SVM, apprentissage de noyaux et filtrage vaste marge

Créatis 2011

Rémi Flamary, Alain Rakotomamonjy, Stéphane Canu

LITIS EA 4108, INSA-Université de Rouen
76800 Saint Etienne du Rouvray, France

10 janvier 2011

Table des Matières

Plan

Le Séparateur à Vaste Marge

Exemple : Détection de piétons

- Systèmes d'aide à la conduite.
- Tâche : apprendre à partir d'exemples pour discriminer des images contenant des piétons.

Qu'est ce que le SVM ?

- Une famille d'algorithme d'apprentissage supervisé.
- Entrée : ensemble d'apprentissage

$$S = \{(x_1, y_1), \cdots, (x_n, y_n)\}$$

d'objets $x_i \in \mathcal{X}$ et leur classe connue $y_i \in \{+1, -1\}$.

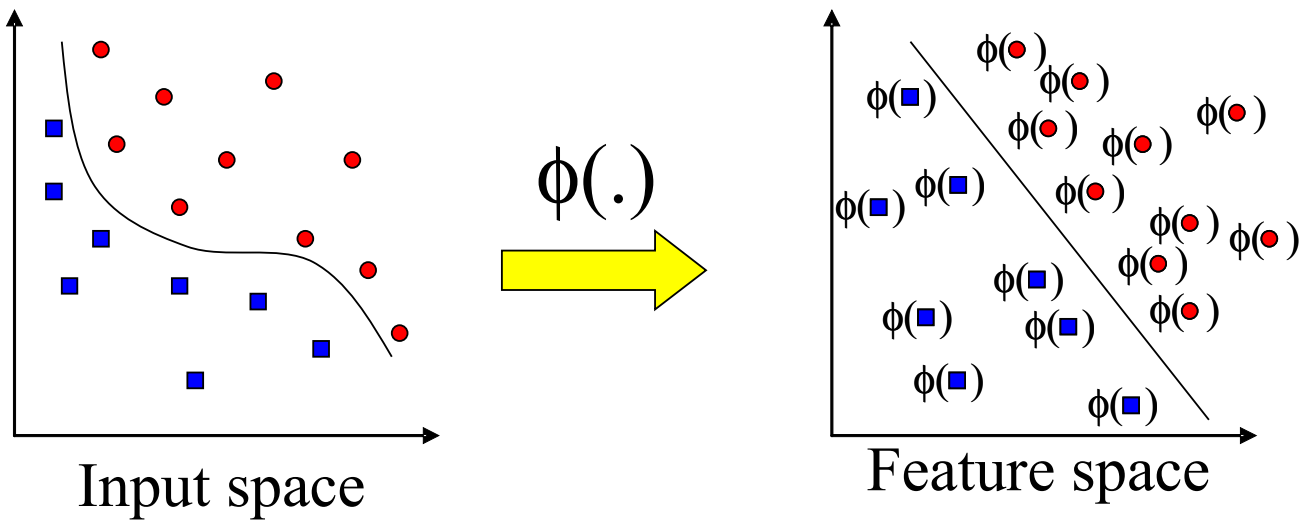
- Sortie : un classifieur $f : \mathcal{X} \longrightarrow \{+1, -1\}$ qui prédit la classe d'un objet $x \in \mathcal{X}$.

Principe des SVM

- Projeter l'ensemble d'apprentissage dans un espace de grande dimension $\mathcal{H}$ en utilisant la projection $\Phi(x)$. Dans l'espace $\mathcal{H}$, trouver un séparateur linéaire

$$f(x) = sign(\langle w, \Phi(x) \rangle_{\mathcal{H}} + b)$$

- ... qui maximise la marge m en classifiant, correctement les exemples.

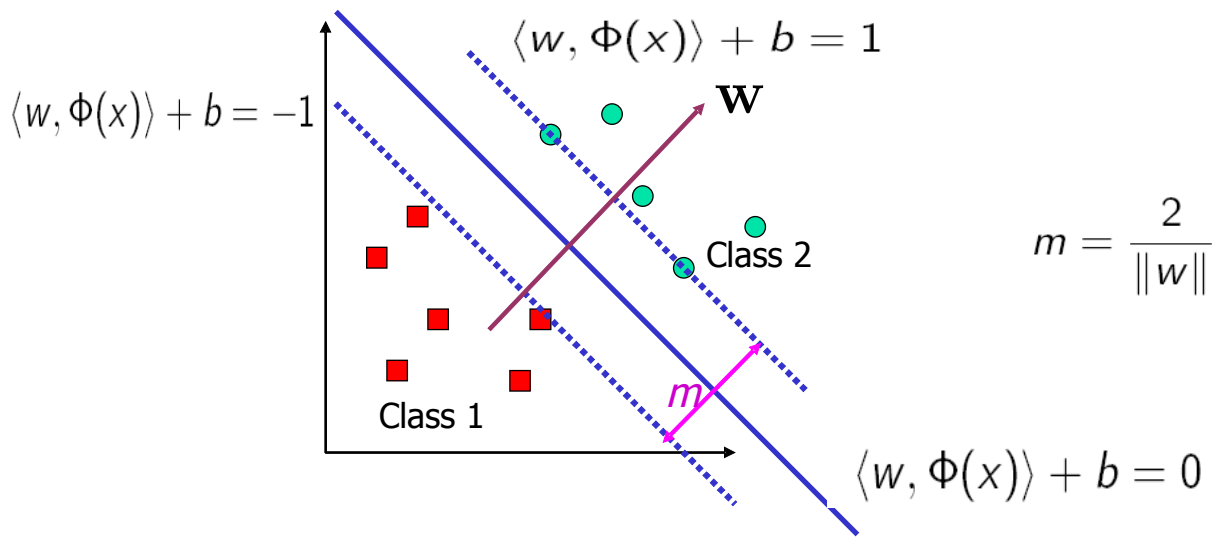


Principe des SVM

- Projeter l'ensemble d'apprentissage dans un espace de grande dimension $\mathcal{H}$ en utilisant la projection $\Phi(x)$. Dans l'espace $\mathcal{H}$, trouver un séparateur linéaire

$$f(x) = sign(\langle w, \Phi(x)\rangle_{\mathcal{H}} + b)$$

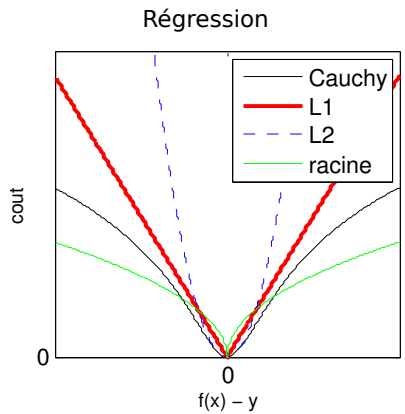
- ... qui maximise la marge m en classifiant, correctement les exemples.



