

Etat de l'art : méthode de classification

DEFINITION :

Classification : Dans une classification, la variable cible est qualitative. Il s'agit de créer des catégories. Elle s'oppose en cela à la régression dont la variable cible est quantitative.

DETAIL DE L'ETAT DE L'ART

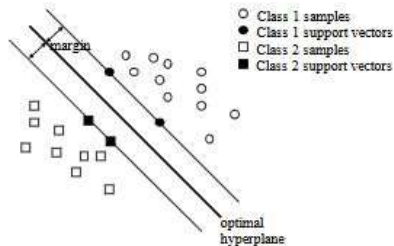
Methode de classification	Apprentissage supervisé ou non supervisé ?	Classification binomiale ou multiclassés ?																																				
CLASSIFICATION NAIVE BAYESIENNE	Supervisé	Multiclasse																																				
Principe : La méthode considère le vecteur x des valeurs des variables prédictives comme une variable aléatoire dont la distribution dépend de la classe. La classification est réalisée à partir d'un classifieur bayésien qui est soumis à l'apprentissage. Le but de l'apprentissage pour le classifieur bayésien est d'estimer la probabilité a priori des classes et d'estimer la densité de probabilités des classes.																																						
Avantages : L'efficacité et la simplicité de l'algorithme, et le peu de données nécessaires pour l'entraîner.																																						
Inconvénients : On est obligé de supposer que les variables prédictives ont des probabilités conditionnelles indépendantes.																																						
Exemple d'application : La classification naïve Bayésienne permet par exemple de déterminer si une personne est un homme ou une femme à partir de son poids et de ses mensurations à partir de données d'entraînement. Données d'entraîneemnt.																																						
<table><tr><th>Sexe</th><th>Taille (cm)</th><th>Poids (kg)</th><th>Pointure (cm)</th></tr><tr><td>masculin</td><td>182</td><td>81.6</td><td>30</td></tr><tr><td>masculin</td><td>180</td><td>86.2</td><td>28</td></tr><tr><td>masculin</td><td>170</td><td>77.1</td><td>30</td></tr><tr><td>masculin</td><td>180</td><td>74.8</td><td>25</td></tr><tr><td>féminin</td><td>152</td><td>45.4</td><td>15</td></tr><tr><td>féminin</td><td>168</td><td>68.0</td><td>20</td></tr><tr><td>féminin</td><td>165</td><td>59.0</td><td>18</td></tr><tr><td>féminin</td><td>175</td><td>68.0</td><td>23</td></tr></table>			Sexe	Taille (cm)	Poids (kg)	Pointure (cm)	masculin	182	81.6	30	masculin	180	86.2	28	masculin	170	77.1	30	masculin	180	74.8	25	féminin	152	45.4	15	féminin	168	68.0	20	féminin	165	59.0	18	féminin	175	68.0	23
Sexe	Taille (cm)	Poids (kg)	Pointure (cm)																																			
masculin	182	81.6	30																																			
masculin	180	86.2	28																																			
masculin	170	77.1	30																																			
masculin	180	74.8	25																																			
féminin	152	45.4	15																																			
féminin	168	68.0	20																																			
féminin	165	59.0	18																																			
féminin	175	68.0	23																																			
Personne à classifier :																																						
<table><tr><th>Sexe</th><th>Taille (cm)</th><th>Poids (kg)</th><th>Pointure (cm)</th></tr><tr><td>Inconnu</td><td>183</td><td>59</td><td>20</td></tr></table>			Sexe	Taille (cm)	Poids (kg)	Pointure (cm)	Inconnu	183	59	20																												
Sexe	Taille (cm)	Poids (kg)	Pointure (cm)																																			
Inconnu	183	59	20																																			
Après analyse on obtient que la postérieure <i>féminin</i> est supérieure à la postérieure <i>masculin</i> donc que l'échantillon est plus probablement de sexe féminin.																																						

LES MACHINES A VECTEURS SUPPORTS

Non supervisé

Binomial

Principe : La méthode consiste à trouver un hyperplan optimal qui sépare les deux catégories. Cet hyperplan doit avoir la distance la plus faible possible avec chacune des catégories. Dans chaque catégorie, les points les plus proches de l'hyperplan sont appelés les vecteurs supports. L'espace entre les deux catégories est appelé la marge.



Avantages : Ces algorithmes fonctionnent sur des problèmes complexes, ie non-linéaire et/ou avec beaucoup de dimension

Inconvénients : L'algorithme est souvent moins performant que les forêts aléatoires et passe difficilement à l'échelle

Exemple d'application : En médecine, la détection du cancer du sein par les machines à vecteurs supports conduit à un taux d'erreur de seulement 3%.

