R package.
2. "Quantization": discretization + grouping.
   1 conference; 1 submitted preprint; 2 packages.
3. Pairwise interactions.
   1 ongoing paper (w. another estimation technique for quantization).
4. Segmentation

Research: reject inference

We have the following available data:



We usually fit a logistic regression using the observed data:

$$\hat{\theta}^f = \text{argmin}\, \text{CRIT}(\theta; \boldsymbol{x}^f, \boldsymbol{y}^f).$$

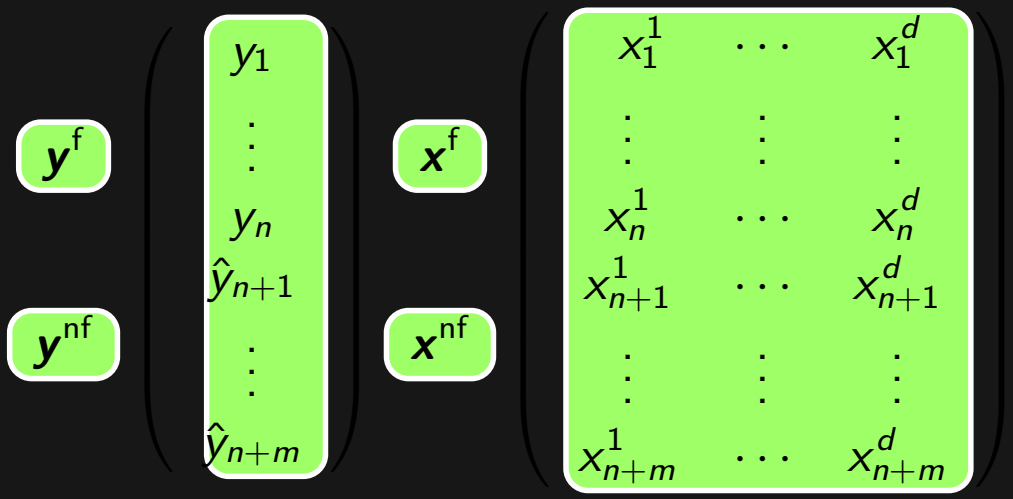
We wish we had:

$$\hat{\theta} = \text{argmin}\, \text{CRIT}(\theta; \boldsymbol{x}, \boldsymbol{y}).$$

Which cannot be computed since we lack $\boldsymbol{y}^{nf}$.
What's the relation between $\hat{\theta}^f$ and $\hat{\theta}$?

