Datamining 2: Classification supervisée Partie 4 : résumé des méthodes précédentes

M2 STD

October 7, 2015

Avantages/inconvénients des méthodes précédentes

	X quanti	X quali	X mélange quanti quali
PPV	ok	bof	bof
SVM	ok	non	non
arbres	ok	ok	ok

Avantages/inconvénients des méthodes précédentes

	Y deux modalités	Y $k > 2$ modalités
PPV	ok	ok
SVM	ok	bof
arbres	ok	ok

SVM si Y a $k > 2$ modalités

Deux méthodes:

1. **One versus all:** on effectue k classifications où on cherche a discriminer avec $Z = 1$ si $Y = i$ contre $Z = -1$ si $Y \neq -i$. PB de désequilibres de tailles de classes. Un nouvel individu est affecté a une classe avec la méthode suivante:
2. **One versus one:** on effectue $k(k-1)/2$ classifications où on discrimine $Y = i$ contre $Y = j$.

SVM si Y a $k > 2$ modalités

One versus all

1. si il existe un seul classifieur (numéro i) qui répond "classe 1" on affecte a la classe i
2. si il existe plusieurs classifieurs qui répondent "classe 1" on distingue les ex-aequo en prenant le SVM qui a donné les marges les plus grandes
3. si il n'existe pas de classigieur qui répondent "classe 1" on ne sait pas

SVM si Y a $k > 2$ modalités

One versus one
On affecte alors un nouveau point a la classe majoritaire... pb : il peut y avoir des ex-aequos

