Proposition de thèse 9 mars 2005

Essays on Classification

Alejandro Karam
HEC-Montréal

HEC MONTRÉAL

Structure de la thèse proposée

- Introduction
- Arbitrary-norm Distance Minimization by VNS
- Misclassification Minimization by VNS
- Application to Credit Scoring

Plan de la présentation

- Motivation et exposition du problème
- Séparation en norme $L_{p:}$
  - Survol d'approches exactes de solution
  - Méthode heuristique proposée
- Expériences numériques
- Credit Scoring:
  - État du travail
  - Défis
  - Tâches

Credit Scoring

- Données sociodémographiques
- Antécédents:
  - Internes
  - Externes
- Autres

- Indice (score)
- Seuil

- Accorder ou refuser le prêt

Classification supervisée

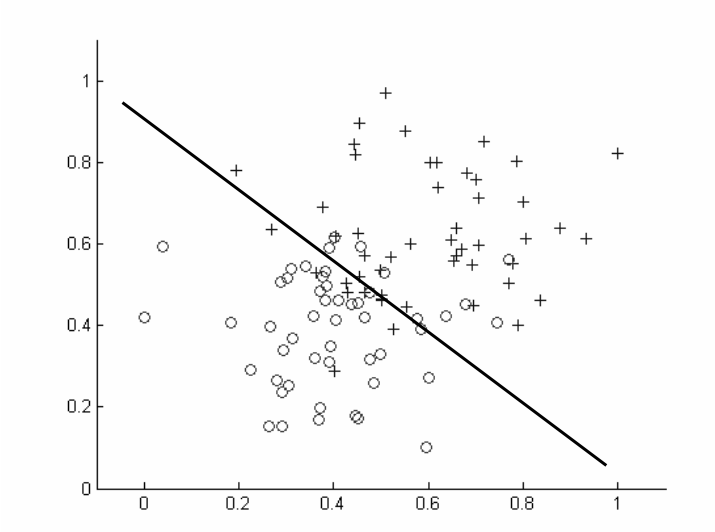
caractéristiques → **modèle** → **classe**

- Données sociodémographiques
- Antécédents:
  - Internes
  - Externes
- Autres

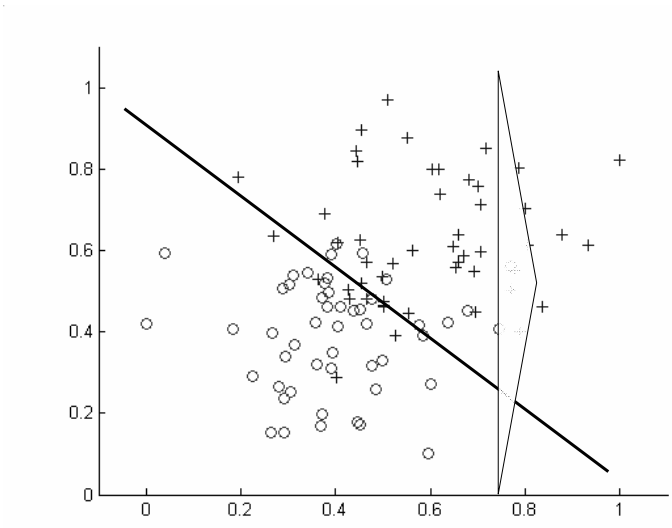
- Indice (score)
- Seuil

- Accorder ou refuser le prêt

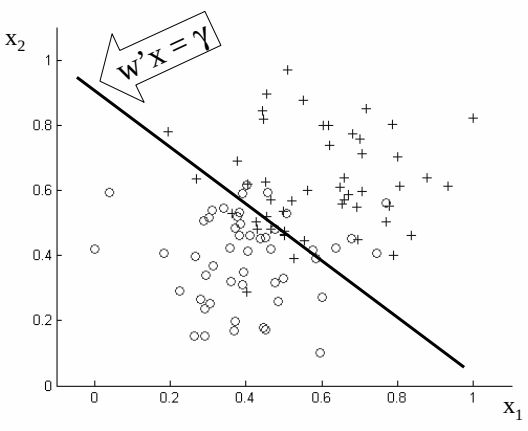
Separation par hyperplans



Separation par hyperplans



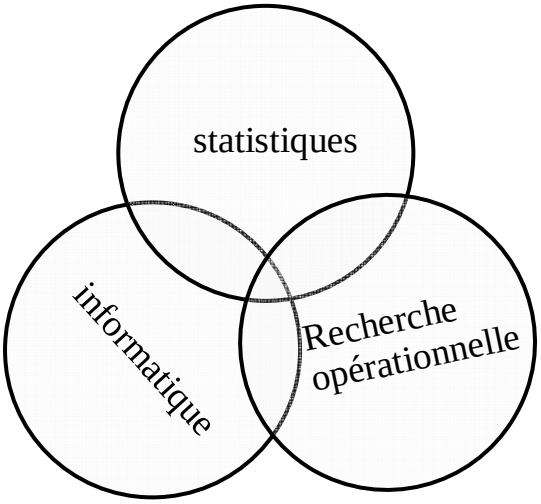
Separation par hyperplans



<u>Discriminant linéaire</u>:

- $w'x > \gamma$ si $x \in \mathcal{A}(+)$
- $w'x < \gamma$ si $x \in \mathcal{B}(o)$

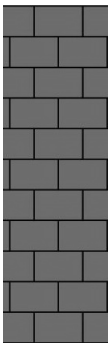
Approches au problème



Un peu d'histoire

Mangasarian:

- 1965 *(Op. Res.)*
- 1968, 1973 *?*
- 1992 *RLP* (avec Bennett)
- 1995 *Breast Cancer*
- 1999 *Arbitrary Norm Separation*



Freed & Glover:

- 1981 *(Dec. Sci.)*
- encore 1981b *(EJOR)*, 1981c (*Dec. Sci.*)
- 1982
- 1986a et 1986b

Glover:

- seul 1988; et al 1988
- 1990

et plein d'autres!

