Thèse CIFRE : modèles prédictifs pour données volumineuses et biaisées

Adrien Ehrhardt

04/04/2019

Table of Contents

Contexte du Credit Scoring

Contexte du Credit Scoring

Job	Home	Time in job	Family status	Wages		Repayment
Craftsman	Owner	20	Widower	2000		0
?	Renter	10	Common-law	1700		0
Licensed professional	Starter	5	Divorced	4000		1
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.

Contexte du Credit Scoring

Job	Home	Time in job	Family status	Wages		Repayment
Craftsman	Owner	20	Widower	2000		0
?	Renter	10	Common-law	1700		0
Licensed professional	Starter	5	Divorced	4000		1
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

Contexte du Credit Scoring

Job	Home	Time in job	Family status	Wages		Repayment
Craftsman	Owner	20	Widower	2000		0
?	Renter	10	Common-law	1700		0
Licensed professional	Starter	5	Divorced	4000		1
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

Contexte du Credit Scoring

Job			Family status	Wages		Repayment
Craftsman			Widower	2000		0
?			Common-law	1700		0
Licensed professional			Divorced	4000		1
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

Contexte du Credit Scoring

Job			Family status	Wages		Repayment
Craftsman			Widower	]1500;2000]		0
?			Common-law	]1500;2000]		0
Licensed professional			Divorced	]2000;∞[		1
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

Contexte du Credit Scoring

Job			Family status	Wages		Repayment
?+Low-qualified			?+Alone	]1500;2000]		0
?+Low-qualified			Union	]1500;2000]		0
High-qualified			?+Alone	]2000;∞[		1
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

Contexte du Credit Scoring

Job			Family status x Wages		Repayment
?+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

Contexte du Credit Scoring

Job			Family status x Wages		Repayment
?+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

Contexte du Credit Scoring

Job			Family status x Wages	Score	Repayment
?+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**

Réintégration des refusés

Réintégration des refusés

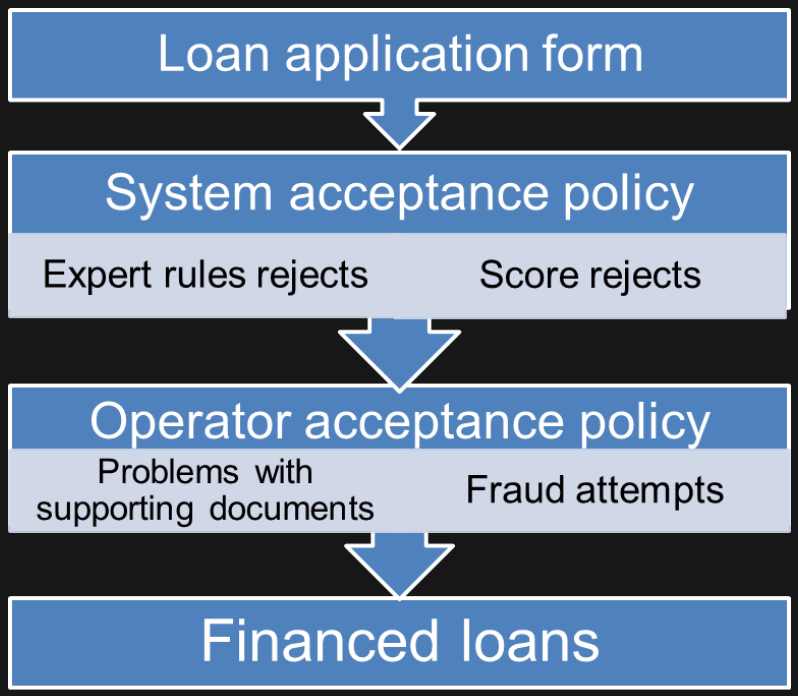


Figure: Simplified Acceptance mechanism in Crédit Agricole Consumer Finance

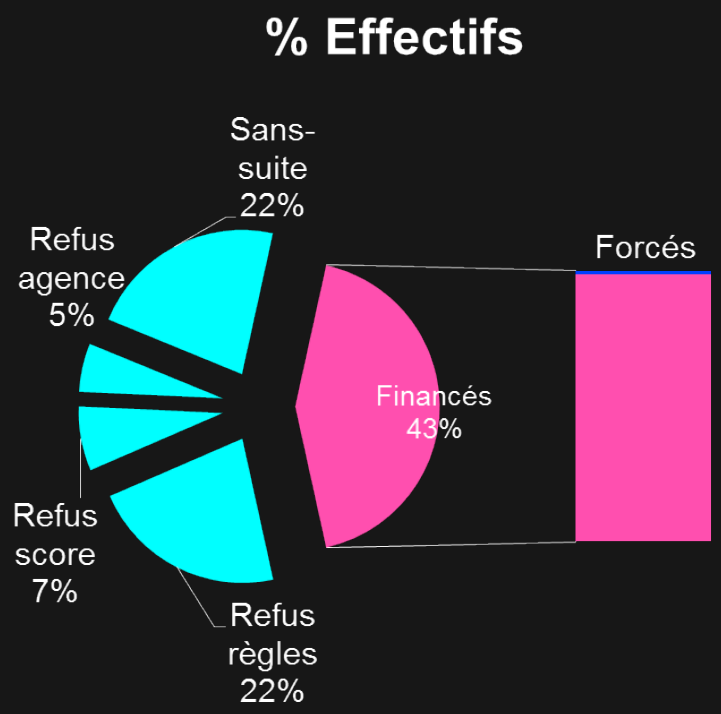
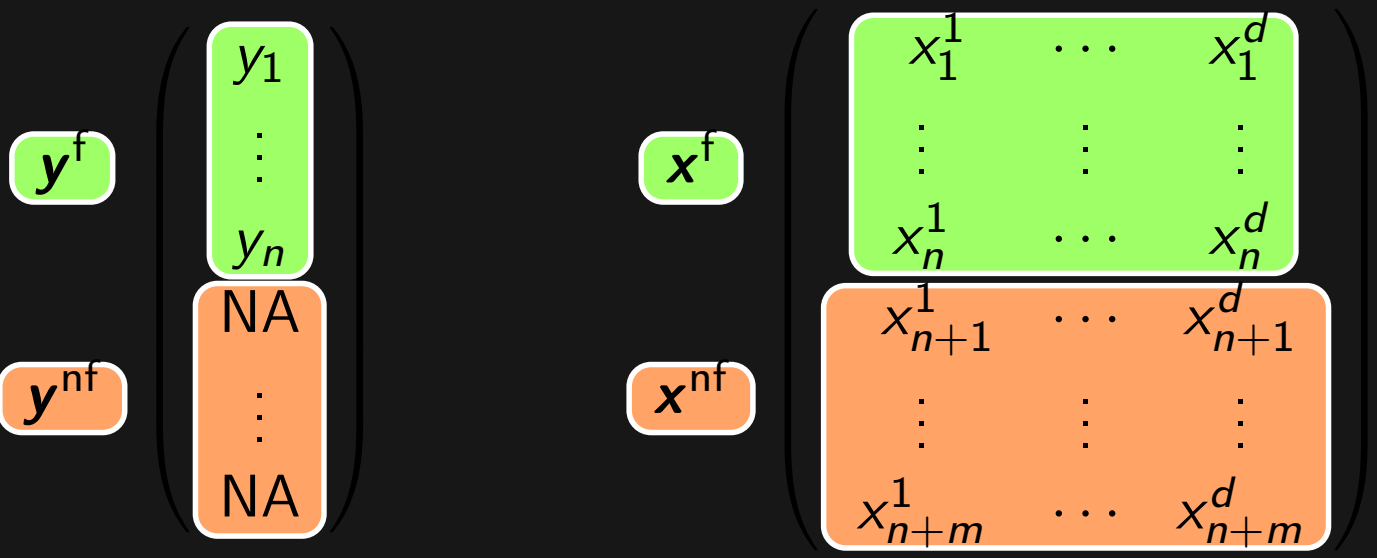