Avantages/inconvénients des méthodes précédentes

	“compréhension” des résultats
PPV	non
SVM	non
arbres	oui

performances attendues

PPV$\geq$SVM$\geq$Arbres

frontières attendues

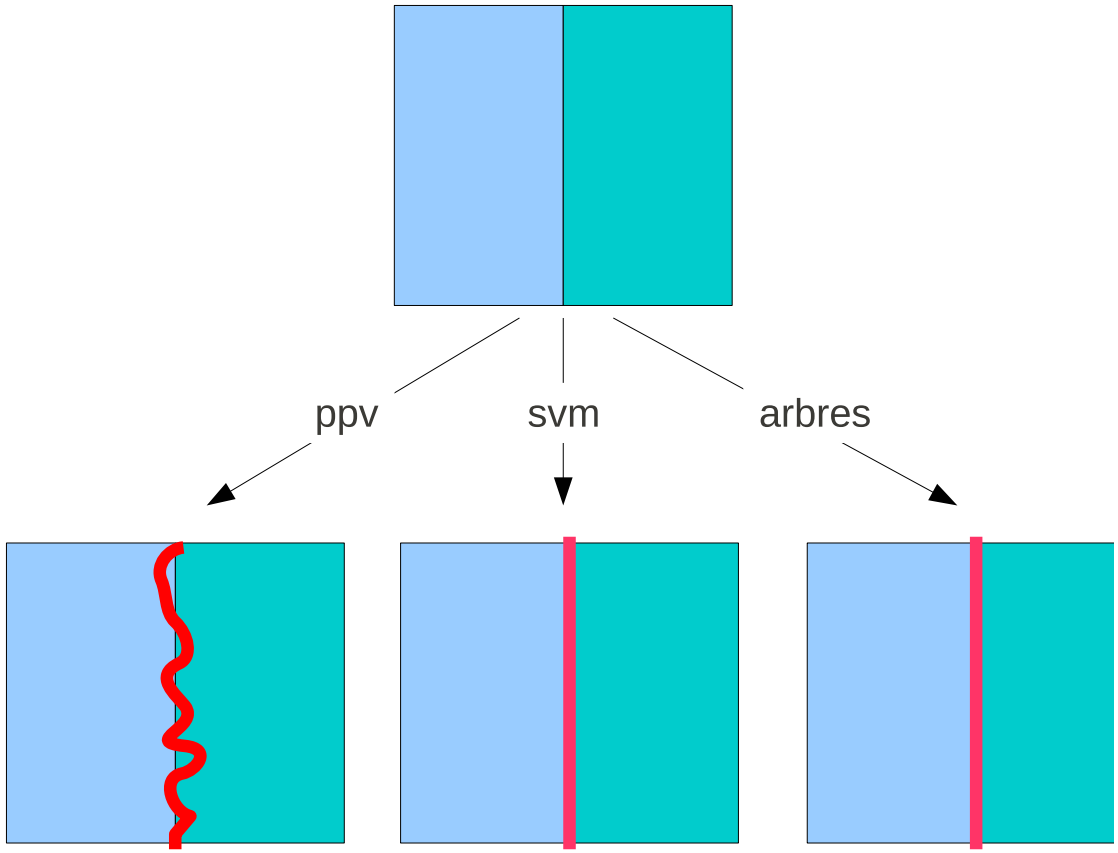
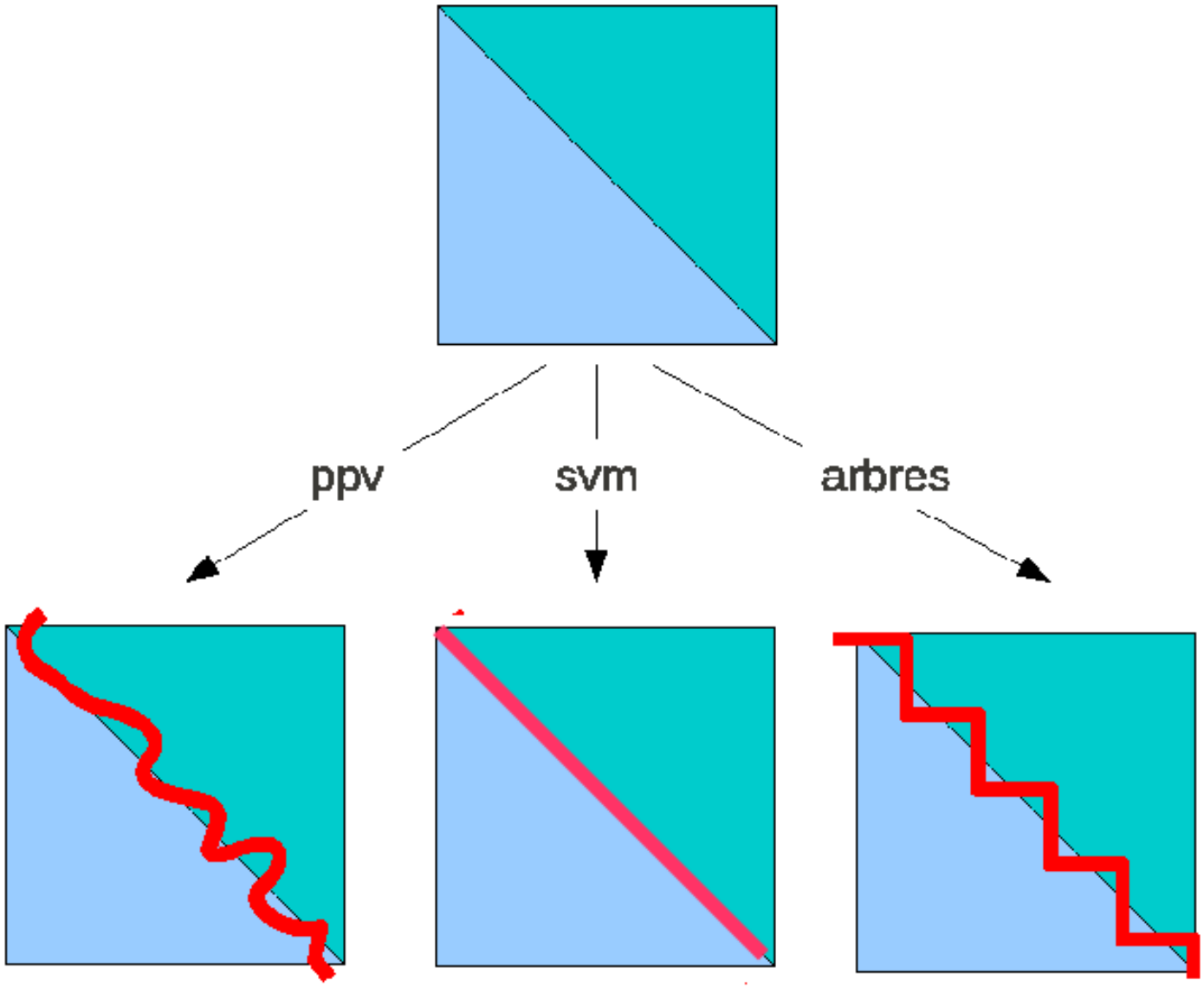
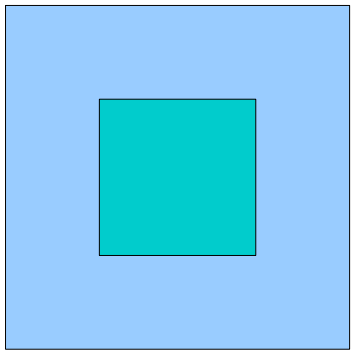