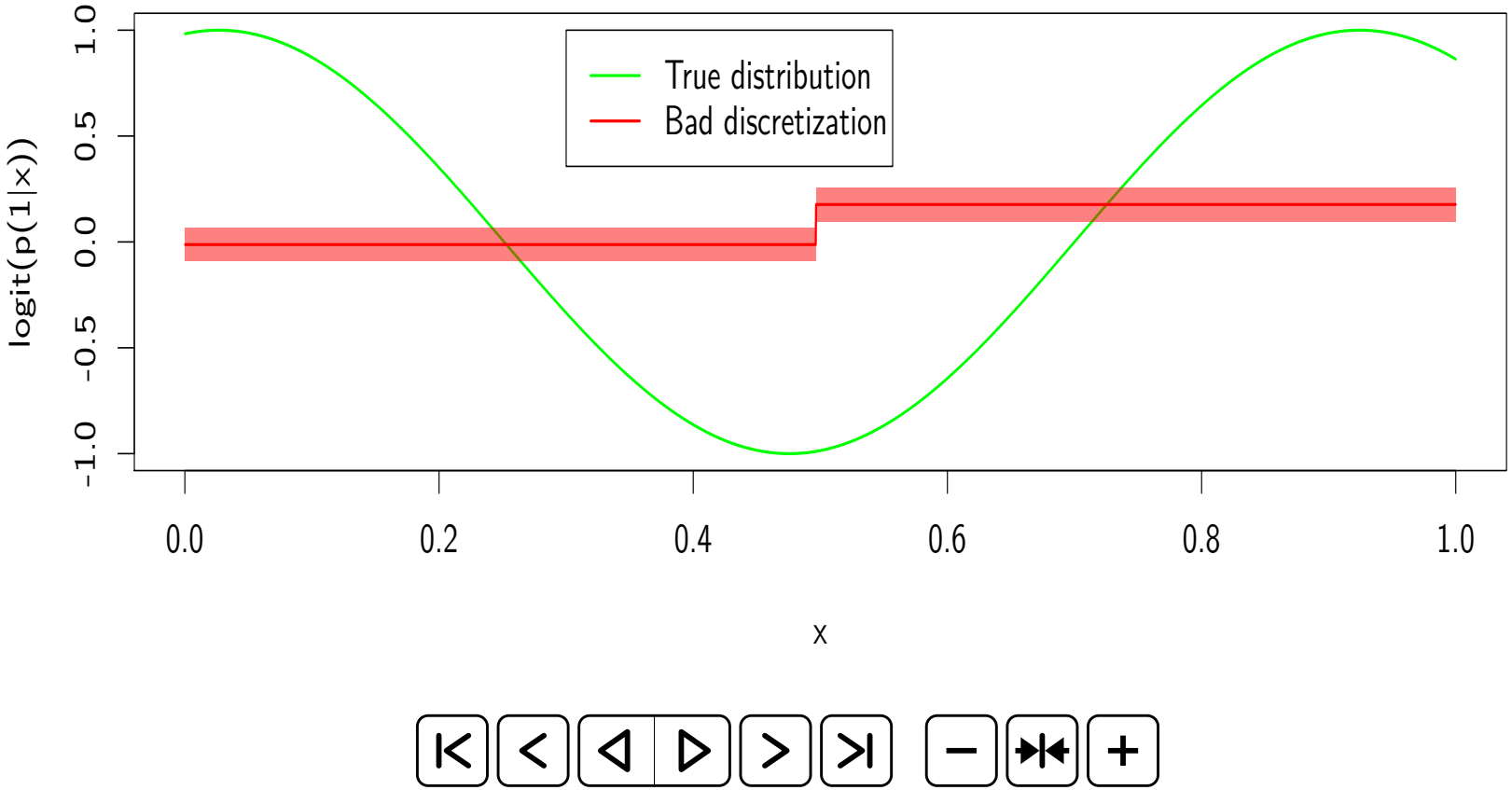
Research: reject inference

For logistic regression, *ad hoc* were proposed:



$\Rightarrow$ formalize and justify these methods, if possible.

Research: quantization



Research: quantization

Problem setting:
"Constrained" representation learning:

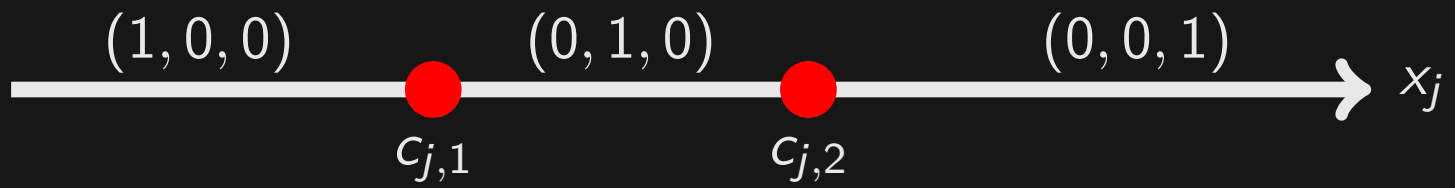
$$(\boldsymbol{q}^\star, \boldsymbol{\theta}^\star) = \underset{q \in \mathcal{Q}, \boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \ \mathrm{CRIT}(\hat{\boldsymbol{\theta}}_q).$$