Une histoire de coût

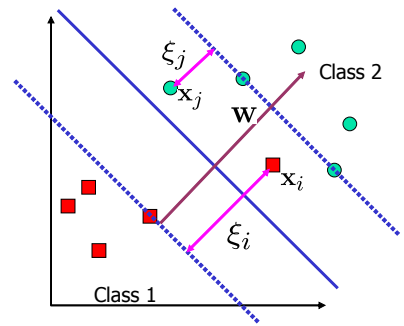
$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \underbrace{\frac{1}{2}\|\mathbf{w}\|^2}_{\text{maximisation de la marge}} + C \underbrace{\sum_{i=1}^{n} H(\mathbf{y}_i, \mathbf{w}^\top \Phi(\mathbf{x}_i) + b)}_{\text{bonne classif. des exemples}}$$

$$H(y, f(x)) = max(0, 1 - y * f(x))$$

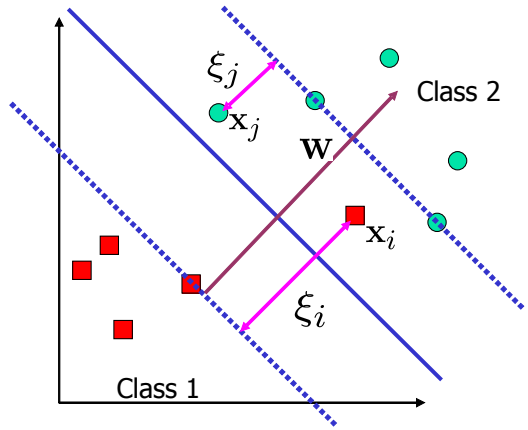


Minimisation du Coût Charnière

- H coût charnière ou hinge.
- Coût de classification non différentiable.



Le problème d'optimisation



Primal

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{avec} & y_i(\mathbf{w}^\top\Phi(\mathbf{x}_i)+b) \geq 1 \\ & i = 1, n \end{cases}$$

Dual

$$\begin{cases} \min\limits_{\alpha\in\mathbb{R}^n} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top\alpha \\ \text{avec} & \mathbf{y}^\top\alpha = 0 \\ \text{et} & 0 \leq \alpha_i \leq C \qquad i = 1, n \end{cases}$$

G matrice des influences
$(G_{i,j} = \mathbf{y}_i\mathbf{y}_j\langle\Phi(x_i),\Phi(x_j)\rangle)$

$$f(\mathbf{x}) = sign(w^\top\Phi(x)+b) = sign(\sum_{i=1}^{n}\alpha_i\, y_i\langle\Phi(x_i),\Phi(x)\rangle + b)$$

Kernel trick

- Les données apparaissent uniquement sous la forme de produits scalaires $\langle \Phi(x_i), \Phi(x_j) \rangle$
- On oublie $\Phi(x)$, on utilise une fonction noyau $k(x, y)$ qui agit comme un produit scalaire $\langle \phi(x_i), \phi(x_j) \rangle$.
- La fonction de décision est :

$$f(x) = sign(\sum_j \alpha_j \mathbf{y}_i K(x, x_j) + b)$$

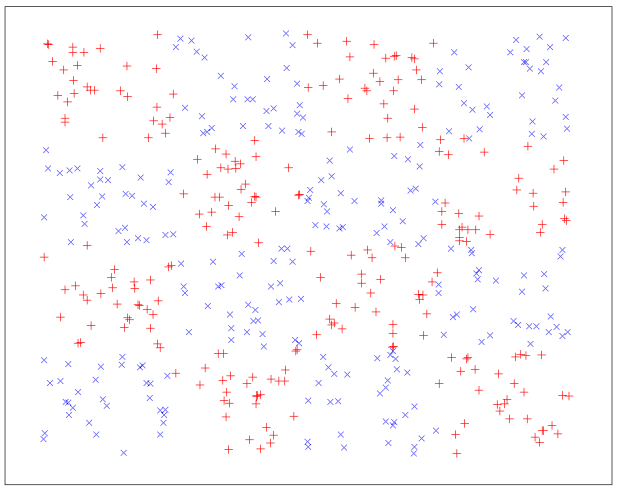
- Pas besoin d'exprimer $\Phi(x)$, les caractéristiques peuvent être de taille infinie.

Noyaux classiques

- Linéaire (espace d'origine), $k(s, t) = s^\top t$
- Polynômial, $k(s, t) = (s^\top t)^p$
- Gaussien, $k(s, t) = \exp\left(-\frac{r^2}{2\sigma^2}\right), \quad r = \|s - t\|$

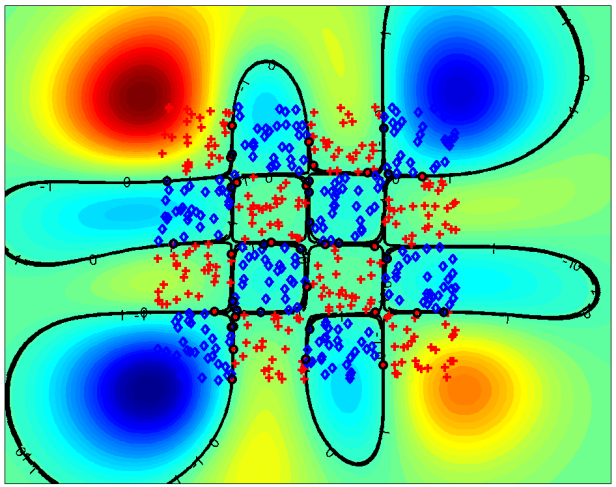
Exemple d'utilisation

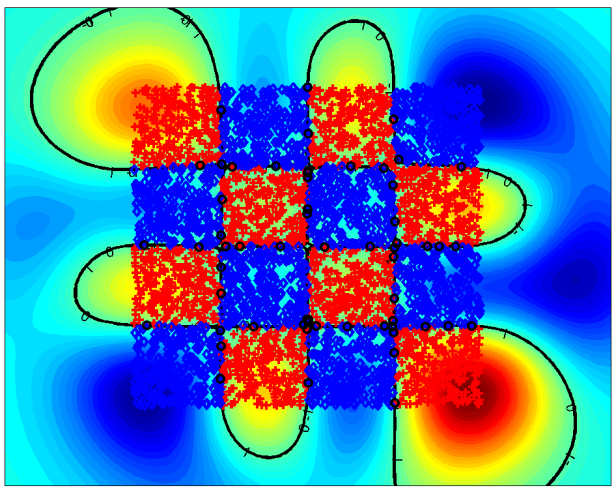
Problème de l'échiquier :



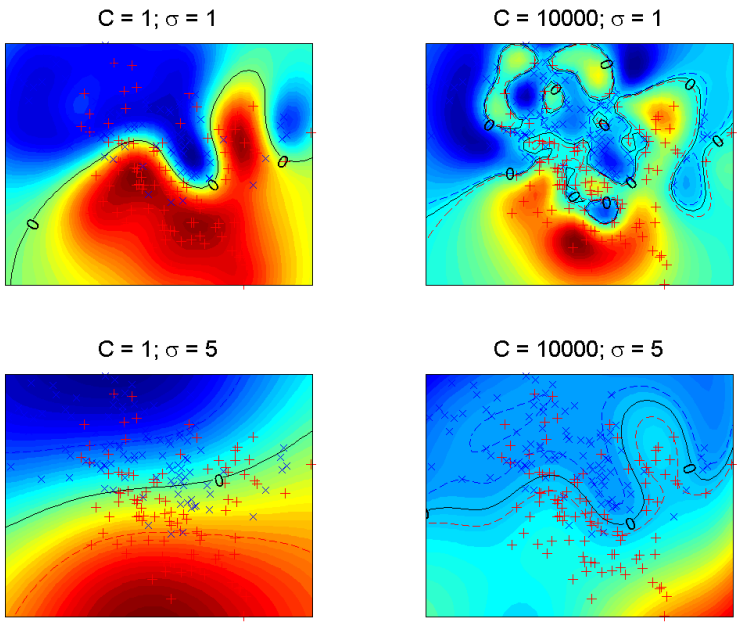
- Données séparables.
- Fortement non-linéaires.
- Très peu de vecteurs supports sélectionnés (en noir).