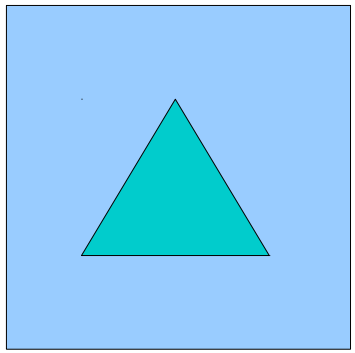
Figure:

frontières attendues



frontières attendues





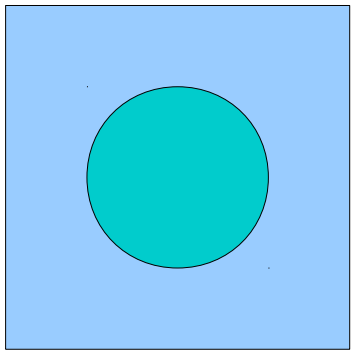


Figure:

méthodologie proposée

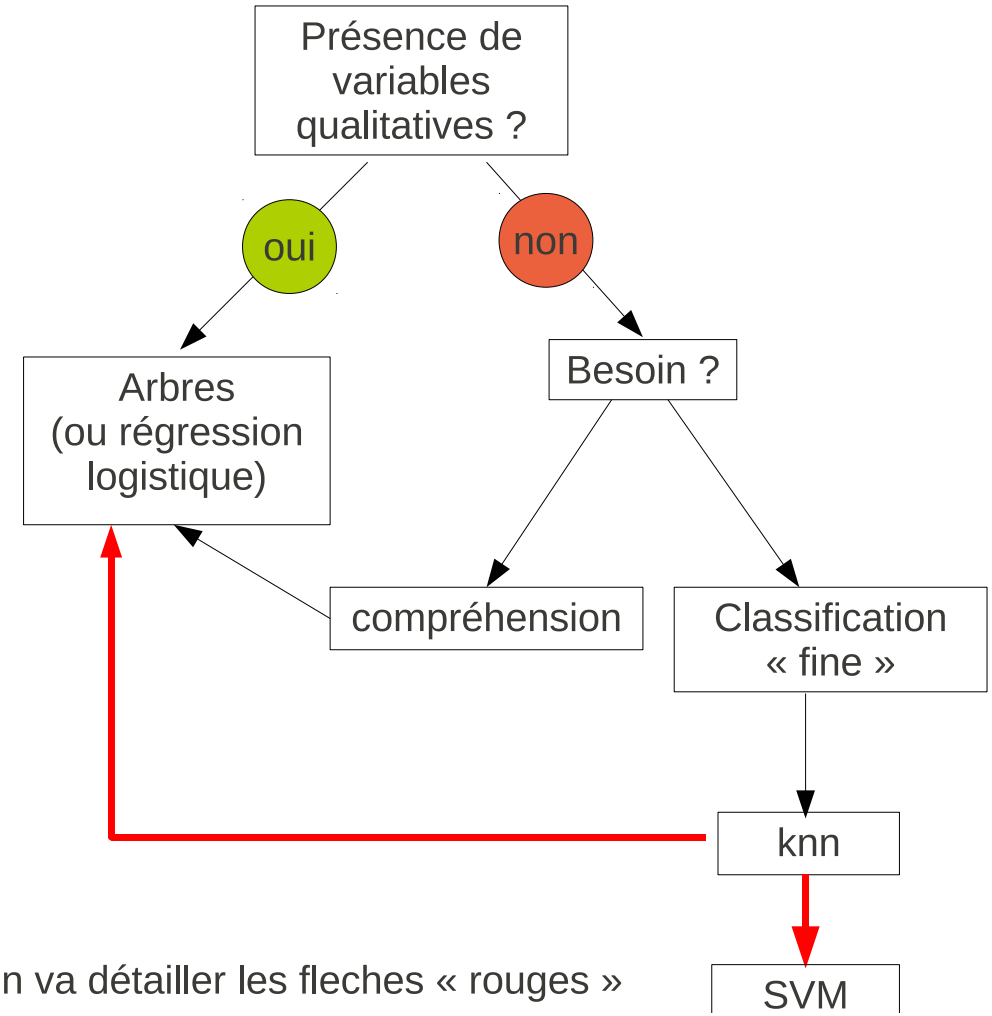


Figure:

Un conseil utiliser les ppv !!!

De par leur universalité (et leur simplicité) les ppv donnent des informations très importantes !!

1. Le taux d'erreur obtenu a l'issus des ppv
2. Possibilité de "débruiter" pour faire ensuite des SVM a marge dure
3. Attention une selection de variable préliminaire peut etre utile !!!!