Figure: Proportion des différentes décisions finales

Réintégration des refusés

On a les données observées suivantes :



On construit la régression logistique sur clients financés :

$$\hat{\theta}^{\mathrm{f}} = \operatorname{argmax} \ell(\theta; \boldsymbol{x}^{\mathrm{f}}, \boldsymbol{y}^{\mathrm{f}}).$$

On voudrait :

$$\hat{\theta} = \operatorname{argmax} \ell(\theta; \boldsymbol{x}, \boldsymbol{y}).$$

On n'a pas $\boldsymbol{y}^{\mathrm{nf}}$.

Réintégration des refusés

Si on oublie un instant le cadre stricte de la LR, on a :

$$p_\gamma(y, z|x) = p_{\beta(\gamma)}(z|y, x)p_{\theta(\gamma)}(y|x).$$

Classiquement, on voudrait :

$$\ell(\gamma; S) = \sum_{i \in \mathrm{F}} \ln p_\gamma(y_i, \mathrm{f}|x_i) + \sum_{i' \in \mathrm{NF}} \ln \left[\sum_{y \in \{0,1\}} p_\gamma(y, \mathrm{nf}|x_{i'}) \right].$$

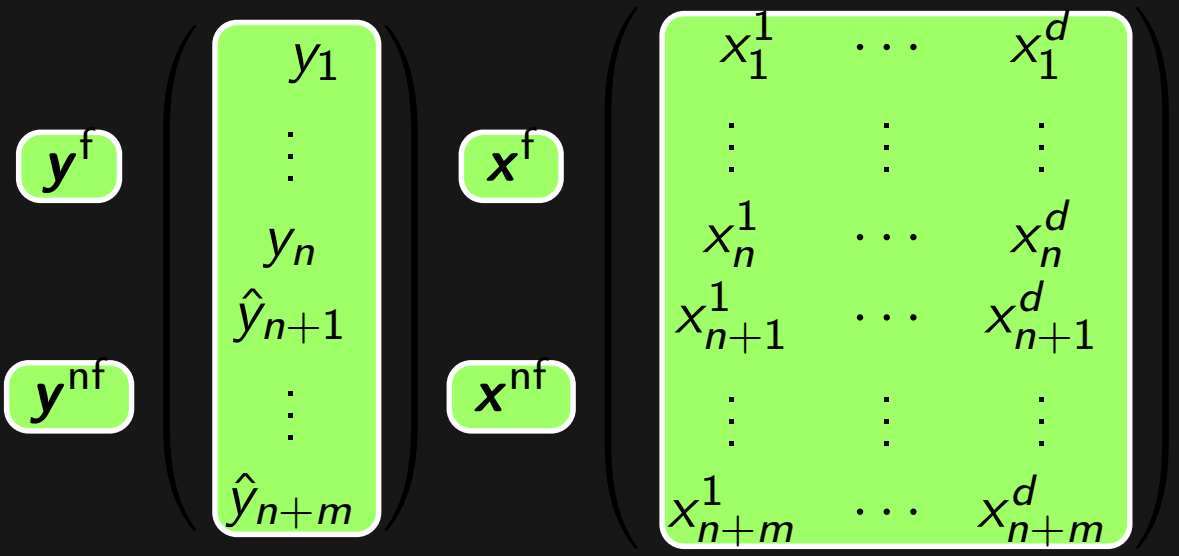
Il faut donc passer à un critère d'information t.q. :

$$\mathrm{AIC} = \ell(\hat{\gamma}; \boldsymbol{x}, \boldsymbol{y}^{\mathrm{nf}}) - |\Gamma|,$$

car un jeu de données test **sur l'ensemble de la population** n'est pas disponible.
On paye un prix supplémentaire incompatible avec la LR et plus restrictive = biais de modèle potentiellement plus important (se voit dans les expériences).

Réintégration des refusés

On se replonge dans la régression logistique ; des méthodes *ad hoc* ont été proposées qui peuvent se résumer à :



$\Rightarrow$ réinterpréter ces méthodes conformément à la slide précédente.

Long story short, les méthodes :

- ne servent à rien ($\hat{\theta}^{\text{méthode}} = \hat{\theta}^{\mathrm{f}}$) ;
- peuvent servir mais font des hypothèses supplémentaires :
  - estimation supplémentaire = biais + variance ;
  - ne peuvent pas être comparées à un "gold standard" car ~~$\boldsymbol{y}^{\mathrm{nf}}$~~ ;
- sont "pires" que le modèle sans réintégration.