Résultats (500 et 5000 points d'apprentissage) :





Conclusion sur les SVM



- Méthode efficace de discrimination (Compétitions, applications, ...).
- Fondement théorique solide (Théorie de Vapnik).
- Possibilité de gérer les problèmes non linéaires et la discrimination d'objets structurés à l'aide de noyaux.
- Peu de paramètres (C,noyau) mais leur choix est déterminant (habituellement validation croisée).

Plan

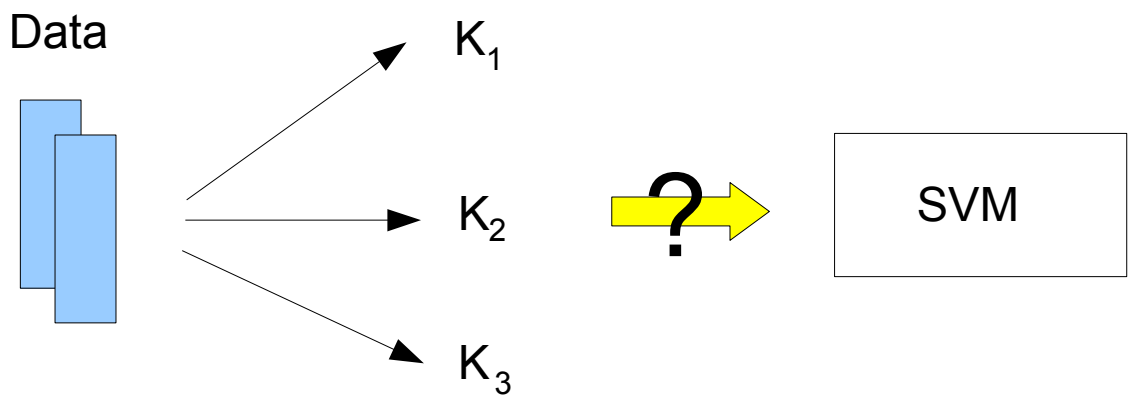
Apprentissage de noyau

Sélection de modèle

- Comment choisir le noyaux ? son/ses paramètre(s) ?
- Laisser l'algorithme apprendre une fonction de décision et apprendre le noyau.

Fusion de sources

- Les données peuvent être représentées selon différentes sources d'information
- Laisser l'algorithme choisir les meilleurs noyaux.



Principe de l'apprentissage de noyau

Problème SVM Dual

$$J(k) = \begin{cases} \min\limits_{\alpha \in \mathbb{R}^n} & \frac{1}{2}\alpha^\top G\alpha - \mathbf{e}^\top \alpha \\ \text{avec} & \mathbf{y}^\top \alpha = 0 \\ \text{et} & 0 \leq \alpha_i \leq C \qquad\qquad i = 1, n \\ & G_{i,j} = \mathbf{y}_i \mathbf{y}_j k(x_i, x_j) \end{cases}$$

But

Apprendre un noyau maximisant la marge :

$$\min_k J(k) \text{ avec } k \in \mathcal{S}$$

Ici $\mathcal{S}$ représente l'ensemble de recherche pour le noyau, il permet d'éviter le sur-apprentissage. Dans la littérature, les ensembles proposés sont :

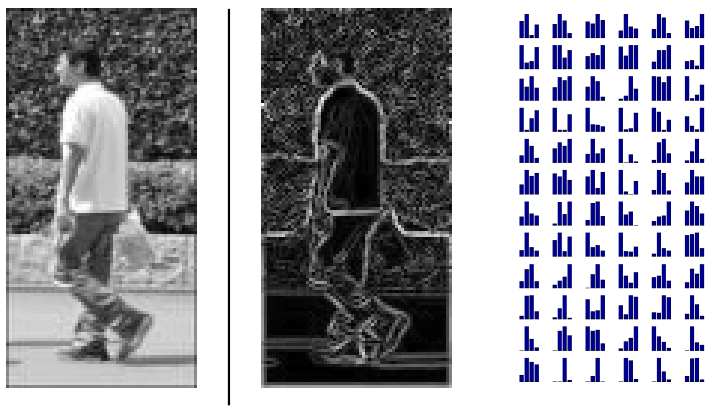
- Une combinaison linéaire de noyaux
  [Lanckriet et al., 2004, Rakotomamonjy et al., 2008, Bach et al., 2004].
- Une multiplication de noyaux [Grandvalet and Canu, 2003, Varma and Babu, 2009].
- L'ensemble des noyaux Gaussiens
  [Grandvalet and Canu, 2003, Chapelle et al., 2002].

Les méthodes à noyaux multiples (MKL)

Problème : Détection de piétons

- Plusieurs types de caractéristiques.
- Gradient, couleur, forme, histogramme local de gradient.
- Plusieurs paramètres possibles pour l'extraction de caractéristiques.

$\Rightarrow$ Comment fusionner/sélectionner ces différentes sources ?

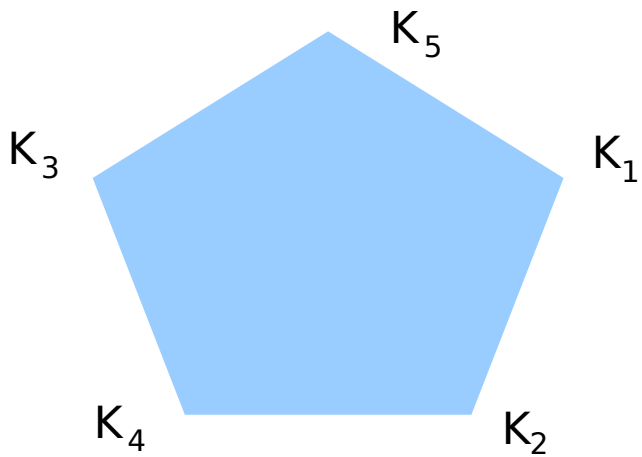


Apprentissage à noyaux multiples

- Un noyau par caractéristique/paramètre.
- Laisser l'algorithme sélectionner la combinaison linéaire optimale.