Un conseil utiliser les ppv !!!

Le taux d'erreur des ppv peut etre vu comme une sorte de limite qu'on ne peut pas trop améliorer sans faire de sur apprentissage.
Une classification donnant un taux d'erreur bien inférieur aux ppv est une méthode qui "sur apprend". Ainsi si on arrive a obtenir un arbre ayant en sortie un taux d'erreur comparable a celui plus proches voisins on peut décider de le garder et de s'arreter là (pas besoin de débloquer l'artillerie des SVM).
Rappel grace a l'arbre on a une compréhension de ce qui se passe, et une regle d'affectation plus rapide a mettre en oeuvre que les ppv

Un conseil utiliser les ppv !!!

Le taux d'erreur des ppv peut etre vu comme une sorte de limite qu'on ne peut pas trop améliorer sans faire de sur apprentissage.
Une classification donnant un taux d'erreur bien inférieur aux ppv est une méthode qui "sur apprend". Si on n'a pas trouvé d'arbre on essaye les SVM.... Ainsi si on arrive a obtenir un SVM ayant en sortie un taux d'erreur comparable a celui plus proches voisins on peut décider de le garder et de s'arreter là.
Rappel grace au SVM on obtient une regle d'affectation plus rapide a mettre en oeuvre que les ppv

Un conseil utiliser les ppv !!!

Si on a suffisament d'individus on peut aussi utiliser les ppv pour "débruiter" les données et ainsi augmenter la performance des SVM en otant tous les individus i tels que : $Y_i = j$ est plus de $x$pourcent de ses plus proche voisins sont d'un label $j \neq i$.

Exemple 1: blood data

Prévoir si un individu a, ou non donné son sang, en mars 2007 les individus sont tous donneurs de sang, les variables sont :

- R (Recency - months since last donation),
- F (Frequency - total number of donation),
- M (Monetary - total blood donated in c.c.),
- T (Time - months since first donation), and

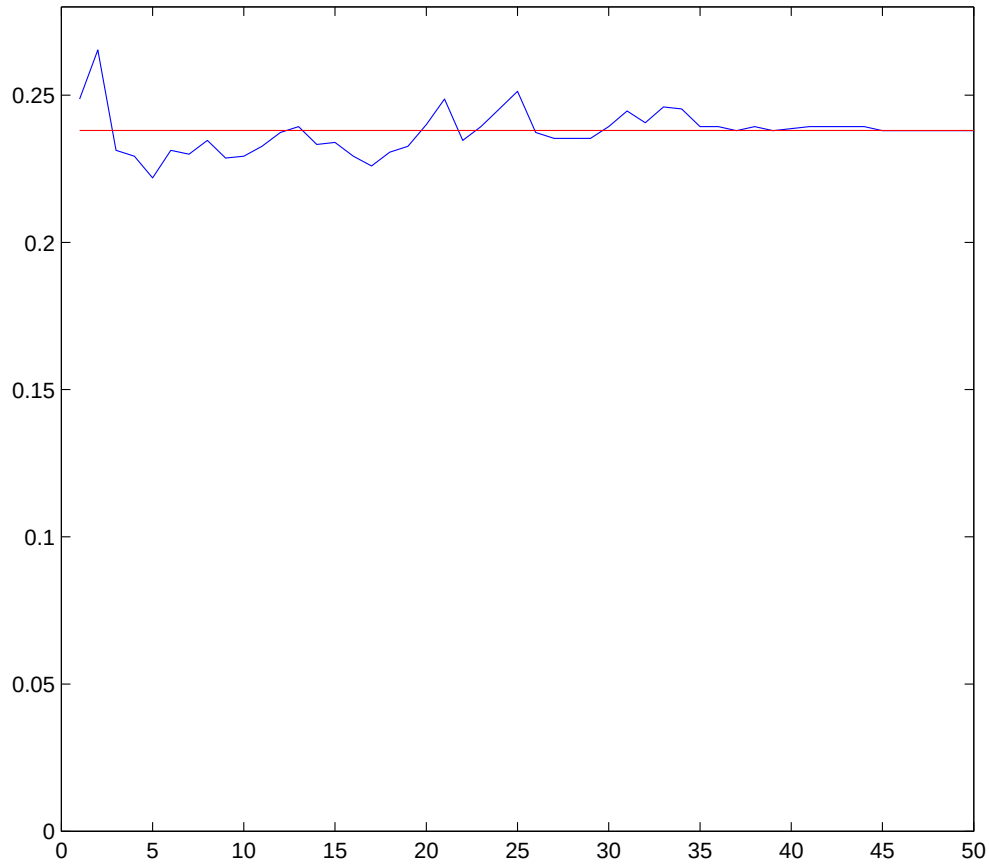
Exemple 1: blood data

dans la base d'apprentissage :