Réintégration des refusés

Publications & livrables :

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

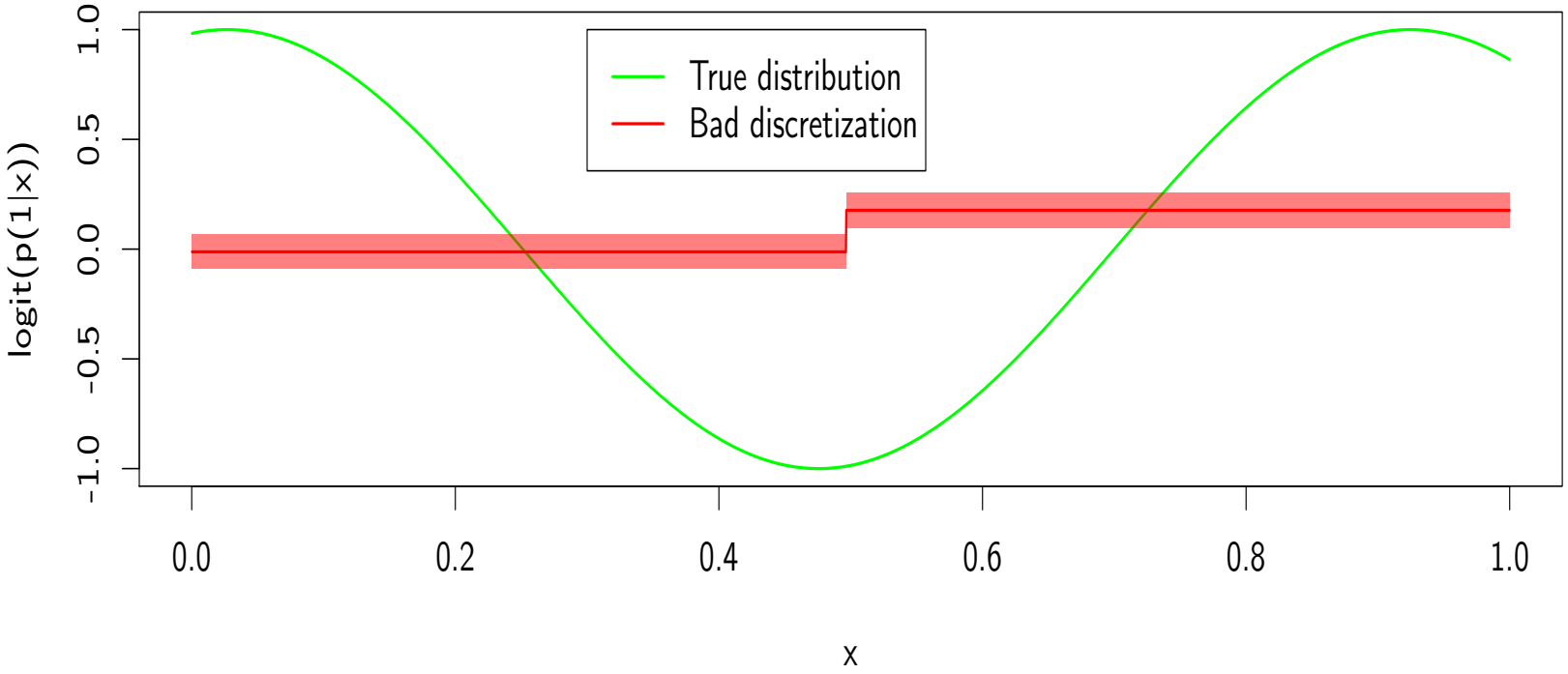
Adrien Ehrhardt et al. "Réintégration des refusés en Credit Scoring". In: 49e Journées de Statistique. Avignon , France, May 2017. URL: https://hal.archives-ouvertes.fr/hal-01653767

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

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

Quantification de variables

Quantification de variables: Un exemple



Quantification de variables: Un exemple

On discrétise par bande d'égale-longueur une variable continue x et la variable cible y dont le logit dépend du sinus de x. On observe :

- le compromis biais-variance contrôlé par le nombre de modalités
- l'importance du choix des endroits de coupure

Quelques notations I

Raw data

$$\boldsymbol{x} = (x_1, \ldots, x_d)$$
$$x_j \in \mathbb{R} \text{ (continuous case)}$$
$$x_j \in \{1, \ldots, l_j\} \text{ (categorical case)}$$
$$y \in \{0, 1\} \text{ (target)}$$

Quantized data

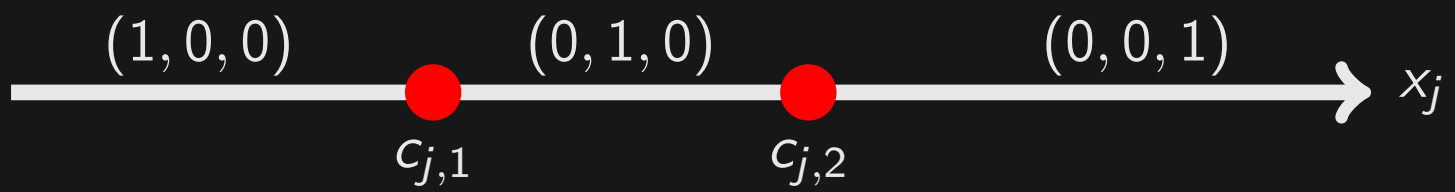
$$\boldsymbol{q}(\boldsymbol{x}) = (\boldsymbol{q}_1(x_1), \ldots, \boldsymbol{q}_d(x_d))$$
$$\boldsymbol{q}_j(x_j) = (q_{j,h}(x_j))_1^{m_j} \text{ (one-hot encoding)}$$
$$q_{j,h}(\cdot) = 1 \text{ if } x_j \in C_{j,h}, 0 \text{ otherwise, } 1 \leq h \leq m_j$$

Quelques notations II

Discretization

$$C_{j,h} = (c_{j,h-1}, c_{j,h}]$$

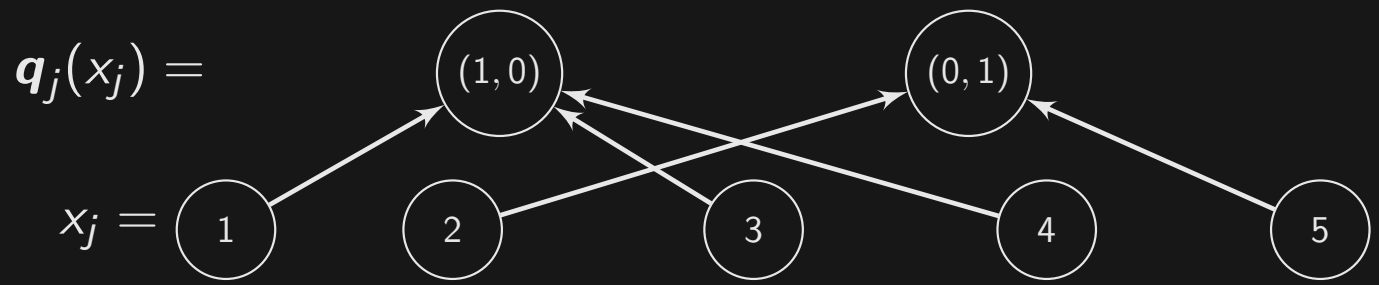
where $c_{j,1}, \ldots, c_{j,m_j-1}$ are increasing numbers called cutpoints, $c_{j,0} = -\infty$ and $c_{j,m_j} = +\infty$.