SimpleMKL [Rakotomamonjy et al., 2008]

Combinaison linéaire convexe de noyaux



$$k(x, y) = \sum_t d_t k_t(x, y)$$

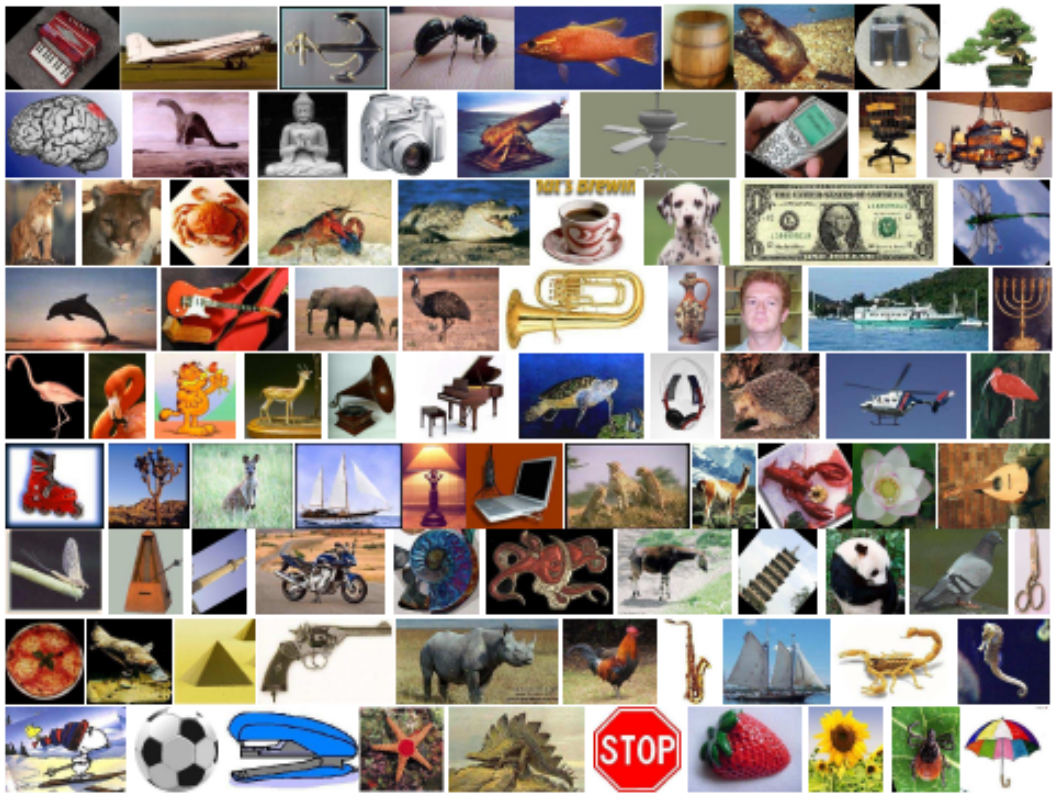
avec $d_t \geq 0$ et $\sum_t d_t = 1$

Problème de type simplex.

Avantages

- Parcimonie, sélection de noyaux.
- Formulation convexe du problème.
- Algorithme simple et efficace.

Exemple pour un problème multiclasse : Caltech 101



- Problème de reconnaissance d'objets : 101 classes + fond
- 4 caractéristiques : 2 histogrammes (forme et gradient), 2 d'apparence (couleur,...)
- Données et noyaux disponibles (Visual Geometry Group, Oxford UK)

Résultats

- 3060 images, 15 images en apprentissage et 15 images en test
- Noyaux gaussiens χ^2.
- 10 tirages, 12 noyaux.
- MKL un-contre-Un pour la sélection de modèle et de caractéristiques.

	Cues				
	Shape 1	Shape 2	App. 1	App. 2	MCMKL
Perf	69.8 ± 0.5	70.6 ± 0.6	71.6 ± 0.6	68.2 ± 0.8	76.6 ± 0.6

Autres méthodes d'apprentissage de noyaux

Noyaux Gaussiens

Noyau isotrope :

$$k(s,t) = \exp\left(-\frac{||s-t||^2}{2\sigma^2}\right)$$

$\sigma \in \mathbb{R}$ largeur de bande du noyau.

Noyau non isotrope :

$$k(s,t) = \exp\left(-\sum_u \frac{(s_u - t_u)^2}{\mathbf{b}_u^2}\right)$$

$\mathbf{b} \in \mathbb{R}^d$ la pondération de chaque caractéristique.

Apprendre σ

- Par descente de gradient [Chapelle et al., 2002].

Apprendre $\mathbf{b}$

- Avec parcimonie [Grandvalet and Canu, 2003].
- Par descente de gradient [Varma and Babu, 2009].

$\Rightarrow$ Problèmes non convexes, nécessité de régulariser.

Conclusion sur l'apprentissage de noyaux

Multiple Kernel Learning

- Bonnes performances, sélection automatique des bons noyaux.
- Problème convexe, algorithmes efficaces
  [Rakotomamonjy et al., 2008, Chapelle and Rakotomamonjy, 2008].
- Parcimonie sur les noyaux.
- Application : fusion de données, sélection d'hyperparamètres, sélection de variables non-linéaire.

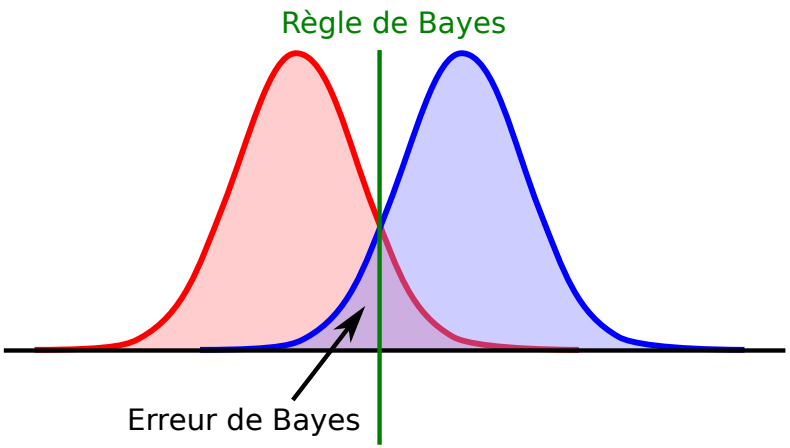
Apprentissage de paramètres

- Recherche continue des paramètres du noyau.
- Se fait généralement par descente de gradient
  [Chapelle et al., 2002, Varma and Babu, 2009].
- Problèmes non convexes, régularisation obligatoire.

Plan

Filtrage Vaste marge

Erreur de Bayes



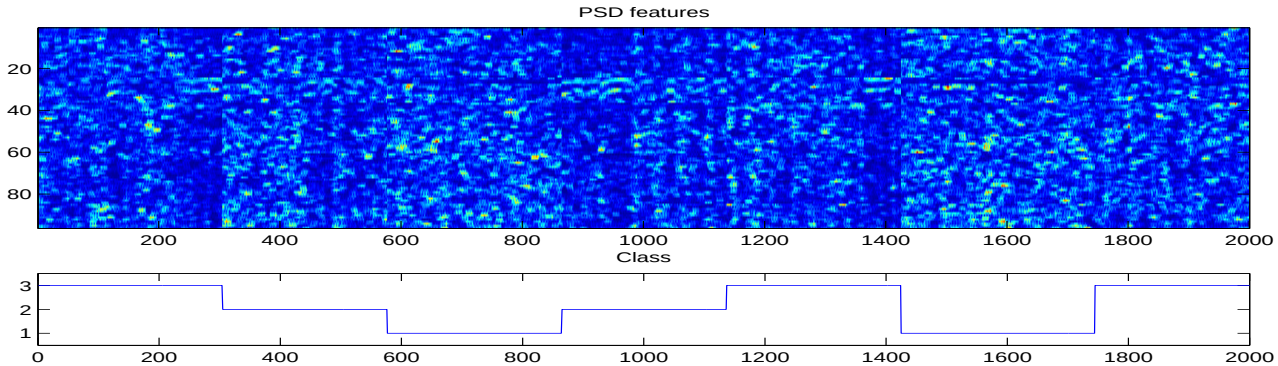
- Si on connaît les distributions de probabilité de chaque classe.
- On peut obtenir le meilleur classifieur : le classifieur de Bayes.
- Celui-ci commet toujours une erreur : l'erreur de Bayes.
- Les SVM convergent vers ce classifieur, mais ils supposent que les échantillons sont IID (indépendants et identiquement distribués).