Glen, Koehler, Stam, Erenguc, Joachimsthaler, Ragsdale, Cavalier, etc.

Distance d'un point au plan en norme L_p

The distance from a point $z \in \mathbb{R}^n$ to a plane $P = \{x \,|\, w'x = \gamma\}$ in norm L_p is

$$\frac{|w'x - \gamma|}{\|w\|'_p} \tag{1}$$

where $\|w\|'_p$ is the dual norm of L_p.

For $1 < p < \infty$, $\|w\|'_p = \|w\|_q$ where q is such that $\frac{1}{p} + \frac{1}{q} = 1$.

We therefore compute it as

$$\|w\|'_p = \sqrt[\frac{p-1}{p}]{\sum w^{\frac{p}{p-1}}} \tag{2}$$

Séparation en norme L_p

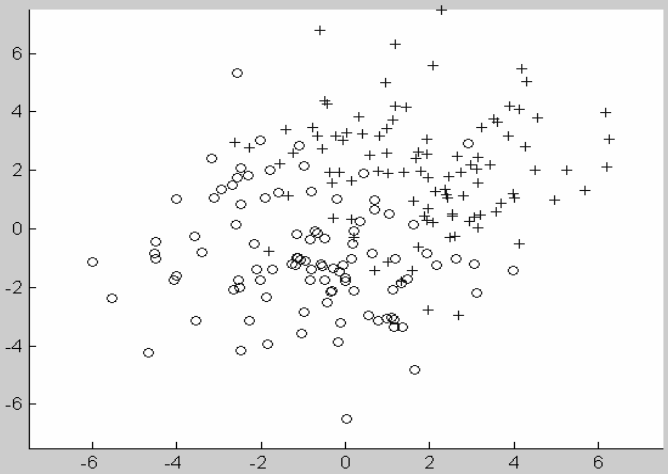
- Le problème est

$$\min_{(w \in \mathbb{R}^n, \gamma \in \mathbb{R})} \left\{ \frac{\sum_{i=1}^{m} \max\{-w'A_i + \gamma, 0\} + \sum_{j=1}^{k} \max\{w'B_j - \gamma, 0\}}{\|w\|'_p} \right\}$$

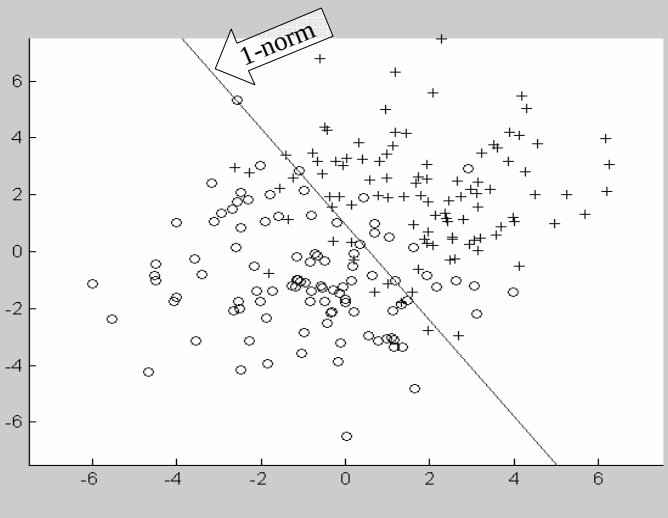
- où $\|w\|'_p = \sqrt[\frac{p-1}{p}]{\sum w^{\frac{p}{p-1}}}$

PAS BEAU!

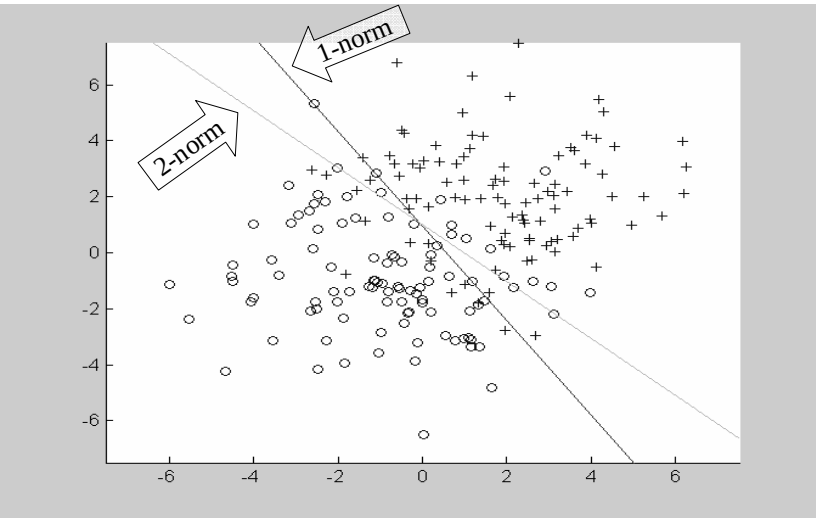
Séparation en norme L_p



Séparation en norme L_p



Séparation en norme L_p



Séparation en norme L_p

- Le problème est

$$\min_{(w\in\mathbb{R}^n,\gamma\in\mathbb{R})} \left\{ \frac{\sum_{i=1}^{m} \max\{-w'A_i+\gamma,0\}+\sum_{j=1}^{k} \max\{w'B_j-\gamma,0\}}{\|w\|'_p} \right\}$$