Quelques notations III

Grouping

$$\bigsqcup_{h=1}^{m_j} C_{j,h} = \{1, \ldots, l_j\}.$$



Quantification de variables: Approximation

$$\boldsymbol{q}_{\boldsymbol{\alpha}_j}(\cdot) = \left(q_{\boldsymbol{\alpha}_{j,h}}(\cdot)\right)_{h=1}^{m_j} \text{ with } \begin{cases} \sum_{h=1}^{m_j} q_{\boldsymbol{\alpha}_{j,h}}(\cdot) = 1, \\ 0 \leq q_{\boldsymbol{\alpha}_{j,h}}(\cdot) \leq 1, \end{cases}$$

For continuous features, we set for $\boldsymbol{\alpha}_{j,h} = (\alpha_{j,h}^0, \alpha_{j,h}^1) \in \mathbb{R}^2$

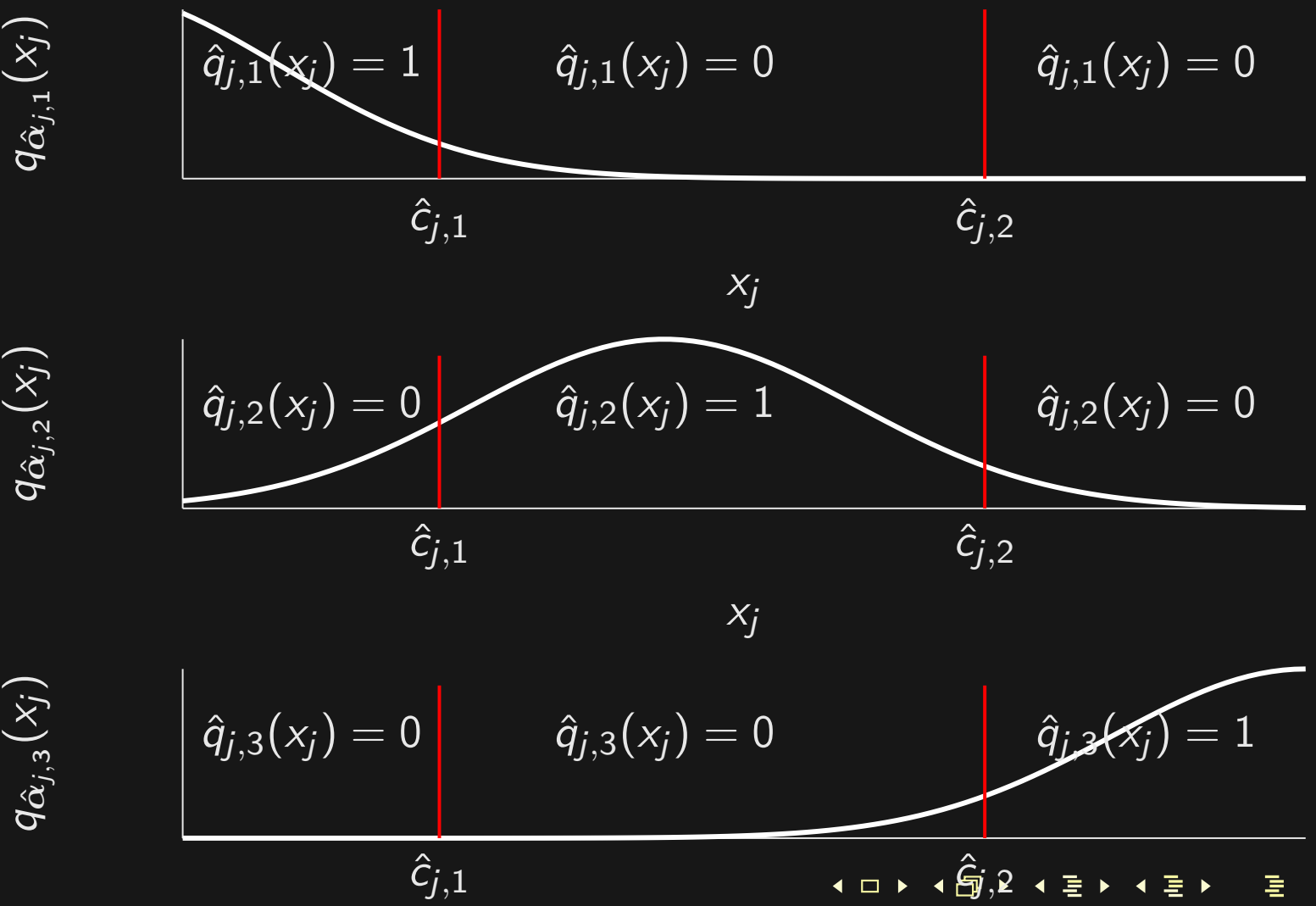
$$q_{\boldsymbol{\alpha}_{j,h}}(\cdot) = \frac{\exp(\alpha_{j,h}^0 + \alpha_{j,h}^1 \cdot)}{\sum_{g=1}^{m_j} \exp(\alpha_{j,g}^0 + \alpha_{j,g}^1 \cdot)}.$$

For categorical features, we set for $\boldsymbol{\alpha}_{j,h} = (\alpha_{j,h}(1), \ldots, \alpha_{j,h}(l_j)) \in \mathbb{R}^{l_j}$

$$q_{\boldsymbol{\alpha}_{j,h}}(\cdot) = \frac{\exp\left(\alpha_{j,h}(\cdot)\right)}{\sum_{g=1}^{m_j} \exp\left(\alpha_{j,g}(\cdot)\right)}.$$

Quantification de variables: Estimation MAP

$$q_{j,h}^{\mathrm{MAP}}(x_j) = 1 \text{ if } h = \underset{1 \le h' \le m_j}{\operatorname{argmax}} q_{\hat{\alpha}_{j,h'}}, 0 \text{ otherwise.}$$



Quantification de variables: SEM-Gibbs

$1^{\text{ère}}$ proposition faite pendant la thèse :

- On considère que $\boldsymbol{q}_\alpha$ est en fait une variable latente ;
- Algorithme EM classique en présence de variables latentes ;
- Espérance sur l'ensemble des quantifications ($\Rightarrow$ intractable) ;
- Solution : tirage aléatoire
- H1 : $p(y|\boldsymbol{x}, \boldsymbol{q}_\alpha) = p(y|\boldsymbol{q}_\alpha)$
- H2 : $p(\boldsymbol{q}_\alpha|\boldsymbol{x}) = \prod_{j=1}^{d} p(\boldsymbol{q}_{\alpha_j}|x_j)$
- H3 : $max_{\theta,\boldsymbol{q}_\alpha} \approx max_{\theta,\alpha} p(y|\boldsymbol{x}, \theta, \alpha)$