Si les échantillons ne sont pas IID ?

- Dans certains cas, ça marche quand même.
- Possibilité d'utiliser cette information.

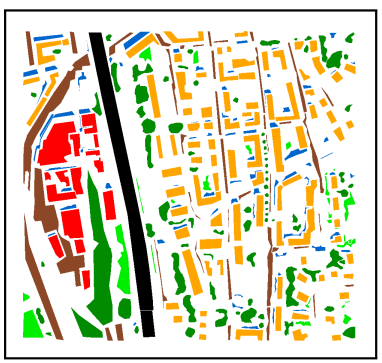
Exemples de données non IID

Signal temporel



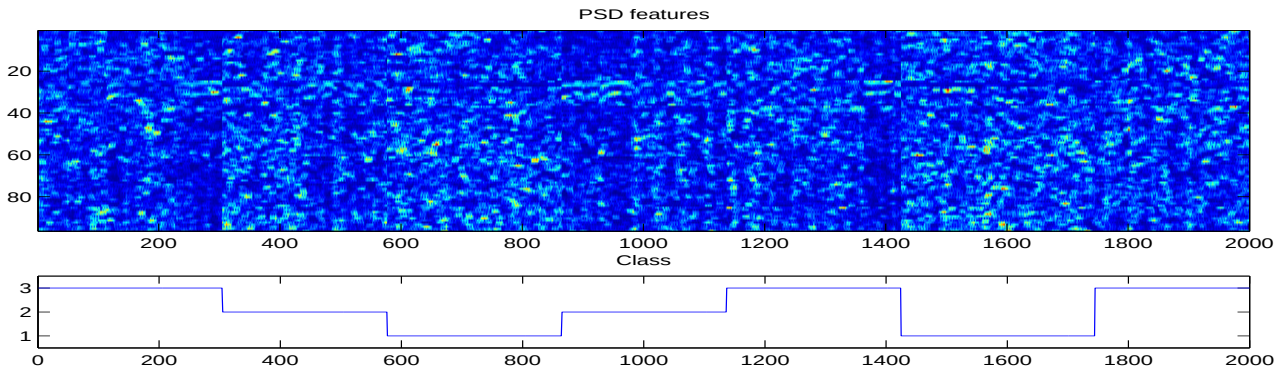
- Décodage d'état mental en BCI.
- Tâche de classification des échantillons temporels, segmentation du signal.

Image



- Détection de bâtiments en imagerie multispectrale.
- Tâche de classification de pixels, segmentation d'image.

Filtrage Vaste Marge



Exemple

Décodage continu d'état mental en BCI

- Le sujet pense au mouvement de sa main droite, gauche ou à un mot.
- Signal multidimensionnel fortement bruité (bruit additif, bruit convolutionnel).

On filtre !

- Soit filtrage fixé a priori.
- Soit on apprend le filtre.

$\Rightarrow$ Apprentissage d'un filtre qui sépare au mieux les classes : le filtrage vaste marge.

Définitions

- $X \in \mathbb{R}^{N \times d}$ matrice des signaux, d canaux N échantillons.
- $\mathbf{y}$ contient les étiquettes.

Filtrage de X par F

$$\widetilde{X}_{i,j} = \sum_{m=1}^{f} F_{m,j} \; X_{i+1-m,j} \tag{1}$$

$F \in \mathbb{R}^{f \times d}$ matrice des filtres (un par canal), d filtres de taille f.

Noyaux

- Linéaire : $\widetilde{K}^F_{i,j} = \widetilde{X}_{i,.}^{\top} \widetilde{X}_{j,.}$
- Gaussien : $\widetilde{K}^F_{i,j} = k(\widetilde{X}_{i,.}, \widetilde{X}_{j,.}) = \exp\left(-\frac{||\widetilde{X}_{i,.} - \widetilde{X}_{j,.}||^2}{2\sigma_k^2}\right)$

Filtrage vaste marge

Principe

Apprendre le filtrage temporel et le classifieur de manière jointe.

Problème

$$\min_{g,F} \quad \frac{1}{2}||g||^2 + C\sum_{i=1}^{n} H(\mathbf{y}_i, g(\widetilde{X}_{i,.})) + \lambda\Omega(F) \tag{2}$$

avec λ un paramètre de régularisation et $\Omega(\cdot)$ une fonction de régularisation pour F

Problème non convexe, mais convexe par rapport à g pour un F fixe (SVM).

Solution

- Minimisation de la fonction objectif par descente de gradient.
- Régularisation

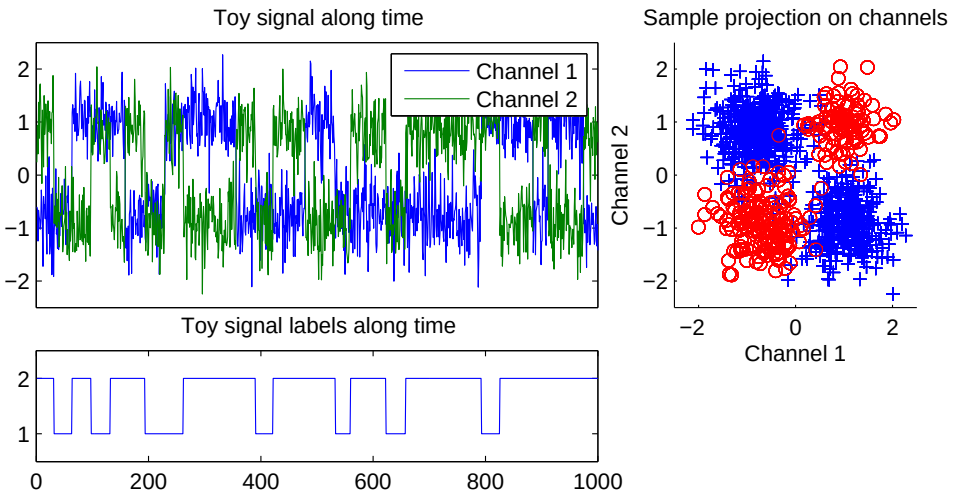
Frobenius :

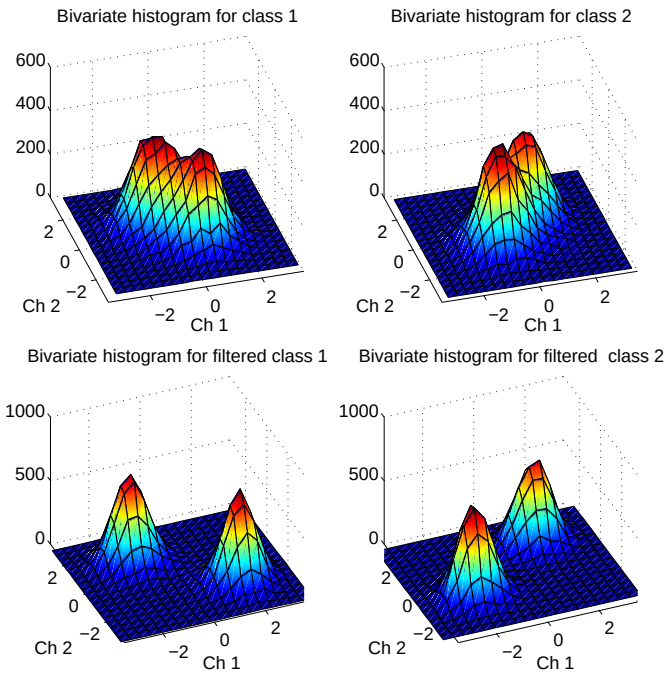
$$\Omega_2(F) = \sum_{u,v}^{f,d} F_{u,v}^2 \tag{3}$$