Research: quantization

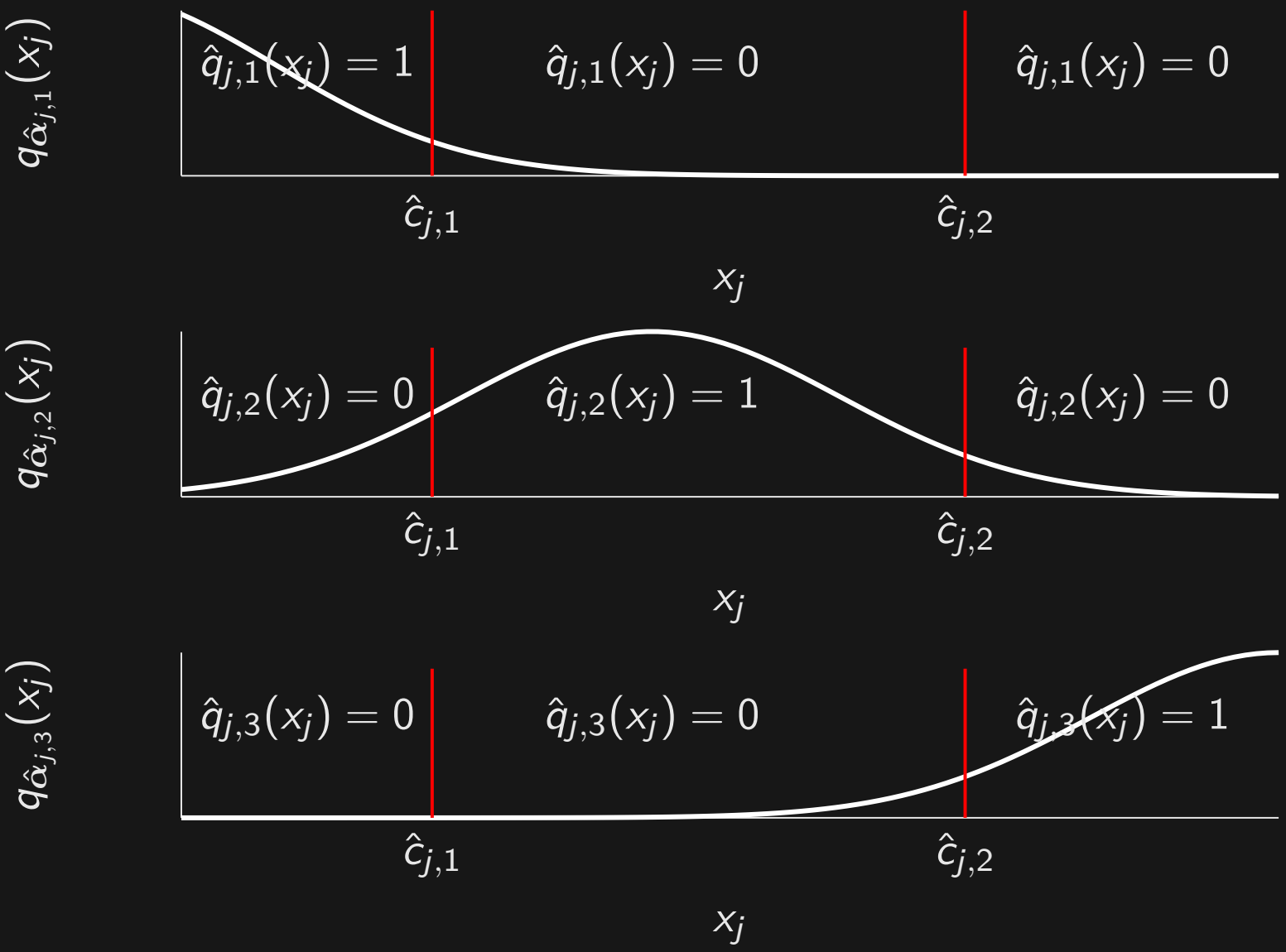
Problem setting:
"Constrained" representation learning:

$$(\boldsymbol{q}^\star, \boldsymbol{\theta}^\star) = \underset{q \in \mathcal{Q}, \boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \ \mathrm{CRIT}(\hat{\boldsymbol{\theta}}_q).$$

Resolution:
Relaxation of the search for the number of cutpoints and their location as a continuous problem.



Research: quantization



Research: interactions

Problem setting:

$$\text{logit}(p_\theta(1|\boldsymbol{x})) = \theta_0 + x'\theta + \sum_{1 \le k < \ell \le d} \delta_{k,\ell} x_k' \theta_{k,\ell} x_\ell$$

Research: interactions

Problem setting:

$$\text{logit}(p_\theta(1|\boldsymbol{x})) = \theta_0 + x'\theta + \sum_{1 \le k < \ell \le d} \delta_{k,\ell} x_k' \theta_{k,\ell} x_\ell$$

$$(\boldsymbol{\theta}^\star, \boldsymbol{\delta}^\star) = \underset{\boldsymbol{\theta}, \boldsymbol{\delta} \in \{0,1\}^{\frac{d(d-1)}{2}}}{\text{argmax}} \text{CRIT}(\hat{\theta}_\delta)$$

Research: interactions

Problem setting:

$$\text{logit}(p_\theta(1|\boldsymbol{x})) = \theta_0 + x'\theta + \sum_{1 \le k < \ell \le d} \delta_{k,\ell} x_k' \theta_{k,\ell} x_\ell$$

$$(\boldsymbol{\theta}^\star, \boldsymbol{\delta}^\star) = \underset{\boldsymbol{\theta}, \boldsymbol{\delta} \in \{0,1\}^{\frac{d(d-1)}{2}}}{\text{argmax}} \text{CRIT}(\hat{\theta}_\delta)$$

Analogous to previous problem: $2^{\frac{d(d-1)}{2}}$ models.