- où $\|w\|'_p = \sqrt[\frac{p-1}{p}]{\sum w^{\frac{p}{p-1}}} = 1$

- Imposition de la contrainte sur le dénominateur
- Linéarisation de l'opérateur *max* {•}

Séparation en norme L_p

- On a donc finalement:

$$\min_{w,\gamma,y,z} \left\{ \sum_{i=1}^{m} y_i + \sum_{j=1}^{k} z_j \;\middle|\; \begin{array}{c} y_i \geq -w'A_i + \gamma \text{ for } i = 1,..,m \\ z_j \geq w'B_j - \gamma \text{ for } j = 1,..,k \\ \|w\|_p' = 1 \end{array} \right\}$$

where $w \in \mathbb{R}^n$, $\gamma \in \mathbb{R}$, $y \in \mathbb{R}^m_+$ and $z \in \mathbb{R}^k_+$.

Stratégies de solution

EXACTE

- méthodes connues pour p = 1 , 2 et ∞
- raisonnable pour instances de taille très modérée en L_2 et L_∞
- pas d'algorithme général pour p arbitraire

HEURISTIQUE

- n'importe quelle p > 1, entier ou fractionnaire
- solutions acceptables et rapides pour instances plus grandes.
- permet d'adresser le cas p = 0

Stratégies de solution

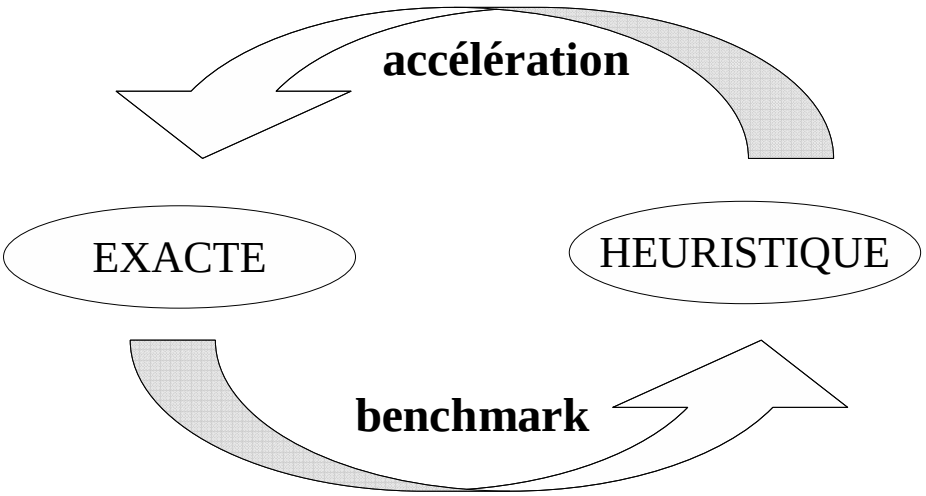
EXACTE

- méthodes connues pour p = 1 , 2 et ∞
- raisonnable pour instances de taille très modérée en L_2 et L_∞
- pas d'algorithme général pour p arbitraire

HEURISTIQUE

- n'importe quelle p > 1, entier ou fractionnaire
- solutions acceptables et rapides pour instances plus grandes
- permet d'adresser le cas p = 0

Stratégies de solution



Plan de la présentation

- Motivation et exposition du problème
- Séparation en norme $L_{p:}$
  - Survol d'approches exactes de solution
  - Méthode heuristique proposée
  - Expériences numériques
- Credit Scoring:
  - État du travail
  - Défis
  - Tâches

Approches exactes: L_1

- Pour p=1

$$\|w\|_1' = \|w\|_\infty = \max_{i=1,\ldots,n} |w_i|$$

- Au moins un w_i doit être exactement 1 ou -1
- On peut donc essayer tous les cas

- 2 *n* programmes linéaires
- Mangasarian (1965) le propose sans savoir ce que c'est
- Astuce pour accélérer

Approches exactes: L_2

- Pour p=2

$$\|w\|_2' = w'w = 1.$$

• Contrainte quadratique non convexe
• Algorithme de Charles Audet, adapté par Sylvain Perron

Approches exactes: L_∞

- Pour p = ∞

$$\|w\|_\infty' = \|w\|_1 = \sum_{l=1}^{n} |w_l| = 1.$$

• Formulation en programmation entière avec *n* variables binaires
• Proposé par Glenn en 1999 sans s'apercevoir que c'était la norme L_∞

Plan de la présentation

- Motivation et exposition du problème
- Séparation en norme $L_{p:}$
  - Survol d'approches exactes de solution
  - Méthode heuristique proposée
  - Expériences numériques
- Credit Scoring:
  - État du travail
  - Défis
  - Tâches