Norme mixte $\ell_1 - \ell_2$:

$$\Omega_{1-2}(F) = \sum_{v}^{d}\left(\sum_{u}^{f} F_{u,v}^2\right)^{\frac{1}{2}} \tag{4}$$

Données Jouet





Problème

- Non linéaire.
- Bruit Gaussien et convolutionnel (délai).
- Le filtrage sépare les classes.

Données BCI (Classification linéaire)

Method	Sub 1	Sub 2	Sub3	Avg
BCI Comp.	0.2040	0.2969	0.4398	0.3135
SVM	0.2877	0.4283	0.5209	0.4123
Filter-SVM				
$f = 8, n_0 = 0$	0.2337	0.3589	0.4937	0.3621
$f = 20, n_0 = 0$	0.2021	0.2693	0.4381	0.3032
$f = 50, n_0 = 0$	**0.1321**	**0.2382**	**0.4395**	**0.2699**
Avg-SVM				
$f = 100, n_0 = 50$	0.1544	0.2235	0.3870	0.2550
Filter-SVM				
$f = 100, n_0 = 50$	**0.0537**	**0.1659**	**0.3859**	**0.2018**

TABLE: Erreurs de test pour le Dataset BCI (BCI Compet. III).

Données

- 3 sujets, 96 canaux, densités spectrales de fréquence (PSD).
- 3 sessions d'apprentissage, 1 session de test.
- Sélection des paramètres par validation.

Visualisations BCI

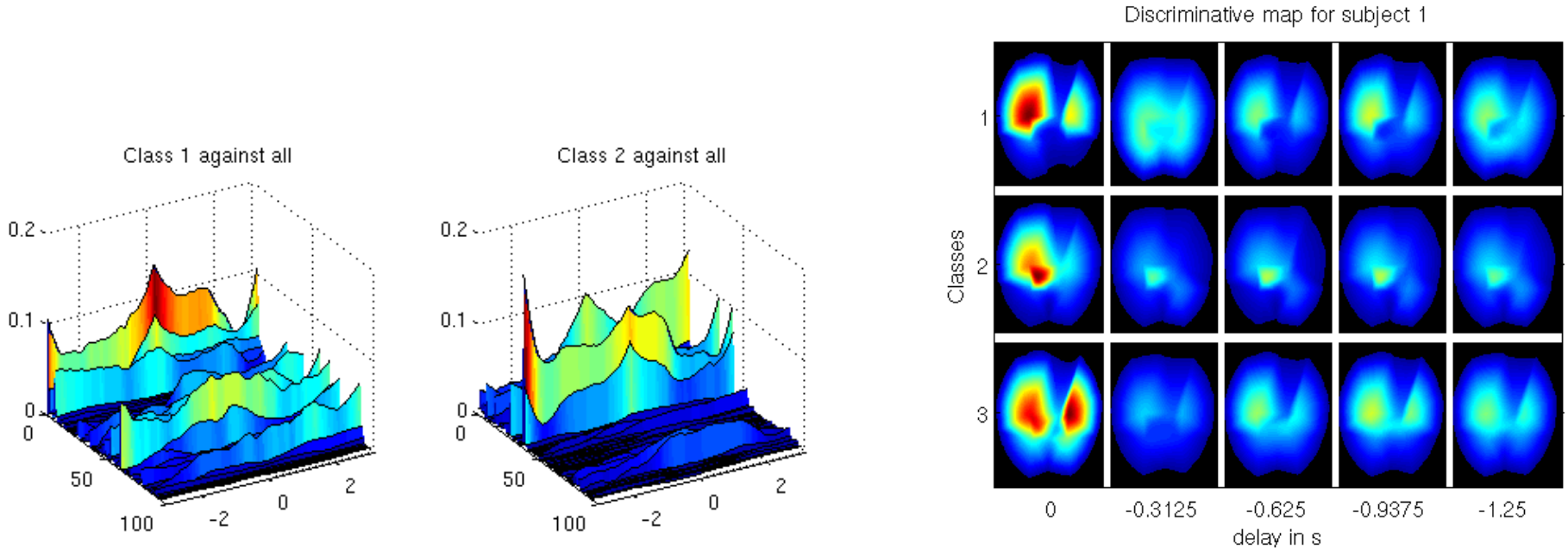
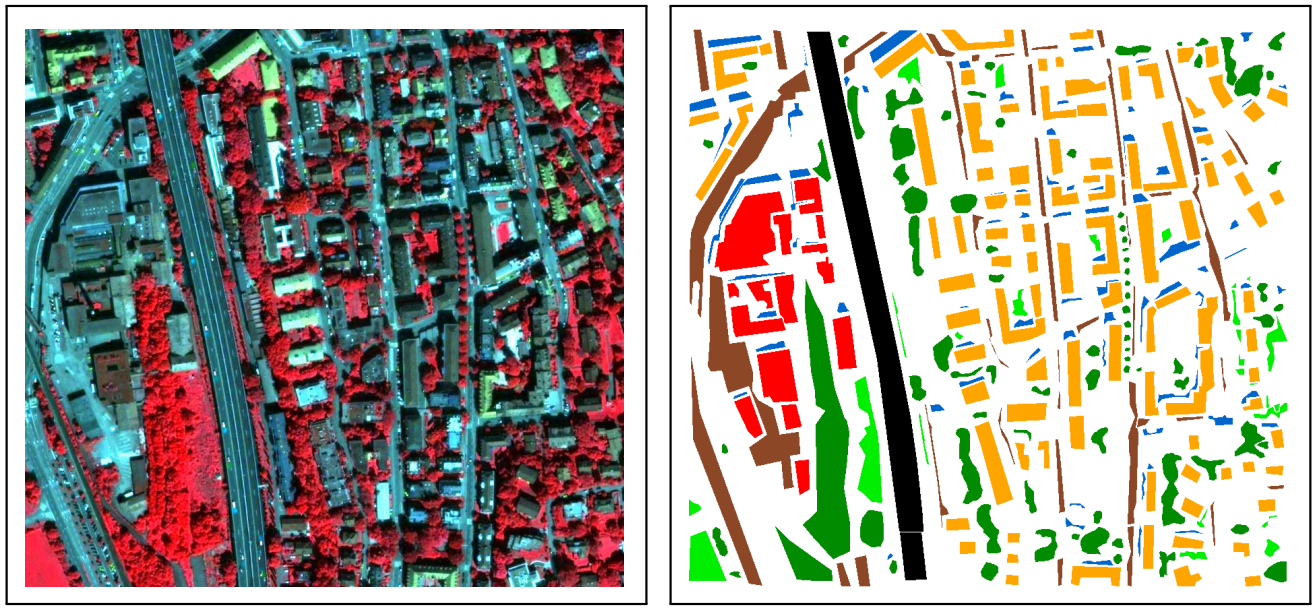


FIGURE: Filtres F pour différentes tâches de discrimination.