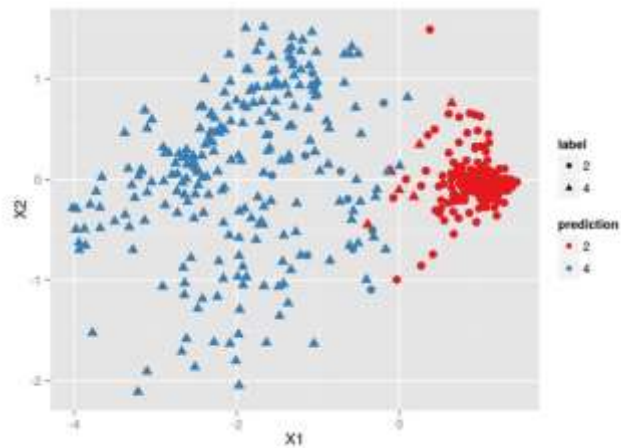
```
# We will perform basic classification on breast cancer dataset
# Using LIBSVM with linear kernel
data(svm_breast_cancer_dataset)

# We can pass either formula or explicitly X and Y
svm <- SVM(X1 ~ ., svm.breastcancer.dataset, core="libsvm", kernel="linear", C=10)

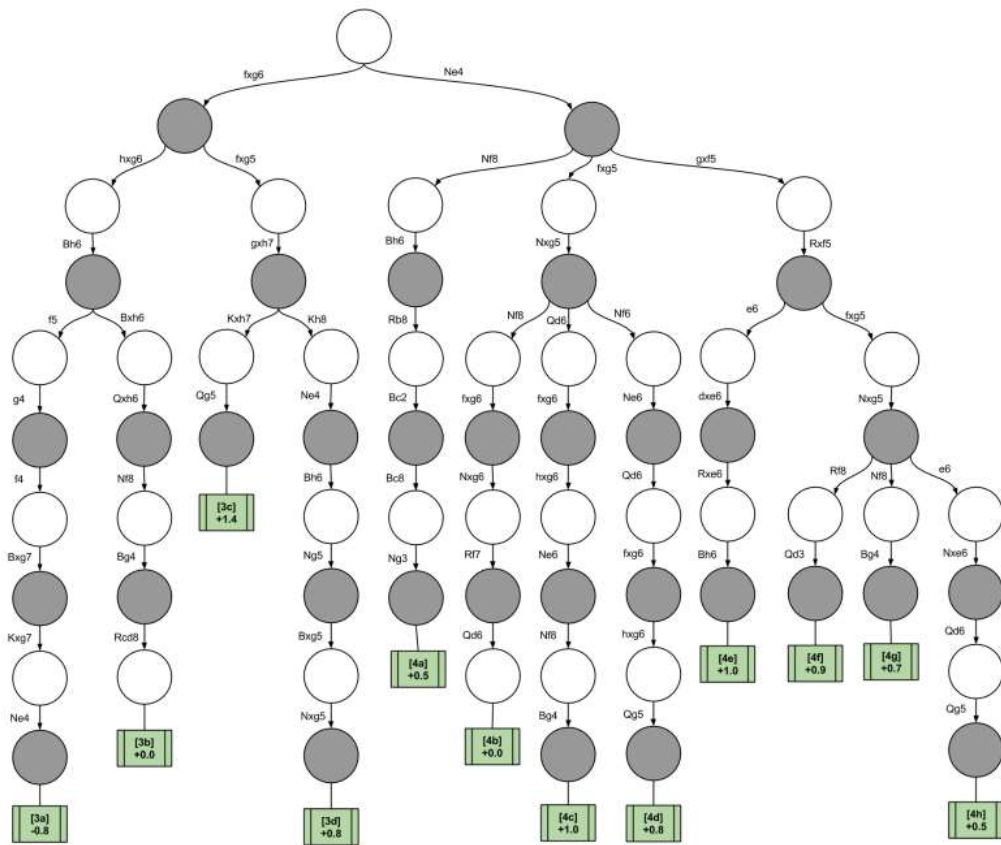
## optimization finished, #iter = 2980

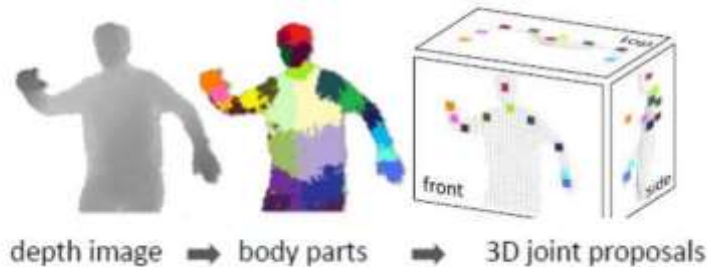
pred <- predict(svm, svm.breastcancer.dataset[,1])

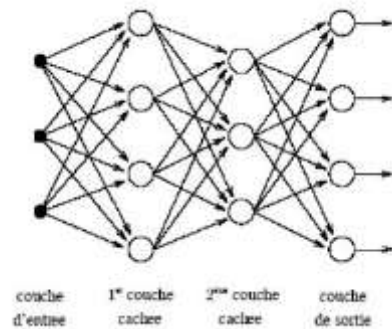
plot(svm, mode="pca")
```