Approche heuristique

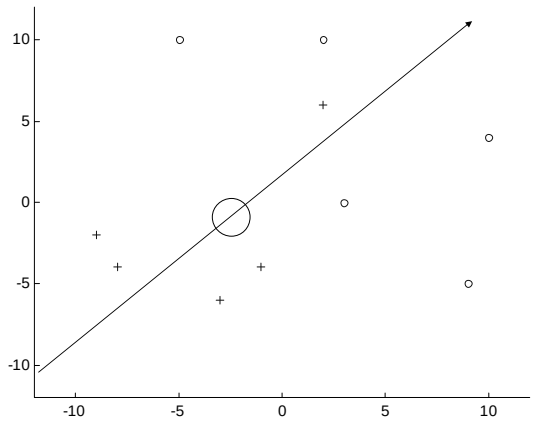
- Découpler la recherche du gradient *w* de celle de l'ordonnée à l'origine γ
- Noter que pour *w* fixe, la conversion par rapport à la norme Euclidienne est un changement d'échelle constante

$$P = \{x \,|w'x = \gamma\}$$

$$\frac{|w'x - \gamma|}{\|w\|'_p}$$

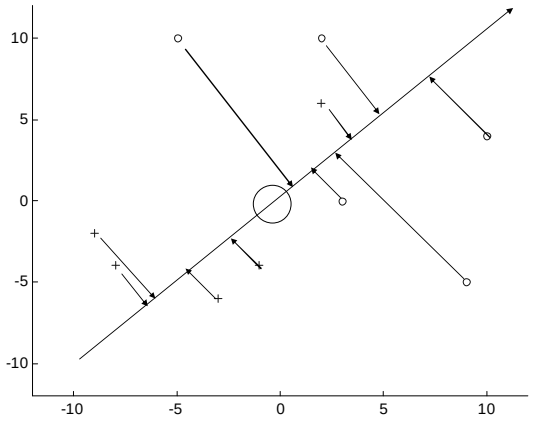
Approche heuristique

La décomposition:
- fixer une direction
- projeter les points sur le rayon
- trouver meilleure position pour plan



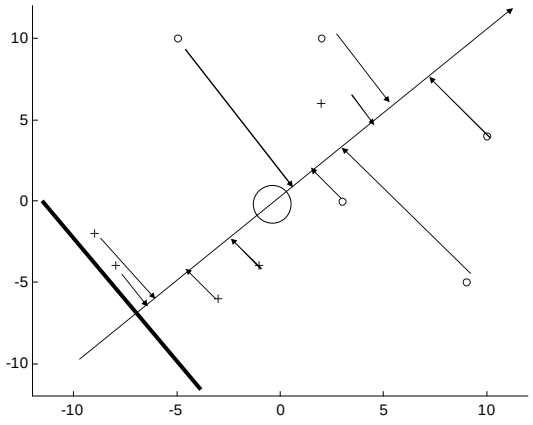
Approche heuristique

La décomposition:
- fixer une direction
- projeter les points sur le rayon
- trouver meilleure position pour plan



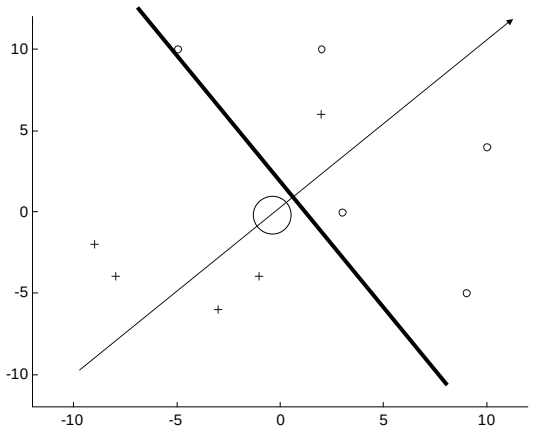
Approche heuristique

La décomposition:
• fixer une direction
• projeter les points sur le rayon
• trouver meilleure position pour plan

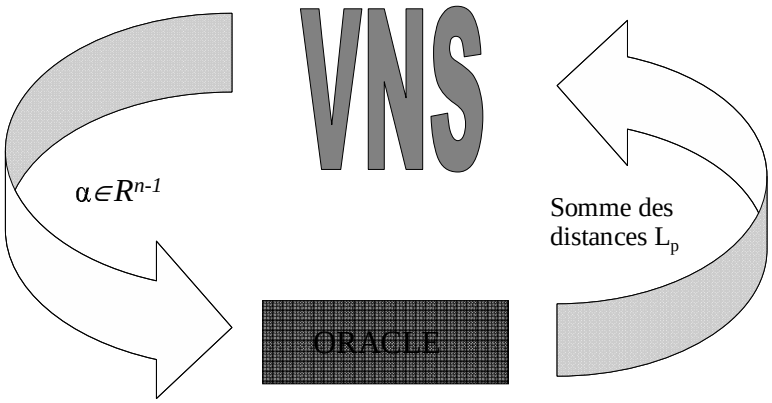


Approche heuristique

La décomposition:
• fixer une direction
• projeter les points sur le rayon
• trouver meilleure position pour plan