Visualisation des filtres

- Un filtre par canal.
- Carte de discrimination spatio/temporelle.

Extension 2D



Discrimination de pixels/ segmentation d'image

- Possibilité d'étendre l'approche à la discriminatiuon de pixels.
- Apprentissage d'un filtre de convolution 2D par canal (couleur).
- Images satellite HR de Zurich.

Résultats en 2D

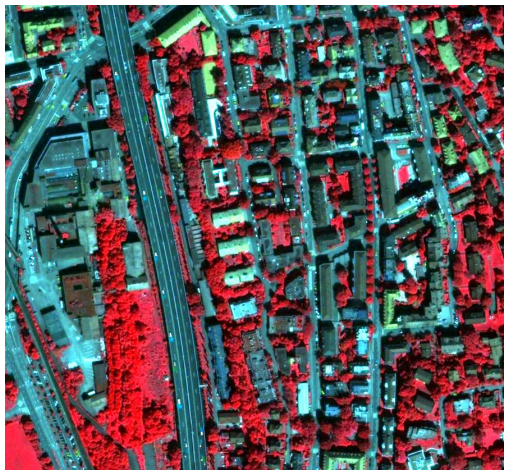
Method	Classes	Filter size	Training Pixels	[%]OA	Kappa
SVM				75.11	0.685
AvgSVM	7	9	$\sim$ 5000	83.68	0.796
WinSVM				82.98	0.785
KF-SVM				**85.32**	**0.816**
SVM				83.04	0.772
AvgSVM	6*	9	$\sim$ 5000	89.48	0.860
WinSVM				**91.71**	**0.889**
KF-SVM				91.45	0.885

Résultats

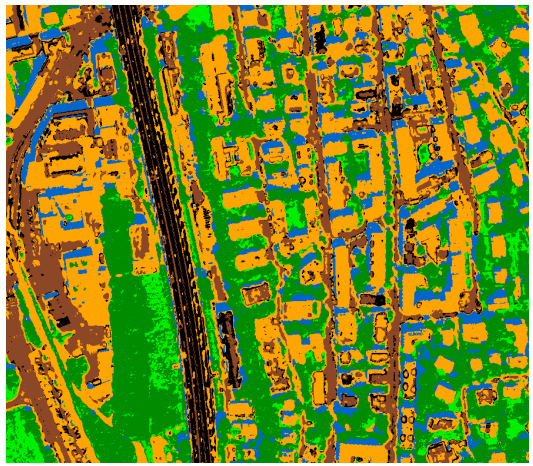
- Résultats équivalents à la classification d'une fenêtre autour du pixel.
- Mais la classification se fait sur un pixel unique.
- Préprocessing optimal, on reste en faible dimension.

Visualisation

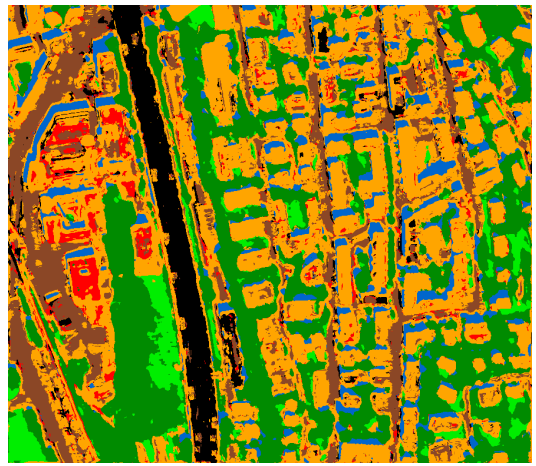
Image

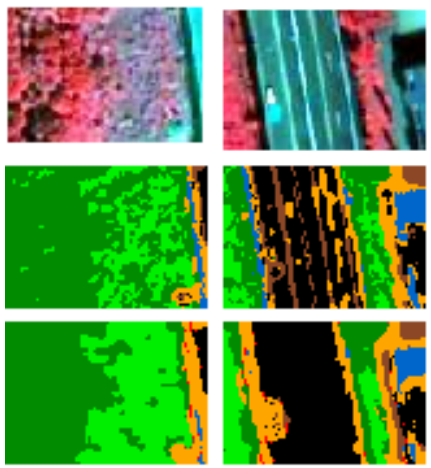


SVM



KF-SVM





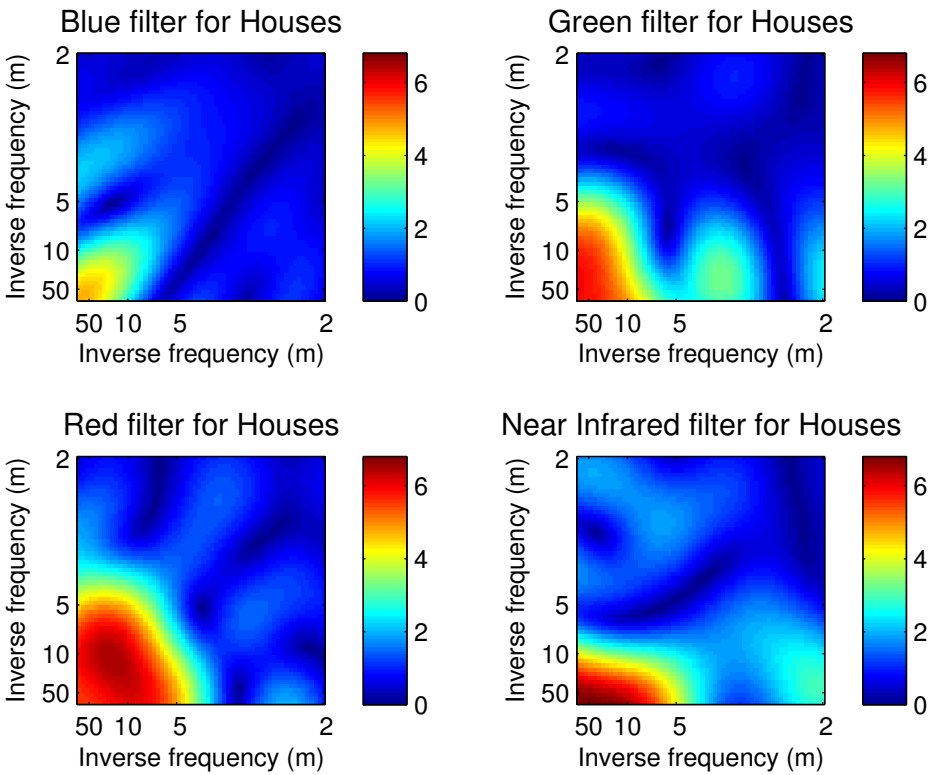
En image

- Frontières plus propres
- On détecte la classe rouge.

Visualisation du filtre (1)

Classe : Maisons, buildings résidentiels

Amplitude de la FFT du filtre pour différentes composantes.