Research: interactions

Problem setting:

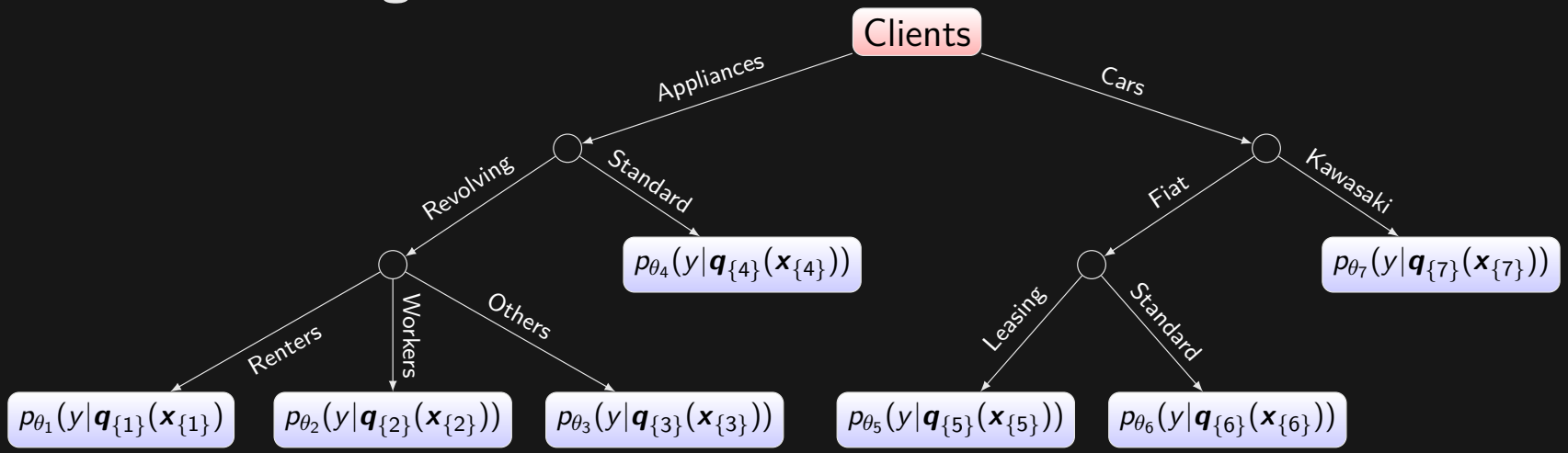
$$\text{logit}(p_\theta(1|\boldsymbol{x})) = \theta_0 + x'\theta + \sum_{1 \le k < \ell \le d} \delta_{k,\ell} x_k' \theta_{k,\ell} x_\ell$$

$$(\boldsymbol{\theta}^\star, \boldsymbol{\delta}^\star) = \underset{\boldsymbol{\theta}, \boldsymbol{\delta} \in \{0,1\}^{\frac{d(d-1)}{2}}}{\text{argmax}} \text{CRIT}(\hat{\theta}_\delta)$$

Analogous to previous problem: $2^{\frac{d(d-1)}{2}}$ models.
δ is latent and hard to optimize over: use a stochastic algorithm!

Research: logistic regression trees

Problem setting:



Research: logistic regression trees

Current methodology: unsupervised generative approaches such as PCA.

State-Of-The-Art: LOTUS[1] / LMT[2] / MOB[3].

[1] Kin-Yee Chan and Wei-Yin Loh. "LOTUS: An algorithm for building accurate and comprehensible logistic regression trees". In: Journal of Computational and Graphical Statistics 13.4 (2004), pp. 826–852.

[2] Niels Landwehr, Mark Hall, and Eibe Frank. "Logistic model trees". In: Machine learning 59.1-2 (2005), pp. 161–205.

[3] Achim Zeileis, Torsten Hothorn, and Kurt Hornik. "Model-based recursive partitioning". In: Journal of Computational and Graphical Statistics 17.2 (2008), pp. 492–514.

Research: logistic regression trees

Current methodology: unsupervised generative approaches such as PCA.

State-Of-The-Art: LOTUS[1] / LMT[2] / MOB[3].