Approche heuristique

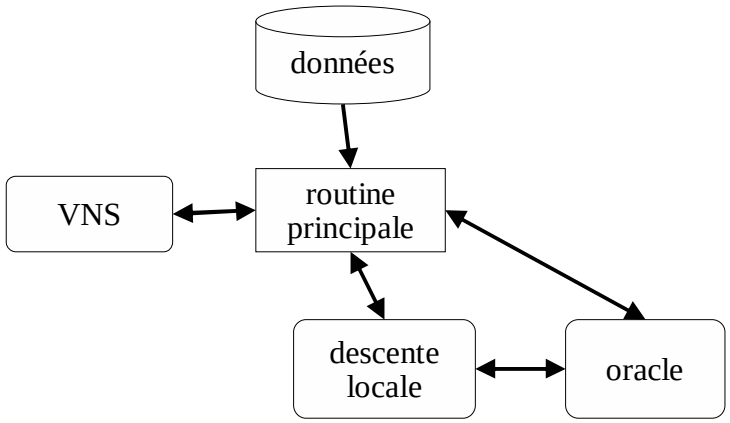


Approche heuristique

VNS

- structure des voisinages
- descente locale
- paramètres

Approche heuristique: Structure du code



Plan de la présentation

- Motivation et exposition du problème
- Séparation en norme $L_{p:}$
  - Survol d'approches exactes de solution
  - Méthode heuristique proposée
  - Expériences numériques
- Credit Scoring:
  - État du travail
  - Défis
  - Tâches

Expériences numériques

- accélération des méthodes exactes
- validation des résultats heuristiques

problèmes aléatoires

- validation des résultats heuristiques
- effet du choix de la norme sur vrais problèmes

UCI repository

Expériences numériques:

précision: taille fixe

		gap		
dim	obs	L1-norm	L2-norm	L∞-norm
4	2000	0.00%	0.00%	0.00%
5	2000	0.00%	0.00%	0.00%
6	2000	0.00%	0.00%	0.00%
7	2000	0.51%	0.00%	0.00%
8	2000	0.53%	0.00%	0.00%
9	2000	0.01%	0.00%	0.05%
10	2000	0.01%	0.00%	0.00%
11	2000	0.00%	0.00%	0.02%
12	2000	2.09%	0.01%	0.04%
13	2000	10.12%	0.01%	0.27%

Expériences numériques

précision: dimension fixe

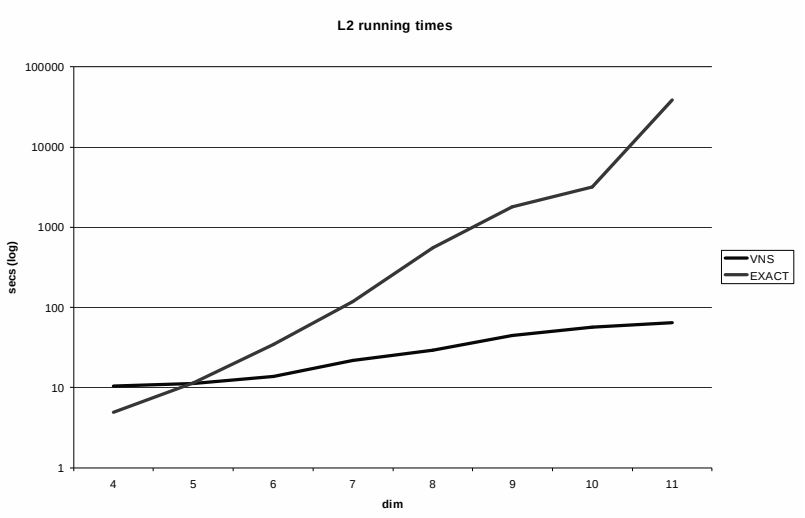
		gap		
dim	obs	L1-norm	L2-norm	L∞-norm
6	2000	0.00%	0.00%	0.00%
6	4000	0.00%	0.00%	0.00%
6	6000	0.21%	0.00%	0.00%
6	8000	0.00%	0.00%	0.00%
6	10000	0.00%	0.00%	0.00%
6	12000	0.00%	0.00%	0.00%
6	14000	0.00%	0.00%	0.00%
6	16000	0.00%	0.00%	0.00%
6	18000	1.27%	0.00%	0.00%
6	20000	0.00%	0.00%	0.00%
6	40000	0.00%		
6	60000	0.00%		
6	80000	3.23%		

Expériences numériques

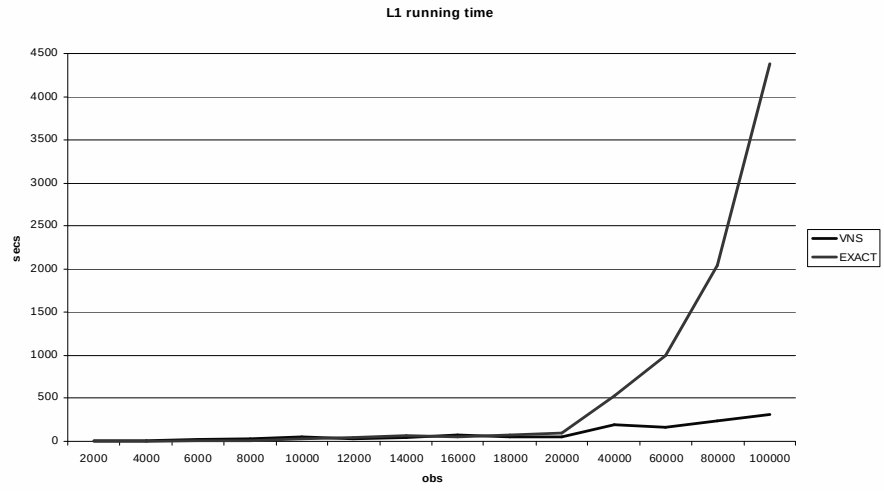
précision : UCI

			gap		
problem	dim	obs	L1-norm	L2-norm	L∞-norm
cancer	9	683	0.00%	0.50%	0.01%
echocardiogram	6	74	0.00%	0.03%	0.00%
glass	9	214	2.33%	0.79%	0.82%
hepatitis	16	150	0.60%	1.17%	3.49%
housing	13	506	0.17%	0.57%	4.50%
pima	8	768	0.00%	0.00%	0.00%