1. 748 individus
2. 178 ont donné leur sang

Exemple 1: blood data

taux d'erreur du classifieur aléatoire qui renvoie la classe "donneur 2007" avec une probabilité de 178/748: 36 pourcent. taux d'erreur du classifieur qui renvoie toujours la classe "non donneur 2007" tout le temps : $23,8$ pourcents



Exemple 2: breath cancer data

dans la base d'apprentissage 569 individus dont les tumeurs sont caractérisées par 30 variables

1. 212 tumeurs malignes
2. 357 tumeurs begnines

taux d'erreur du classifieur aléatoire : 47 pourcents taux d'erreur du classifieur tout est benin : 37 pourcents

Exemple 2: breath cancer data

résultats des k plus proches voisins :

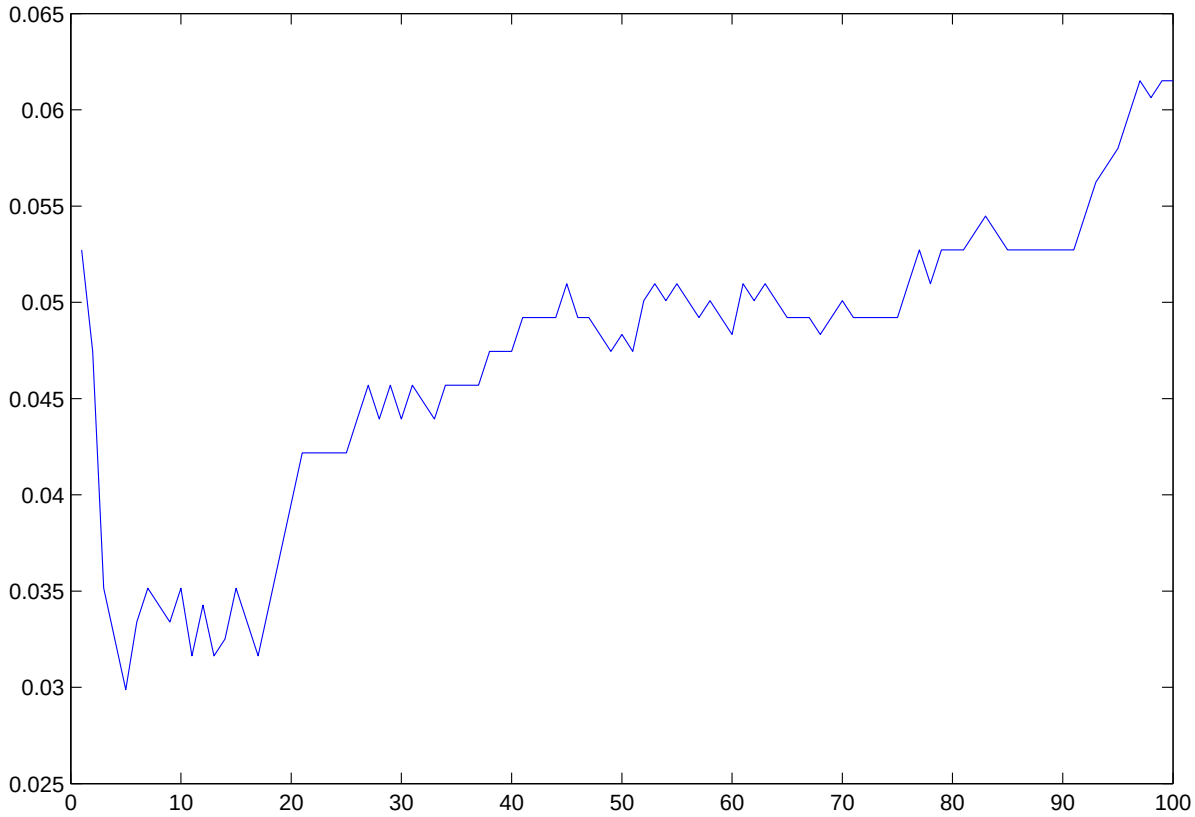


Figure: super !

Arbre

l'arbre est très simple.... si $X_1 > 0.5$ on décide tumeur maligne sinon tumer benine... le taux d'erreur est de 15 pourcents (ie bien mieux que les classifieurs naifs 37 mais bien moins bien que les ppv 3)

On obtient un taux d'erreur de 5 pourcents