LES ARBES DE DECISIONS	Supervisé	Multiclasse
<p>Principe :</p> <p>La méthode consiste à classer une observation au moyen d'une succession de tests concernant les valeurs des variables prédictives. Chaque test est représenté par un nœud de l'arbre. Chaque branche correspond à une réponse possible à la question posée. La classe de la variable est déterminée par la feuille à laquelle parvient l'observation à l'issue de la suite de tests.</p> <p>La phase d'apprentissage consiste donc à trouver les bons tests pour classer correctement les observations par rapport à leur valeur pour la variable cible. L'objectif est le suivant : les feuilles doivent être homogènes en ne contenant que les observations appartenant à une seule et même classe</p>		
<p>Avantages : Fonctionne sur des problèmes complexe (non-linéaire, multiclasse). Peu de préparation de données nécessaires.</p>		
<p>Inconvénients : Risque important de surapprentissage. Le critère du premier nœud influe énormément l'ensemble du modèle de prédiction</p>		
<p>Exemple d'application : Les arbres de décisions sont utilisés pour programmer les robots intelligents du jeu d'échec comme l'illustre l'arbre de décision ci-dessous :</p>		



LES FORÊTS ALEATOIRES	Supervisé	Multiclasse
<p>Principe : A partir d'un échantillon initial de N observations dont chacune est décrite par p variables prédictives, on crée artificiellement B nouveaux échantillons de même taille N par tirage avec remise. On entraîne alors B arbres de décisions différents. Parmi les p variables prédictives, on n'en utilise qu'un nombre $m < p$ choisies au hasard. Elles sont alors utilisées pour faire la meilleure segmentation possible. L'algorithme combine plusieurs algorithmes faibles (les B arbres de décision) pour en constituer un plus puissant en procédant par vote : pour classer une nouvelle observation, on la fait passer par les B arbres et on sélectionne la classe majoritaires parmi les B prédictions.</p> <p>Avantages : Possède les avantages des arbres de décision</p> <p>Inconvénients : Peu intelligible, complexe à comprendre et à implémenter.</p> <p>Exemple d'application : Ce sont les mêmes applications que les decisions tree. Les forêts aléatoire sont particulièrement plus performantes dans certains domaines de pointes. Par exemple, la squelettisation numérique de personne.</p> <ul style="list-style-type: none"> • Squelettisation de personne avec Kinect 		
 <p>depth image → body parts → 3D joint proposals</p>		
LES RESEAUX DE NEURONES	Non supervisé	Multiclasse
<p>Principe : Les réseaux de neurones consistent en un réseau orienté composé de neurones artificiels organisés en couches.</p>		



Les neurones d'une couche donnée sont liés à tous les neurones de la couche précédente et de la couche suivante par des relations pondérées. De ces poids dépend le comportement du réseau, et leur adaptation au problème considéré et l'objectif de la phase d'apprentissage. Chaque neurone a une sortie qui est obtenue par l'application d'une fonction non linéaire de la somme pondérée des entrées, qui sont elles mêmes les sorties des neurones de la couche précédente. Afin de calculer le vecteur poids pour chaque neurone, des algorithmes de rétropropagation ont été développés

Avantages : Les réseaux de neurones permettent de traiter des problèmes de classification non linéaires complexes.

Inconvénients : Le choix de la structure du réseau de neurone est compliqué. Il existe un risque de tomber dans un minimum local lors de l'apprentissage.

Exemple d'application : Les réseaux de neurones sont par exemple utilisés dans la bourse pour identifier les tendances. Une explication est donnée par dans le papier suivant : <https://dumas.ccsd.cnrs.fr/dumas-01064660/document>

BIBLIOGRAPHIE :

<http://r.gmum.net/samples/svm.basic.html>

<http://dSPACE.univ-tlemcen.dz/bitstream/112/4013/1/classification%20des%20tumeurs%20du%20cancer%20du%20sein%20par%20approche%20SVM>

<https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-svm-old.pdf>

http://georges.gardarin.free.fr/Surveys_DM/Survey_SVM.pdf

http://dpt-info.u-strasbg.fr/~nicolas.lachiche/CNAM_NFE212/arbresDecision.pdf

<http://www.grappa.univ-lille3.fr/polys/apprentissage/sortie004.html>

<https://www.lri.fr/~antoine/Courses/ENSTA/Tr.%20Cours%20ID3x9.pdf>

<https://kevinbinz.com/2015/02/26/decision-trees-in-chess/>

<http://www.univ-orleans.fr/log/Doc-Rech/Textes-PDF/1997-1.pdf>

https://www.math.ens.fr/enseignement/telecharger_fichier.php?fichier=821

<https://cran.r-project.org/doc/Rnews/>

<https://dumas.ccsd.cnrs.fr/dumas-01064660/document>