Expériences numériques: vitesse (taille fixe)



Expériences numériques: vitesse (dimension fixe)



Expériences numériques: accélération exacte

L_∞-norm results on S6k series

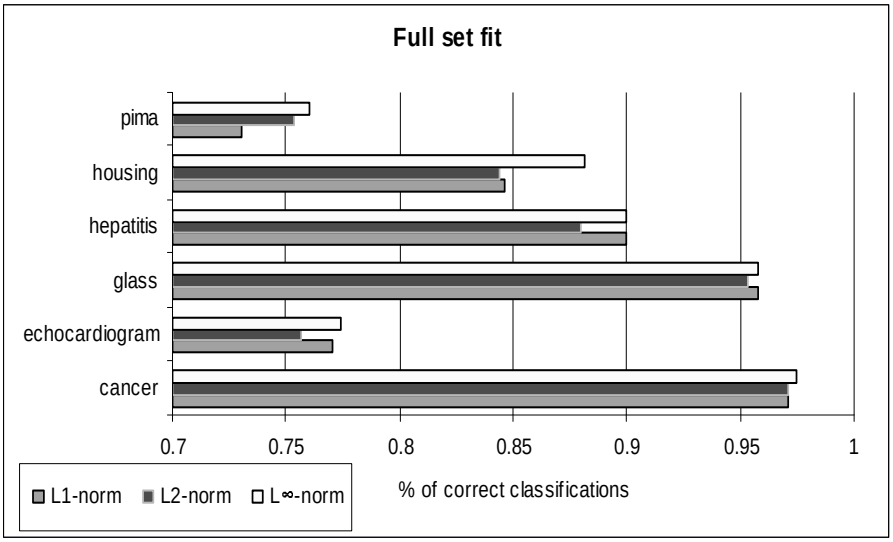
Problem size		Exact solution				Global time reduction
		Obj	Fit	Time		
n	$m+k$			Without init. sol.	With init. sol.	
6	10000	26.2138	88.96%	156.3	96.7	27%
6	20000	49.6972	88.74%	786.4	461.4	37%
6	30000	73.1058	89.11%	1774.4	1197.7	30%
6	40000	100.85	89.01%	3972.7	2326.4	39%
6	50000	116.317	88.95%	6004.2	3428.2	41%
6	60000	137.682	89.07%	7655.7	5050.1	32%
6	70000	159.39	89.23%	15215.0	7495.1	50%
6	80000	190.945	88.86%	21864.7	10639.1	51%
6	90000	202.645	89.02%	20072.5	11650.6	41%
6	100000	218.939	88.99%	35206.3	14020.3	60%

Expériences numériques: accélération exacte

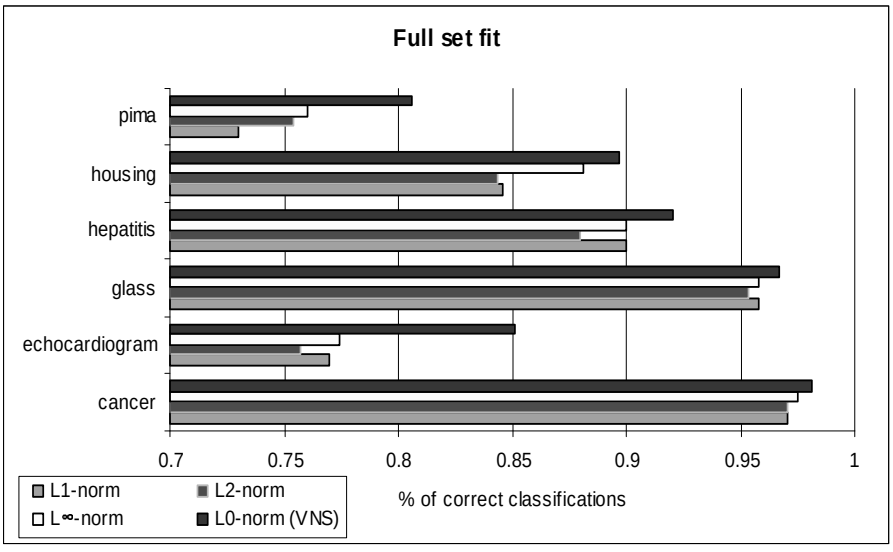
L_2-norm results on 2k series

Problem size		Exact solution				Global time reduction
		Obj	Fit	Time		
n	$m+k$			Without init. sol.	With init. sol.	
4	2000	3.10957	94.41%	5.0	6.3	-238.0%
5	2000	3.45771	93.84%	11.4	17.4	-149.2%
6	2000	4.33043	93.23%	34.5	38.1	-50.8%
7	2000	5.03871	91.75%	118.3	95.6	0.8%
8	2000	5.96413	90.16%	557.4	223.1	54.7%
9	2000	6.29355	90.17%	1796.3	680.1	59.7%
10	2000	6.48006	89.55%	3194.4	1728.4	44.1%
11	2000	11.27714	83.74%	38530.5	17491.7	54.4%
12	2000	7.26097	89.02%	-	18970.8	-
13	2000	9.49367	86.15%	-	60818.1	-