$p(\boldsymbol{q}_{\alpha_j}|\boldsymbol{x}, y, \boldsymbol{q}_{\alpha_{-\{j\}}}) \propto p_\theta(y|\boldsymbol{q}_\alpha) p(\boldsymbol{q}_{\alpha_j}|x_j)$

Quantification de variables: SEM-Gibbs

Publications & livrables :

Adrien Ehrhardt et al. “Supervised multivariate discretization and levels merging for logistic regression”. In: (). 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

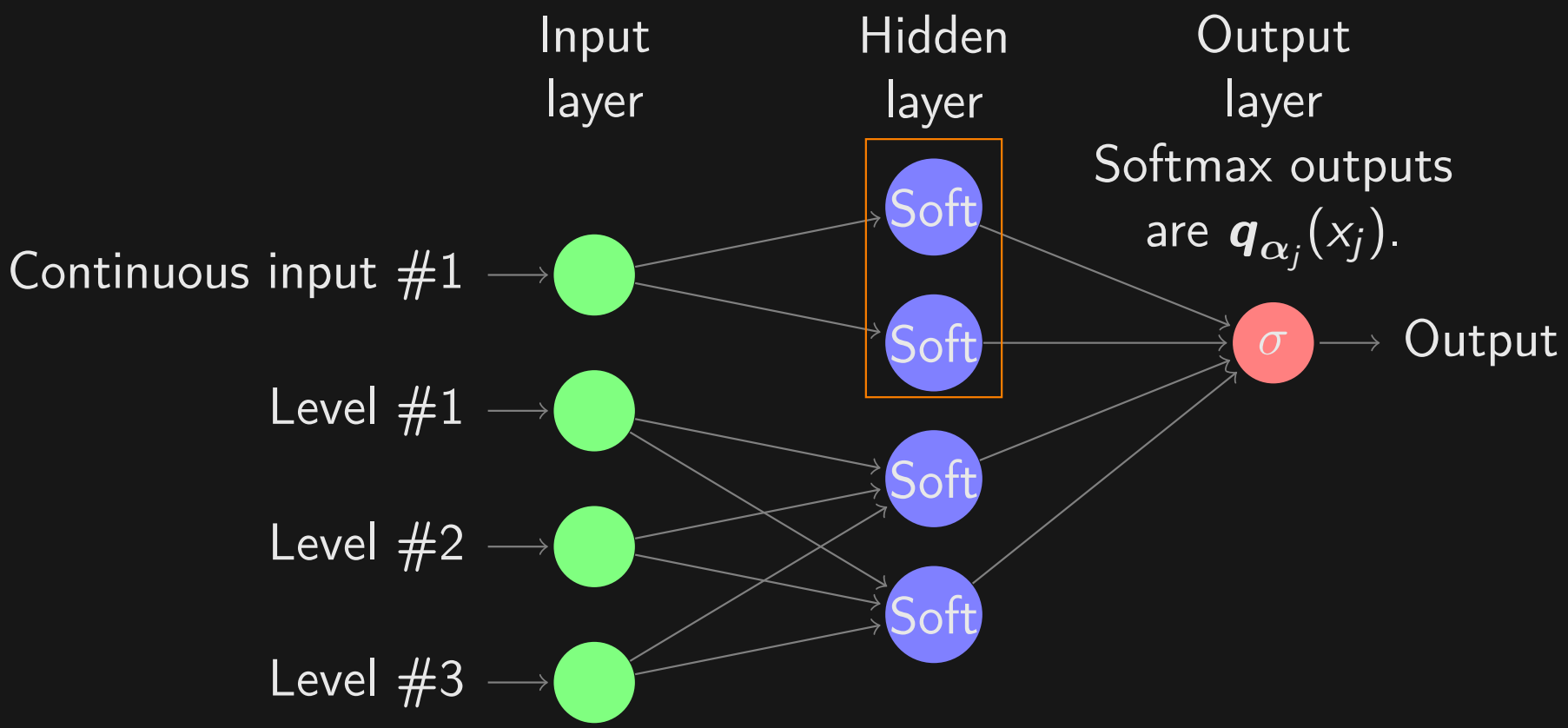
Quantification de variables: Réseaux de neurones I

Alternativement à la proposition SEM, on peut maximiser directement :

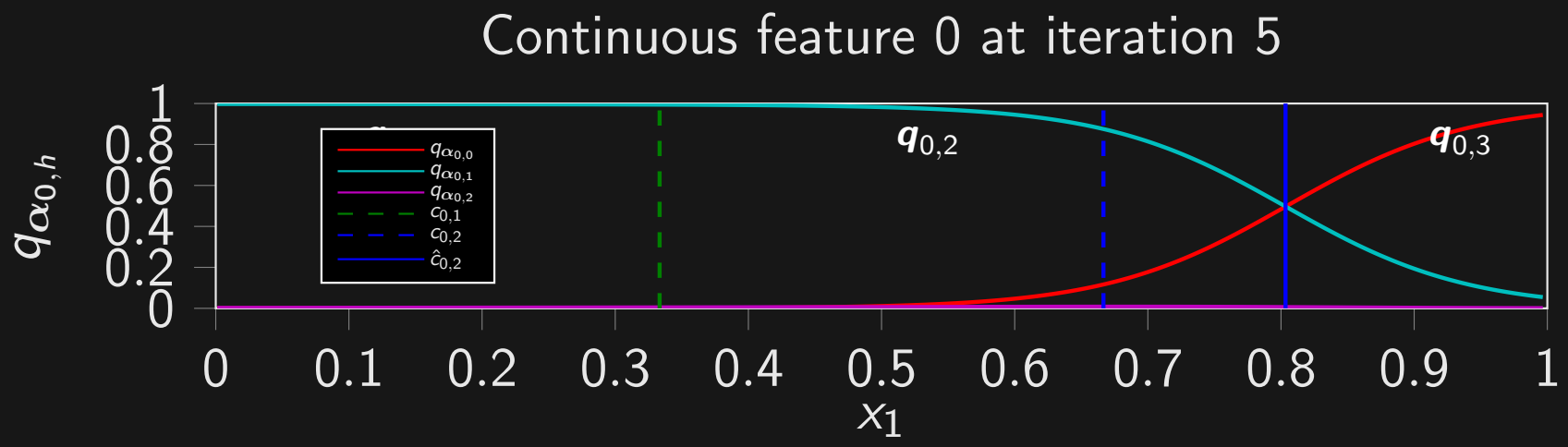
$$\ell(\theta, \alpha; \boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} p_\theta(y_i | \boldsymbol{q}_\alpha(x_i)),$$

par **descente de gradient**. Le problème n'est pas convexe : pas de garantie de convergence dans le cas général contrairement à la LR "classique".

