Data Mining 2: Modélisation statistique et apprentissage	**G. Cohen**
Exercice 05 mars 2018	
ETheme 5 Machine à vecteurs supports	*Solution*

1 Exercices

1. On considère l'ensemble de données suivant (fig. 1), qui représente la table de vérité de l'opérateur logique ET. On vous demande de calculer un classifieur SVM à marge maximum.

A	B	classe
0	0	0
0	1	0
1	0	0
1	1	1

Figure 1: Ensemble de données. ET

   (a) Dessiner les données sur un plan en dimension 2.

   (b) Essayer de deviner à quoi devrait ressembler la ligne optimale

   (c) Prouver que la ligne que vous avez devinée comme correcte, est en effet la solution optimale. Vous pouvez pour cela utiliser des propriétés géométriques très simples. Vous n'avez pas besoin de résoudre des programmes d'optimisation complexes.

Solution: En premier, il est important de visualiser le jeu de données sur le plan, comme illustré sur la Figure 2. Après quelques essais on "estime" que la marge optimale est définie par la ligne bleue. La ligne séparatrice (qui se trouve au milieu de la ligne bleue) est alors $x/1.5 + y/1.5 = 1$. Comment prouver cette optimalité?

Comme on peut le voir, on a pas besoin de la formule de la droite pour le faire (soyons malin!). Noter la ligne rouge qui croise le point rouge. Puisque le point rouge $(1, 1)$ est le seul the point étiqueté '+' du jeu de données, une des droites de marge optimale doit absoluement passer par ce point (i.e., ce point est un vecteur support).

La prochaine étape est de trouver l'autre droite de marge, qui correspond à la ligne rouge touchant un point étiqueté '-'.

Par définition, cette droite doit-être parallèle à la première droite de marge. On veut prouver que toutes paires de droite rouge telles que celles dessinées sur la figure, conduiront toujours à une plus grande marge que la bleue.

On note la distance de marge des deux lignes bleues BE, et la distance de marge des deux lignes rouges BC. On observe alors que le triangle ABC doit être un triangle rectangle. Ainsi, il vient que $AB > BC$. De plus, à moins que la droite rouge inférieure est tangente à la droite inférieure bleue , on doit avoir $BE > AB$. Et donc, il découle que $BE > BC$, CQFD.

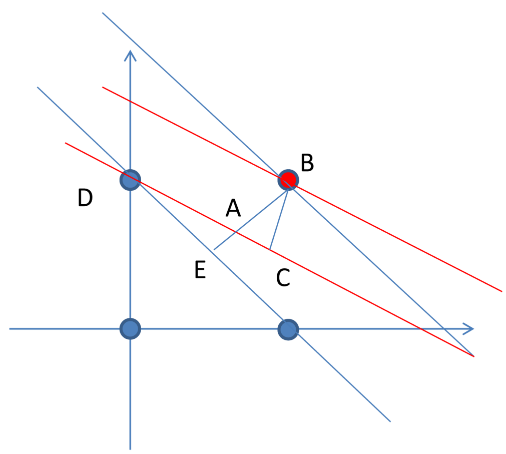

Figure 2: Visualisation de la solution

2. On considère l'ensemble de données suivant (fig. 3). Dans ce problème on utilise ces données afin d'apprendre un séparateur de la forme $f(x) = sign(w_1x_1 + w_2x_2 + w_3)$, avec $\|w\|| = 1$

   (a) Quelle est la valeur de w que l'algorithme SVM va donner avec ces données ? (Contrôler que la valeur de w retournée vérifie $\|w\| = 1$).

x_1	x_2	classe
3	-1	+
2	0	+
1	1	+
0	2	+
0	0	-
0	-4	-
-4	0	-

Figure 3: Ensemble de données.

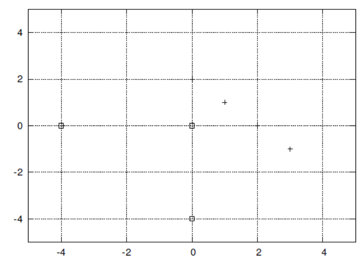

Figure 4: Les données

Solution:
On note d'après la figure que l'on a 5 vecteurs support qui sont : $sv_1 = (0,0), sv_2 = (0,2), sv_3 = (1,1), sv_4 = (2,0), sv_5 = (3,-1)$.

On recherche le vecteur directeur de la droite passant par les vecteurs supports des positifs. On prend par exemple les deux points support vecteur $sv_3 = (2,0), sv_4 = (3,-1)$. $x_{\overrightarrow{sv_3sv_4}} = x_{sv_4} - x_{sv_3} = 3 - 2 = 1$ et $y_{\overrightarrow{sv_3sv_4}} = y_{sv_4} - y_{sv_3} = -1 - 0 = -1$ donc $\overrightarrow{sv_3sv_4}$ $(1;1)$ est un vecteur directeur de (d_+). On a alors $w_1 = 1$ et $-w_2 = -1$ soit $w_2 = 1$ donc la droite admet une équation cartésienne de la forme $1 * x + 1 * y + w_3 = 0$. $sv_4 \in (d_+)$ d'où $3 - 1 + w_3 = 0$ et donc $w_3 = -2$ l'équation cartésienne est alors : $x + y - 2 = 0$.

Pour la droite (d_-) l'équation est $x + y = 0$ le biais b valant 0 et pour la droite séparatrice l'équation cartésienne est: $x+y-1 = 0$. Si $w_1x + w_2y + w_3 = 0$ est une équation de (d) alors, pour tout k non nul, $kw_1x + kw_2y + kw_3 = 0$ est aussi une équation de d on ajoute la contrainte suivante $\|w\| = 1$ d'où $w = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}\right)$

(b) Quel est l'ensemble d'erreur sur cet ensemble d'apprentissage (exprimé comme le pourcentage de points d'apprentissage mals classés)?
Solution: 0% tout est correctement classé

(c) Calculer une SVM k-validation croisée avec $k = 7$ sur les données. Expliquer votre raisonnement.
Solution: 6/7 correct. Le seul vecteur support qui change l'hyperplan est $(0, 0)$

3. Pour une SVM, si on retire un des vecteurs support de l'ensemble d'apprentissage est ce-que la marge diminue, reste la même, ou augmente?
Solution: La marge reste la même ou augmente.