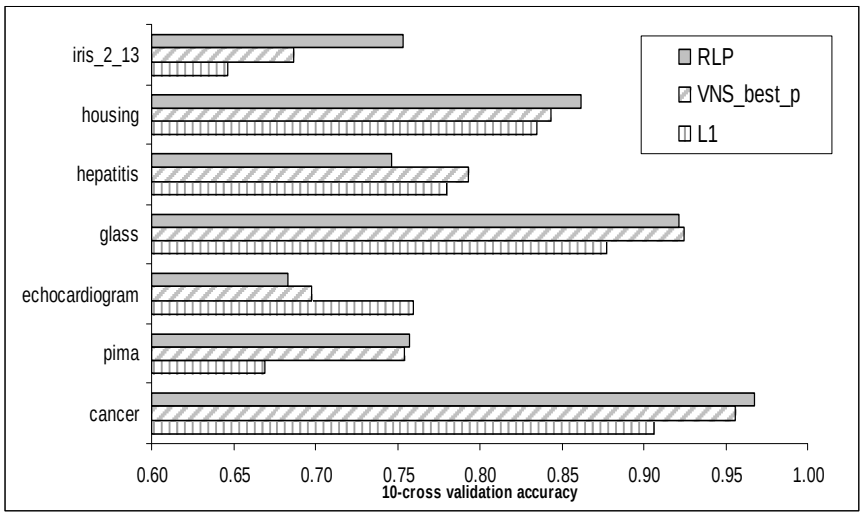
Expériences numériques: ajustement



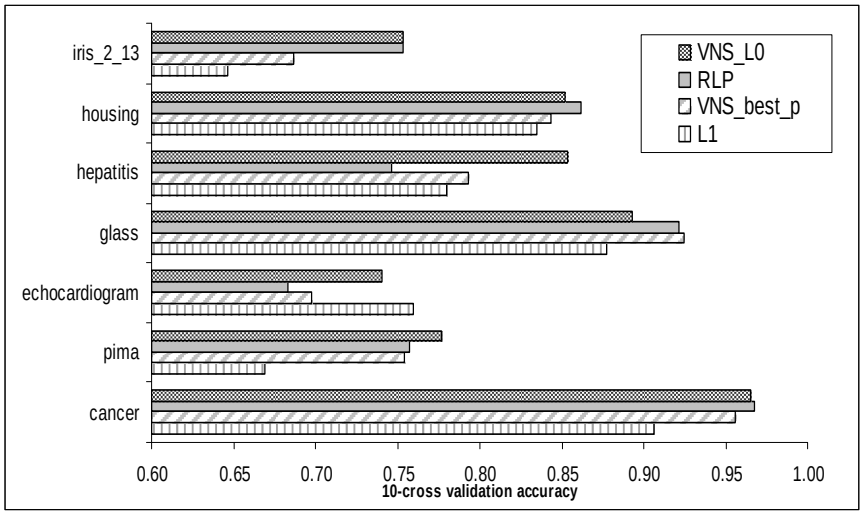
Expériences numériques: ajustement



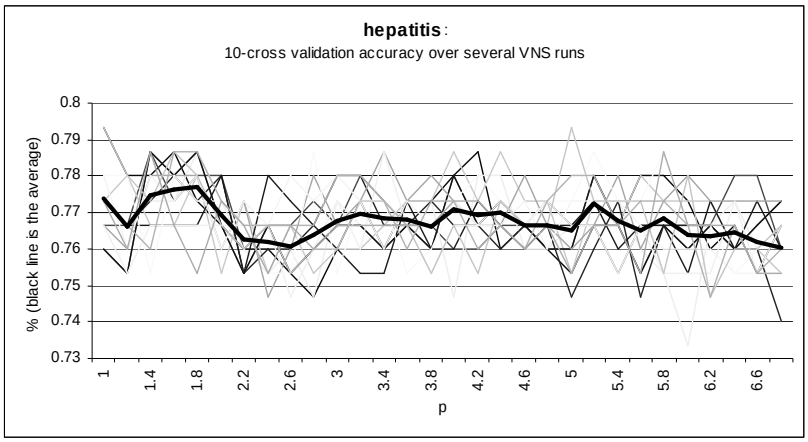
Expériences numériques: généralisation



Expériences numériques: généralisation

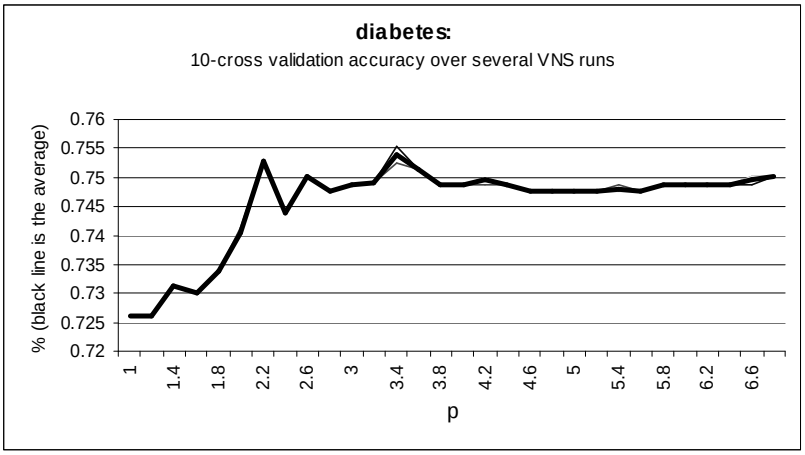


Expériences numériques: p arbitraire



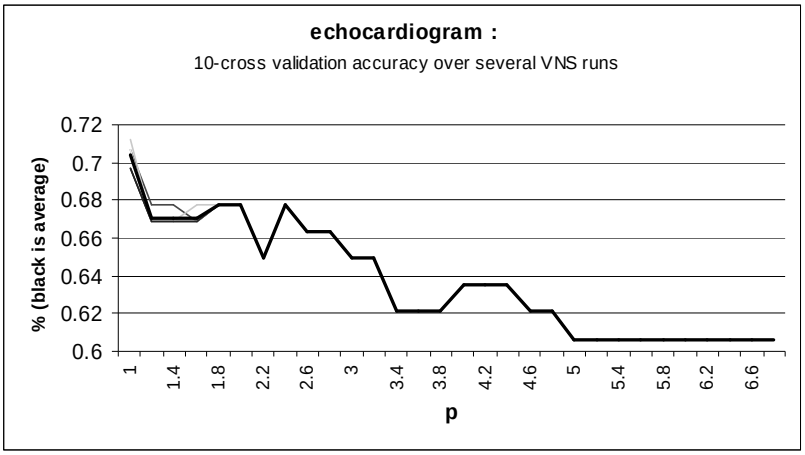
Pour certains problèmes le choix de la norme ne paraît pas important

Expériences numériques: p arbitraire



Mais pour d'autres la norme semble importante.

Expériences numériques: p arbitraire



Mais pour d'autres la norme semble importante.

Séparation en norme L_p Conclusions

- La méthode semble marcher très bien sur des problèmes à faible dimension, même avec des tailles plus grandes que les méthodes exactes ne peuvent gérer.
- L'utilisation pour accélération des méthodes exactes en normes 1, 2 et ∞ est intéressante pour les gros problèmes

Séparation en norme L_p Conclusions

- On peut pour la première fois explorer les conséquences d'un choix vraiment arbitraire de la norme.
- Pour certains problèmes cela paraît pertinent, mais l'interprétation reste un mystère.

Séparation en norme L_p Conclusions

- L'heuristique pour la norme 0 semble marcher très bien et elle est concurrentielle par rapport aux méthodes alternatives de discrimination linéaire pour les problèmes à faible dimension, même avec un grand nombre de points

Séparation en norme L_p Conclusions

- Hélas! Les problèmes de credit scoring sont typiquement à haute dimension
- Alors …

Plan de la présentation

- Motivation et exposition du problème
- Séparation en norme $L_{p:}$
  - Survol d'approches exactes de solution
  - Méthode heuristique proposée
  - Expériences numériques
- Credit Scoring:
  - État du travail