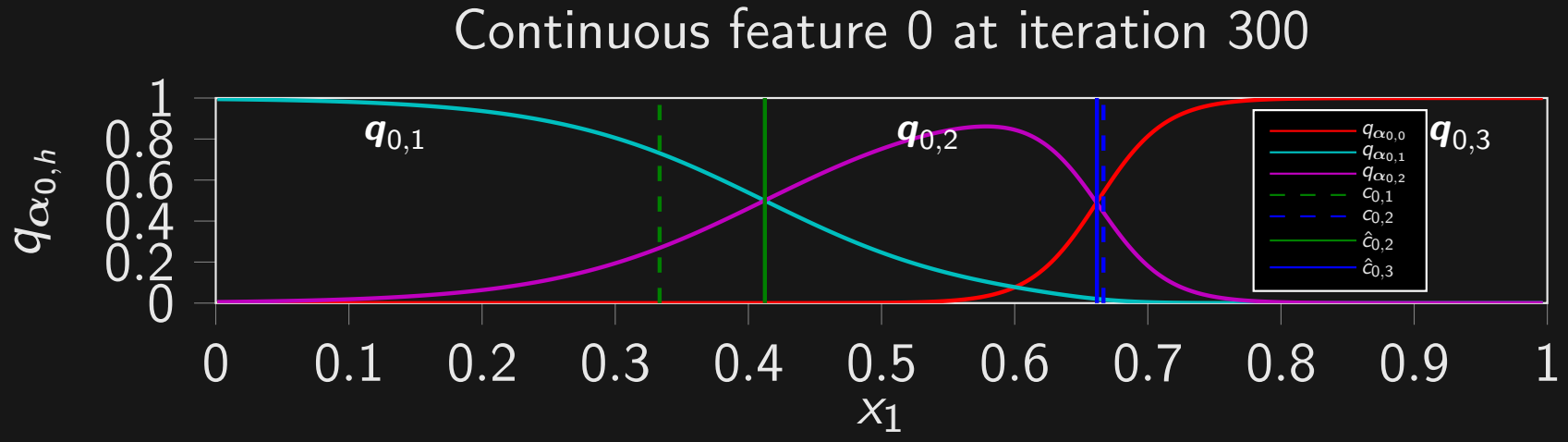
Quantification de variables: Réseaux de neurones II



Estimation *via* neural networks



(a) Quantization $\hat{\boldsymbol{q}}_1^{(s)}(x_1)$ resulting from the MAP at iter $t = 5$ and $m_{\max} = 3$.



(b) Quantizations $\hat{\boldsymbol{q}}_1^{(s)}(x_1)$ resulting from the MAP at iter $t = 300$ and $m_{\max} = 3$.

Quantification de variables: Réseaux de neurones

Publications & livrables :

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

Quantification de variables: Choisir la meilleure

New model selection criterion

We have drastically restricted the search space to clever candidates $\boldsymbol{q}^{\text{MAP}(1)}, \ldots, \boldsymbol{q}^{\text{MAP}(\text{iter})}$ resulting from the the gradient descent steps.

$$(\boldsymbol{q}^\star, \theta^\star) = \underset{\hat{\boldsymbol{q}} \in \{\boldsymbol{q}^{\text{MAP}(1)}, \ldots, \boldsymbol{q}^{\text{MAP}(\text{iter})}\}, \theta \in \Theta_{\boldsymbol{m}}}{\text{argmin}} \text{BIC}(\hat{\theta}_{\hat{\boldsymbol{q}}})$$

Quantification de variables: Choisir la meilleure

New model selection criterion

We have drastically restricted the search space to clever candidates $\boldsymbol{q}^{\text{MAP}(1)}, \ldots, \boldsymbol{q}^{\text{MAP}(\text{iter})}$ resulting from the the gradient descent steps.

$$(\boldsymbol{q}^\star, \theta^\star) = \underset{\hat{\boldsymbol{q}} \in \{\boldsymbol{q}^{\text{MAP}(1)}, \ldots, \boldsymbol{q}^{\text{MAP}(\text{iter})}\}, \theta \in \Theta_{\boldsymbol{m}}}{\text{argmin}} \text{BIC}(\hat{\theta}_{\hat{\boldsymbol{q}}})$$

We would still need to loop over candidates $\boldsymbol{m}$!

Quantification de variables: Choisir la meilleure

New model selection criterion

We have drastically restricted the search space to clever candidates $\boldsymbol{q}^{\text{MAP}(1)}, \ldots, \boldsymbol{q}^{\text{MAP}(\text{iter})}$ resulting from the the gradient descent steps.

$$(\boldsymbol{q}^\star, \theta^\star) = \underset{\hat{\boldsymbol{q}} \in \{\boldsymbol{q}^{\text{MAP}(1)}, \ldots, \boldsymbol{q}^{\text{MAP}(\text{iter})}\}, \theta \in \Theta_{\boldsymbol{m}}}{\text{argmin}} \text{BIC}(\hat{\theta}_{\hat{\boldsymbol{q}}})$$

We would still need to loop over candidates $\boldsymbol{m}$!

In practice if $\forall i,\ q_{\alpha_{j,h}}(x_j) \ll 1$, then level h disappears while performing the argmax.

Quantification de variables: Choisir la meilleure

New model selection criterion

We have drastically restricted the search space to clever candidates $\boldsymbol{q}^{\mathrm{MAP}(1)}, \ldots, \boldsymbol{q}^{\mathrm{MAP}(\mathrm{iter})}$ resulting from the the gradient descent steps.

$$(\boldsymbol{q}^\star, \theta^\star) = \underset{\hat{\boldsymbol{q}} \in \{\boldsymbol{q}^{\mathrm{MAP}(1)}, \ldots, \boldsymbol{q}^{\mathrm{MAP}(\mathrm{iter})}\}, \theta \in \Theta_{\boldsymbol{m}}}{\operatorname{argmin}} \mathrm{BIC}(\hat{\theta}_{\hat{\boldsymbol{q}}})$$

We would still need to loop over candidates $\boldsymbol{m}$!

In practice if $\forall i,\ q_{\alpha_{j,h}}(x_j) \ll 1$, then level h disappears while performing the argmax.

Start with $\boldsymbol{m} = (m_{\max})_1^d$ and "wait" ...