Proposed method: smooth stochastic relaxation similar to the quantization problem.

Results: Better than SOTA on simulated data.
Works well on real data where SOTA methods are intractable.

---

[1] Kin-Yee Chan and Wei-Yin Loh. "LOTUS: An algorithm for building accurate and comprehensible logistic regression trees". In: Journal of Computational and Graphical Statistics 13.4 (2004), pp. 826–852.

[2] Niels Landwehr, Mark Hall, and Eibe Frank. "Logistic model trees". In: Machine learning 59.1-2 (2005), pp. 161–205.

[3] Achim Zeileis, Torsten Hothorn, and Kurt Hornik. "Model-based recursive partitioning". In: Journal of Computational and Graphical Statistics 17.2 (2008), pp. 492–514.

Bonus: high dimensional unstructured data

The banking industry is in a *Big Data* era: gotta catch'em all!

1. Navigation data on Sofinco's website.
2. Transactional data (credit cards).

How to use these "dynamic" features, stored on evolving Hadoop clusters, in traditional, parametric models such as logistic regression alongside "classical" features, stored in relational databases?

First few ideas: supervised generative clustering techniques, *e.g.* functional PCA.

Teaching

Teaching: past experience I

DUT STID 1st year - Université de Lille

Programmation statistique - R
Aim: give an overview of data ingestion, data types, standard libraries and graphs, notebooks and basic statistics.

GIS3 - Polytech' Lille

Statistique inférentielle - R
Aim: rephrase an industrial problem with toy data into statistical tests, restitution as reports.

Régression linéaire - R
Aim: perform linear regression (univariate then multivariate regression) and goodness-of-fit tests.

Teaching: past experience II

GIS4 - Polytech' Lille

Projet statistique - R
Aim: perform an end-to-end analysis of a toy dataset (classification or clustering), structure it into a report and a presentation.

Semaine d'Études Mathématiques-Entreprises

Aim: provide PhD students with industry experience; proposed subject: profit estimation.

G3 - Centrale Lille

Projet Impact - Python
Aim: perform a "real-life" project in partnership with a company; proposed subject: quantization + interactions.

Teaching: possible contribution

1. Courses in the Bachelor program

   $\approx$ similar level to what I've taught so far *e.g.* CSE 101, 102, 204, Computer Science projects, which could benefit from my "industrial" toolkit (R, Python, git, Docker) and potential projects with real datasets.

Teaching: possible contribution

1. Courses in the Bachelor program

   $\approx$ similar level to what I've taught so far *e.g.* CSE 101, 102, 204, Computer Science projects, which could benefit from my "industrial" toolkit (R, Python, git, Docker) and potential projects with real datasets.

2. Data Science track of the 3A / M1 Herbrand

   In particular INF554, MAP573, MAP566 which are close to the day-to-day activities of a Data Scientist: pipeline, preprocessing, supervised learning, visualization, ...

Thank you!

Publications

Adrien Ehrhardt et al. Credit Scoring : biais déchantillon ou réintégration des refusé. 2017. URL: https://adimajo.github.io/assets/publications/EHRHARDT_RJS_REINTEGRATION.pdf

Adrien Ehrhardt et al. "Réintégration des refusés en Credit Scoring". In: 49e Journées de Statistique. Avignon , France, May 2017. URL: https://hal.archives-ouvertes.fr/hal-01653767

Adrien Ehrhardt et al. "Reject Inference methods in Credit Scoring: a rational review". in preparation

Adrien Ehrhardt. scoring: Credit Scoring tools (version 0.1). 2018. URL: http://www.github.com/adimajo/scoring

Adrien Ehrhardt et al. Supervised multivariate discretization and levels merging for logistic regression. in preparation

Adrien Ehrhardt et al. Model-based multivariate discretization for logistic regression. Data Science Summer School. 2017. URL: http://2017.ds3-datascience-polytechnique.fr/wp-content/uploads/2017/08/DS3_posterID_049.pdf

Adrien Ehrhardt et al. "Supervised multivariate discretization and levels merging for logistic regression". In: 23rd International Conference on Computational Statistics. Iasi, Romania, Aug. 2018. URL: https://hal.archives-ouvertes.fr/hal-01949128

Adrien Ehrhardt et al. Supervised multivariate discretization and levels merging for logistic regression. Séminaire EA2496. 2018

Adrien Ehrhardt et al. "Feature quantization for parsimonious and interpretable predictive models". In: arXiv preprint arXiv:1903.08920 (2019)