  - Défis
  - Tâches

Application au Credit Scoring

- Application dans un sens ample du mot:
  - discrimination linéaire
  - par programmation mathématique
- Exploitation des cadres conceptuelles et intuitions
- Possible adaptation des outils

Bases de données

- Banque commerciale:
  - automobile:
    - 60,000 demandes
    - 24,000 dossiers
  - hypothèques:
    - 19,000 demandes
    - 5,000 dossiers
  - carte de crédit:
    - 121,300 demandes
    - 52,200 dossiers

Bases de données

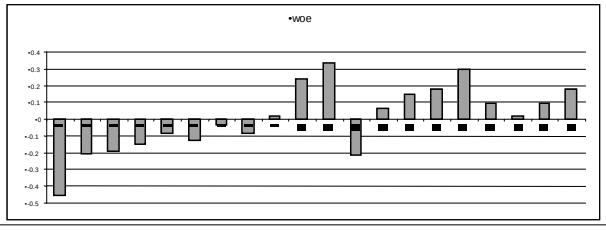
- Société hypothécaire :
  - données de plusieurs institutions financières, avec portefeuilles fondés en deuxième étage par la société
  - 200,000 demandes
    - nombre de caractéristiques variable (35 à 120)
  - 140,000 dossiers
- Benchmark industriel de pérformance

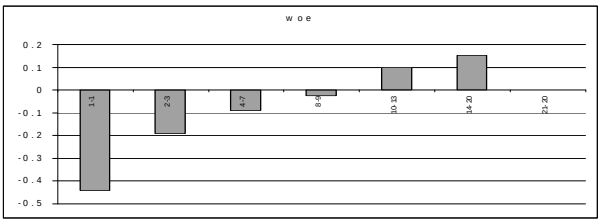
Premières expériences

- Base de données automobile
- Pre-processing:
  - détection d'erreurs et omissions
- Traitement des variables catégoriques:
  - catégorisation des autres!
  - bricolage avec woe (weights of evidence)
- Approches naïves

Catégorisation des variables numériques

exemple: property to loan ratio





Les défis

Performance gap

Feature selection

Lignes à explorer

- Hybrid classification
  - Chen & Mangasarian (1996)
- Train set editing
  - Duda, Hart & Stork (2000)
- Formulations qui adressent la sélection de colonnes:
  - Bennett, Demiriz et al. (2000)
  - Bradley, Mangasarian & Street (1997)

Conclusion

- Sujet fascinant
- Contribution bien définie dans l'application de VNS à un problème de classification
- Idées pour approcher le problème du crédit scoring

Structure de la thèse proposée

Introduction

Arbitrary-norm Distance Minimization by VNS

Misclassification Minimization by VNS

Application to Credit Scoring