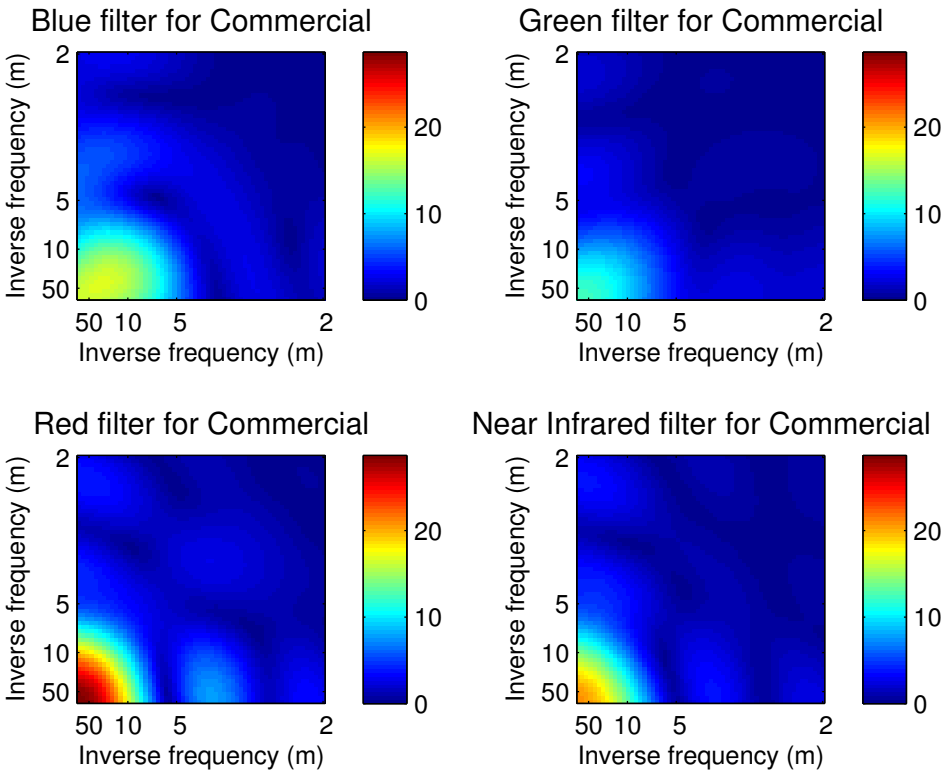
(Orange)

- Passe-bas large bande (maisons petites).

Filter Visualization (2)

Classe : Buildings commerciaux

Amplitude de la FFT du filtre pour différentes composantes.



(Red)

- Passe-bas petite bande passante pour promouvoir les gros building..

Conclusion sur le filtrage vaste marge

Conclusion

- Apprentissage d'un filtrage vaste marge.
- Discrimination d'échantillons.
- Sélection/pondération automatique des canaux.
- Visualisation des filtres.

Travaux futurs et en cours

- Gérer les données avec beaucoup de points d'apprentissage, sous échantillonnage.
- Nouveaux types de régularisation selon les connaissances a priori.

Biblio SVM : kernel-machines.org

- [Shawe-Taylor and Cristianini, 2004] John Shawe-Taylor and Nello Cristianini Kernel Methods for Pattern Analysis, Cambridge University Press, 2004
- [Schölkopf and Smola, 2001] Bernhard Schölkopf and Alex Smola. Learning with Kernels. MIT Press, Cambridge, MA, 2002.
- [Hastie et al., 2001] Trevor Hastie, Robert Tibshirani and Jerome Friedman, The Elements of Statistical Learning :. Data Mining, Inference, and Prediction, springer, 2001

- [Bottou, 2007] Léon Bottou, Olivier Chapelle, Dennis DeCoste and Jason Weston Large-Scale Kernel Machines (Neural Information Processing, MIT press 2007
- [Chapelle et al., 2006] Olivier Chapelle, Bernhard Scholkopf and Alexander Zien, Semi-supervised Learning, MIT press 2006

- [Vapnik, 1995] Vladimir Vapnik. The Nature of Statistical Learning Theory. Springer, 1995.

- [Wahba, 1990] Grace Wahba. Spline Models for Observational Data. SIAM CBMS-NSF Regional Conference Series in Applied Mathematics vol. 59, Philadelphia, 1990
- [Berlinet and Agnan, 2004] Alain Berlinet and Christine Thomas-Agnan, Reproducing Kernel Hilbert Spaces in Probability and Statistics,Kluwer Academic Publishers, 2003
- [Atteia and Gaches, 1999] Marc Atteia et Jean Gaches , Approximation Hilbertienne - Splines, Ondelettes, Fractales, PUG, 1999

Bibliographie I

[Atteia and Gaches, 1999] Atteia, M. and Gaches, J. (1999).
Approximation hilbertienne : Splines, Ondelettes, Fractales.
Presses Universitaires de Grenoble.

[Bach et al., 2004] Bach, F., Lanckriet, G., and Jordan, M. (2004).
Multiple kernel learning, conic duality, and the SMO algorithm.
In *Proceedings of the 21st International Conference on Machine Learning*, pages 41–48.

[Berlinet and Agnan, 2004] Berlinet, A. and Agnan, C. T. (2004).
Reproducing Kernel Hilbert Spaces in Probability and Statistics.
Kluwer Academic Publishers.

[Bottou, 2007] Bottou, L. (2007).
Large-scale kernel machines.
Mit Pr.

[Chapelle and Rakotomamonjy, 2008] Chapelle, O. and Rakotomamonjy, A. (2008).
Second order optimization of kernel parameters.
In *NIPS Workshop on Automatic Selection of Optimal Kernels*.

[Chapelle et al., 2006] Chapelle, O., Scholkopf, B., and Zien, A., editors (2006).
Semi-Supervised Learning.
MIT Press.

[Chapelle et al., 2002] Chapelle, O., Vapnik, V., Bousquet, O., and Mukerjhee, S. (2002).
Choosing multiple parameters for SVM.
Machine Learning, 46(1-3) :131–159.

[Grandvalet and Canu, 2003] Grandvalet, Y. and Canu, S. (2003).
Adaptive scaling for feature selection in svms.
In *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

Bibliographie II

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. (2001).
The Elements of Statistical Learning.
Springer-Verlag.

[Lanckriet et al., 2004] Lanckriet, G., Cristianini, N., El Ghaoui, L., Bartlett, P., and Jordan, M. (2004).
Learning the kernel matrix with semi-definite programming.
Journal of Machine Learning Research, 5 :27–72.

[Rakotomamonjy et al., 2008] Rakotomamonjy, A., Bach, F., Grandvalet, Y., and Canu, S. (2008).
SimpleMKL.
Journal of Machine Learning Research, 9 :2491–2521.

[Schölkopf and Smola, 2001] Schölkopf, B. and Smola, A. (2001).
Learning with Kernels.
MIT Press.

[Shawe-Taylor and Cristianini, 2004] Shawe-Taylor, J. and Cristianini, N. (2004).
Kernel methods for pattern analysis.
Cambridge Univ Pr.

[Vapnik, 1995] Vapnik, V. (1995).
The Nature of Statistical Learning Theory.
Springer, N.Y.

[Varma and Babu, 2009] Varma, M. and Babu, B. (2009).
More generality in efficient multiple kernel learning.
In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1065–1072. ACM.

[Wahba, 1990] Wahba, G. (1990).
Spline Models for Observational Data.
Series in Applied Mathematics, Vol. 59, SIAM.