Quantification de variables: Résultats

Données simulées

Table: For different sample sizes n, (A) CI of $\hat{c}_{j,2}$ for $c_{j,2} = 2/3$. (B) CI of $\hat{m}$ for $m_1 = 3$. (C) CI of $\hat{m}_3$ for $m_3 = 1$.

n	(A) $\hat{c}_{j,2}$	(B)	$\hat{m}_1$	(C)	$\hat{m}_3$
		1		60	
1,000	$[0.656, 0.666]$	90		32	
		9		8	
		0		88	
10,000	$[0.666, 0.666]$	100		12	
		0		0	

Quantification de variables: Résultats

Données UCI

Table: Gini indices (the greater the value, the better the performance) of our proposed quantization algorithm *glmdisc* and two baselines: ALLR and MDLP / χ^2 tests obtained on several benchmark datasets from the UCI library.

Dataset	ALLR	MDLP/χ^2	*glmdisc*
Adult	81.4 (1.0)	85.3 (0.9)	80.4 (1.0)
Australian	72.1 (10.4)	84.1 (7.5)	92.5 (4.5)
Bands	48.3 (17.8)	47.3 (17.6)	58.5 (12.0)
Credit	81.3 (9.6)	88.7 (6.4)	92.0 (4.7)
German	52.0 (11.3)	54.6 (11.2)	69.2 (9.1)
Heart	80.3 (12.1)	78.7 (13.1)	86.3 (10.6)

Quantification de variables: Résultats

Données CACF

Table: Gini indices (the greater the value, the better the performance) of our proposed quantization algorithm *glmdisc*, the two baselines of Table 3 and the current scorecard (manual / expert representation) obtained on several portfolios of Crédit Agricole Consumer Finance.

Portfolio	ALLR	Current	MDLP/χ^2	*glmdisc*
Automobile	59.3 (3.1)	55.6 (3.4)	59.3 (3.0)	58.9 (2.6)
Renovation	52.3 (5.5)	50.9 (5.6)	54.0 (5.1)	56.7 (4.8)
Standard	39.7 (3.3)	37.1 (3.8)	45.3 (3.1)	44.0 (3.1)
Revolving	62.7 (2.8)	58.5 (3.2)	63.2 (2.8)	62.3 (2.8)
Mass retail	52.8 (5.3)	48.7 (6.0)	61.4 (4.7)	61.8 (4.6)
Electronics	52.9 (11.9)	55.8 (10.8)	56.3 (10.2)	72.6 (7.4)

Interactions bivariées

Interactions bivariées: Notations

Upper triangular matrix with $\delta_{k,\ell} = 1$ if $k < \ell$ and features p and q "interact" in the logistic regression.

$$\text{logit}(p_{\boldsymbol{\theta}_f}(1|\boldsymbol{q}(\boldsymbol{x}))) = \theta_0 + \sum_{j=1}^{d} \theta_j^{\boldsymbol{q}_j(x_j)} + \sum_{1 \le k < \ell \le d} \delta_{k,\ell} \theta_{k,\ell}^{\boldsymbol{q}_k(x_k) f_\ell(x_\ell)}$$

Interactions bivariées: Notations

Upper triangular matrix with $\delta_{k,\ell} = 1$ if $k < \ell$ and features p and q "interact" in the logistic regression.

$$\text{logit}(p_{\theta_f}(1|\boldsymbol{q}(\boldsymbol{x}))) = \theta_0 + \sum_{j=1}^{d} \theta_j^{\boldsymbol{q}_j(x_j)} + \sum_{1 \le k < \ell \le d} \delta_{k,\ell} \theta_{k,\ell}^{\boldsymbol{q}_k(x_k) f_\ell(x_\ell)}$$

Imagine for now that the discretization $\boldsymbol{q}(\boldsymbol{x})$ is fixed. The criterion becomes:

$$(\theta^\star, \delta^\star) = \underset{\theta, \boldsymbol{\delta} \in \{0,1\}^{\frac{d(d-1)}{2}}}{\text{argmax}} \sum_{i=1}^{n} \ln p_\theta(y_i|\boldsymbol{q}(\boldsymbol{x}_i), \delta) - \text{penalty}(n; \theta)$$

Interactions bivariées: Notations

Upper triangular matrix with $\delta_{k,\ell} = 1$ if $k < \ell$ and features p and q "interact" in the logistic regression.

$$\text{logit}(p_{\theta_f}(1|\boldsymbol{q}(\boldsymbol{x}))) = \theta_0 + \sum_{j=1}^{d} \theta_j^{\boldsymbol{q}_j(x_j)} + \sum_{1 \leq k < \ell \leq d} \delta_{k,\ell} \theta_{k,\ell}^{\boldsymbol{q}_k(x_k) f_\ell(x_\ell)}$$

Imagine for now that the discretization $\boldsymbol{q}(\boldsymbol{x})$ is fixed. The criterion becomes:

$$(\theta^\star, \delta^\star) = \underset{\theta, \boldsymbol{\delta} \in \{0,1\}^{\frac{d(d-1)}{2}}}{\text{argmax}} \sum_{i=1}^{n} \ln p_\theta(y_i|\boldsymbol{q}(\boldsymbol{x}_i), \delta) - \text{penalty}(n; \theta)$$

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

Interactions bivariées: Model proposal

δ is latent and hard to optimize over: use a stochastic algorithm!

Interactions bivariées: Model proposal

δ is latent and hard to optimize over: use a stochastic algorithm!

Strategy used here: Metropolis-Hastings algorithm.

Interactions bivariées: Model proposal

δ is latent and hard to optimize over: use a stochastic algorithm!

Strategy used here: Metropolis-Hastings algorithm.

Which transition proposal $q : (\{0,1\}^{\frac{d(d-1)}{2}}, \{0,1\}^{\frac{d(d-1)}{2}}) \mapsto [0;1]$?

$2^{d(d-1)}$ probabilities to calculate...

Interactions bivariées: Model proposal

δ is latent and hard to optimize over: use a stochastic algorithm!

Strategy used here: Metropolis-Hastings algorithm.

Which transition proposal $q : (\{0,1\}^{\frac{d(d-1)}{2}}, \{0,1\}^{\frac{d(d-1)}{2}}) \mapsto [0;1]$?

$2^{d(d-1)}$ probabilities to calculate...

Proposal: gain/loss in BIC between bivariate models with / without the interaction.

Interactions bivariées: Model proposal

δ is latent and hard to optimize over: use a stochastic algorithm!

Strategy used here: Metropolis-Hastings algorithm.
Which transition proposal $q : (\{0,1\}^{\frac{d(d-1)}{2}}, \{0,1\}^{\frac{d(d-1)}{2}}) \mapsto [0;1]$?
$2^{d(d-1)}$ probabilities to calculate...
Proposal: gain/loss in BIC between bivariate models with / without the interaction.

