Les Support Vector Machines

2020-10-07

Exercice I. Introduction aux C-SVM

Afin d'illustrer le fonctionnement des méthodes à noyaux, nous allons considérer un problème de discrimination très simple. On suppose que les données évoluent dans un espace à une dimension. Supposons que l'on dispose de 5 observations réparties en deux classes $+1$ et -1 comme suit :

$$\mathcal{S} = \{(\mathbf{x}_1 = 1, y_1 = 1), (\mathbf{x}_2 = 2, y_2 = 1), (\mathbf{x}_3 = 4, y_3 = -1), (\mathbf{x}_4 = 5, y_4 = -1), (\mathbf{x}_5 = 6, y_5 = 1)\}$$

Les commandes suivantes ont été utilisées pour générer la Figure 1 illustrative des données à analyser.

```
x = c(1, 2, 4, 5, 6)
y = c(1, 1, 2, 2, 1)

plot(x, rep(0, 5), pch = c(21, 22)[y], bg = c("red", "green3")[y],
     cex = 1.5, ylim = c(-1.7, 1), xlim = c(0, 8), ylab = "",
     xlab = "x", las = 2)

grid()

text(matrix(c(1.5, 4.3, 7, 0.5, 0.5, 0.5), 3, 2),
     c("class 1", "class -1", "class 1"),
     col = c("red", "green3", "red"))

abline(h=0) ; abline(v=c(3, 5.5))
```

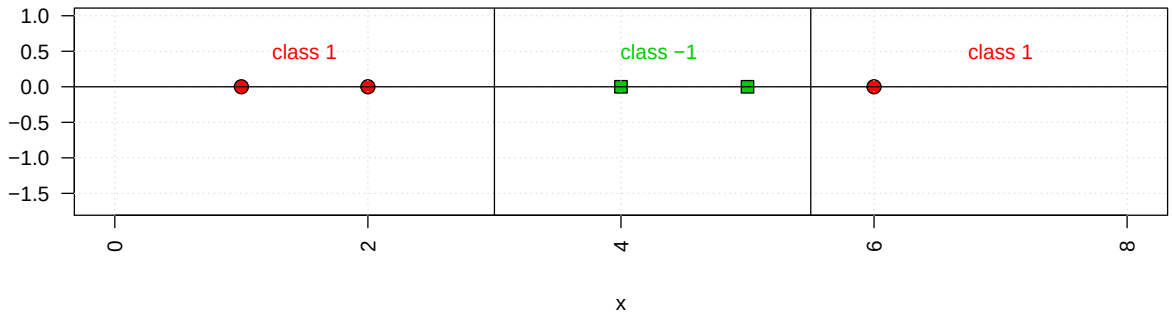

Figure 1. Visualisation des données à analyser

Il n'est évidemment pas possible de trouver une séparatrice linéaire permettant de séparer les observations labelisées 1 de celles labelisées -1. On se propose donc d'entraîner un modèle de type SVM combiné à une fonction noyau polynomial d'ordre 2 définie comme suit :

$$k(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x_1}^\top \mathbf{x}_2 + 1)^2.$$

Question 1. Ecrire la formulation duale du problème d'optimisation associé aux Support Vector Machines.

Question 2. Spécifier les arguments de la fonction `ipop` du package `kernlab` pour résoudre ce problème d'optimisation.

Question 3. En choisissant $C = 100$, montrer que le programme de résolution de problème quadratique retourne la solution $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_5)$ suivante :

$$\alpha_1 = 0, \alpha_2 = 2.5, \alpha_3 = 0, \alpha_4 = 7.333 \text{ et } \alpha_5 = 4.833$$

Question 4. En déduire que la frontière de décision correspondante est de la forme :

$$f(\mathbf{x}) = w_2\mathbf{x}^2 + w_1\mathbf{x} + w_0$$

et préciser les valeurs des paramètres de cette quadratique.

Question 5. Ajouter, à la Figure 1, la frontière de décision déterminée en question 4.

Exercice II : Support Vector Machines et validation croisée

L'exemple compagnon de cet exercice est le jeu de données «Banana »disponible sur Edunao.

Question 1. Importer et visualiser le jeu de données Banana.

Question 2. Entraîner des SVM non-linéaires combinés à un noyau gaussien[1] de paramètre $\sigma = 5$ et de paramètre de régularisation $C = 5$.

Vous pouvez utiliser la fonction `ksvm()` du package `kernlab`.

Question 3. Visualiser le modèle SVM généré à l'aide de la fonction `plot.ksvm()`. Visualiser et commenter l'effet de C et σ sur la séparatrice et le nombre de vecteurs de support.

Question 4. Tracer l'évolution du taux d'erreur par validation croisée en fonction de C et σ. En déduire les valeurs du couple optimal (C^*, σ^*).

Question 5. 'A partir de l'échantillon d'apprentissage, construire le modèle SVM associé au couple optimal (C^*, σ^*) et tester le modèle sur l'échatnillon test. Reporter le taux d'erreur obtenu sur l'échantillon de test.

[1] On rappelle que, dans `kernlab`, un noyau gaussien est défini par :

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\Big(- \sigma \|\mathbf{x}_i - \mathbf{x}_j\|^2 \Big) \tag{1}$$