Trick: alternate one discretization / grouping step and one "interaction" step.

Interactions bivariées: Résultats

Données UCI

Table: Gini indices (the greater the value, the better the performance) of our proposed quantization algorithm *glmdisc* and two baselines: ALLR and MDLP / χ^2 tests obtained on several benchmark datasets from the UCI library.

Dataset	ALLR	*ad hoc* methods	Our proposal: *glmdisc*-NN	Our proposal: *glmdisc*-SEM	*glmdisc*-SEM w. interactions
Adult	81.4 (1.0)	**85.3** (0.9)	80.4 (1.0)	81.5 (1.0)	81.5 (1.0 - no interaction)
Australian	72.1 (10.4)	84.1 (7.5)	92.5 (4.5)	**100** (0)	**100** (0 - no interaction)
Bands	48.3 (17.8)	47.3 (17.6)	58.5 (12.0)	**58.7** (12.0)	**58.8** (13.0)
Credit	81.3 (9.6)	88.7 (6.4)	**92.0** (4.7)	87.7 (6.4)	87.7 (6.4 - no interaction)
German	52.0 (11.3)	54.6 (11.2)	**69.2** (9.1)	54.5 (10)	
Heart	80.3 (12.1)	78.7 (13.1)	**86.3** (10.6)	82.2 (11.2)	84.5 (10.8)

Interactions bivariées: Résultats

Données benchmark médecine

Table: Gini indices of our proposed quantization algorithm *glmdisc*-SEM and two baselines: ALLR and ALLR with all pairwise interactions on several medicine-related benchmark datasets.

	Pima	Breast	Birthwt
ALLR	**73.0**	94.0	34.0
ALLR LR w. interactions	60.0	51.0	15.0
glmdisc	57.0	93.0	18.0
glmdisc w. interactions	62.0	**95.0**	**54.0**

Interactions bivariées: Résultats

Données CACF

Table: Gini indices (the greater the value, the better the performance) of our proposed quantization algorithm *glmdisc*, the two baselines of Table 3 and the current scorecard (manual / expert representation) obtained on several portfolios of Crédit Agricole Consumer Finance.

Portfolio	ALLR	Current performance	*ad hoc* methods	Our proposal: *glmdisc*-NN	Our proposal: *glmdisc*-SEM	*glmdisc*-SEM w. interactions
Automobile	59.3 (3.1)	55.6 (3.4)	59.3 (3.0)	58.9 (2.6)	57.8 (2.9)	**64.8** (2.0)
Renovation	52.3 (5.5)	50.9 (5.6)	54.0 (5.1)	**56.7** (4.8)	55.5 (5.2)	55.5 (5.2)
Standard	39.7 (3.3)	37.1 (3.8)	45.3 (3.1)	43.8 (3.2)	36.7 (3.7)	**47.2** (2.8)
Revolving	62.7 (2.8)	58.5 (3.2)	63.2 (2.8)	62.3 (2.8)	60.7 (2.8)	**67.2** (2.5)
Mass retail	52.8 (5.3)	48.7 (6.0)	61.4 (4.7)	**61.8** (4.6)	61.0 (4.7)	60.3 (4.8)
Electronics	52.9 (11.9)	55.8 (10.8)	56.3 (10.2)	**72.6** (7.4)	62.0 (9.5)	63.7 (9.0)

Segmentation : arbres de régressions logistiques

Segmentation : arbres de régressions logistiques

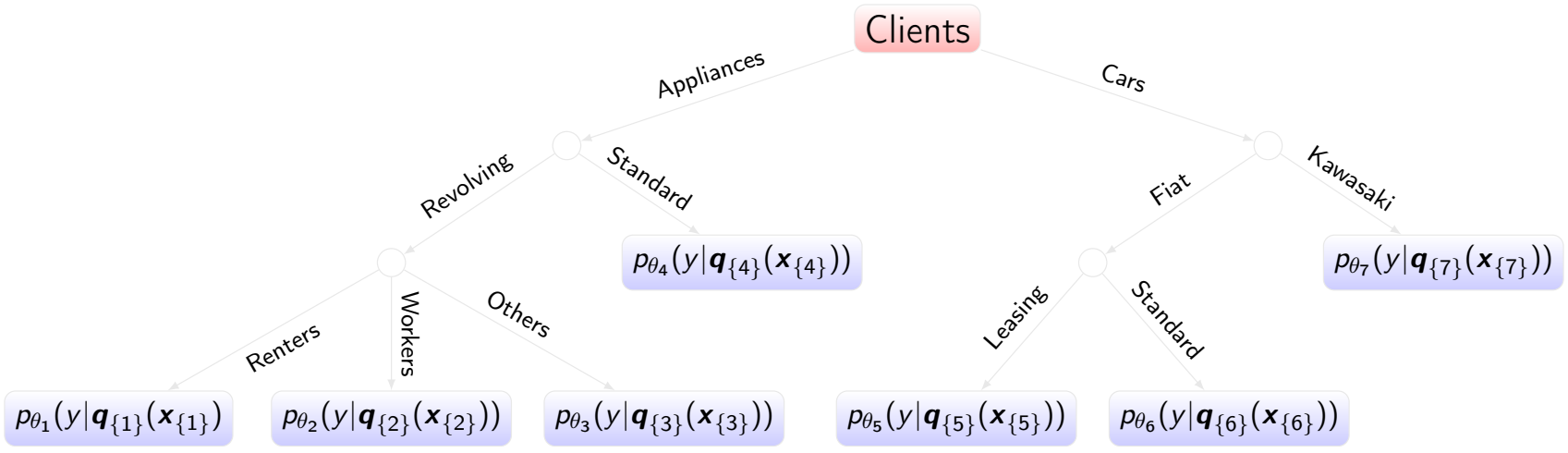


Figure: Scorecards tree structure in acceptance system.

Segmentation : arbres de régressions logistiques: Quelques résultats

	Oracle = ALLR	*glmtree*-SEM	FAMD	PLS	LMT	MOB
Gini	69.7	**69.7**	65.3	47.0	**69.7**	64.8

	Oracle	ALLR	*glmtree*-SEM	FAMD	PLS	LMT	MOB
Gini	69.7	25.8	**69.7**	17.7	48.4	65.8	**69.7**

Merci de votre attention !

References I


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


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


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


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.

References II


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: (). In preparation.


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 regress Séminaire EA2496. 2018.


Adrien Ehrhardt et al. “Supervised multivariate discretization and levels merging for logistic regression”. In: (). In preparation.