Université d'Angers
Département de Mathématiques

M2 Data Science

Statistique en Grande Dimension et Apprentissage
TP "SVMs et Méthodes à noyaux"

Exercice 1

1. Un exemple simulé linéairement séparable.
   (a) On fabrique à la main deux jeux de données dans $\mathbb{R}^2$ séparés par la droite $y = x$. Par exemple, on peut simuler aléatoirement x selon une loi uniforme sur $[-3, 3]$ puis simuler b selon une loi uniforme sur $[-3, 3]$ et classer $+1$ les points $(x_i, x_i + b_i)$ qui sont au-dessus de la droite $y = x$ et classer -1 ceux qui sont en dessous de la droite. En d'autres termes, on a ici clairement un jeu de données parfaitement linéairement séparables. Fabriquez ainsi un dataframe ayant pour deux premières colonnes les abcisses et ordonnées des points et pour 3ème colonne la classe de chacun des points.
   (b) Entraîner un prédicteur `svm.svc` sur un échantillon d'entraînement de taille 200 individus équilibré avec le noyau linéaire.
   (c) Explorez les différents paramètres associés au module `svc`.
   (d) Affichez les vecteurs supports et leurs indices, les coefficients associés et β_0^*. Quelle est l'équation de l'hyperplan séparateur ?
   (e) Fabriquez un graphe représentant l'hyperplan séparateur, les données d'entraînement et les différentes zones associées au problème.
   (f) Quel rôle joue le paramètre C ? Comment faire pour retrouver un SVM à marge non flexible ? Faites varier les valeurs de C et regardez l'effet sur les vecteurs supports.
   (g) Calculez l'erreur d'entraînement et l'erreur test sur un échantillon de taille 100 (équilibré).
   (h) Affichez les scores associés à votre échantillon test.
   (i) Que se passe-t-il si l'on enlève l'option sur le noyau ? Quel est le noyau par défaut ? Testez la méthode avec ce nouveau noyau.
2. Un exemple non linéairement séparable.
   (a) Importez le jeu de données `make_moon`.
   (b) Entraînez un prédicteur à noyau polynomial sur ces données (après avoir séparé votre échantillon en parties train/test.
   (c) Affichez le graphe des régions de décision.
   (d) Faites une recherche sur grille pour entraîner le meilleur modèle sur cet exemple.
   (e) Dans les questions qui suivent, on suppose que $p = 2$.

Exercice 2 (Un exemple multi-classes)
Testez le modèle svm sur la base de données `iris` (N.B. Pour l'affichage, on peut utiliser le module `DecisionBoundaryDisplay`).

Exercice 3 (SVM et régression)

Comme mentionné dans la partie cours, les SVM peuvent aussi permettre d'aborder des problèmes de régression. On propose dans cet exercice de tester cette méthode sur le jeu de données `ozone` accessible sur Moodle. L'objectif, sur ces données, est d'améliorer la prévision calculée par les services de MétéoFrance (MOCAGE) de la concentration d'ozone dans certaines stations de prélèvement à partir de cette prévision et en s'aidant d'autre variables également prévues par MétéoFrance. Il s'agit d'un problème dit d'adaptation statistique d'une prévision déterministe pour une amélioration locale de modèles à trop grande échelle. Plus précisément, deux variables peuvent être prédites : soit la concentration quantitative d'ozone, soit le dépassement (qualitatif) d'un certain seuil fixé à 150 μg. On s'intéressera en priorité au premier problème.

Exercice 4 (Un second exemple multi-classes)

On revient ici sur la base `MNIST` (digits numbers). A nouveau, on a un modèle multi-classes (qui sert également de base de données usuelle pour les tests d'algorithmes). Mettre en oeuvre la méthode `svm`. Comment fonctionne l'algorithme ? Peut-on ajuster les paramètres ? *N.B. Attention, la base `MNIST` contient un nombre de points importants. Le temps de calcul risque d'exploser avec n. Testez dans un premier temps des petites valeurs de n (en tirant au hasard un échantillon de taille n).*

Les SVM sont applicables et appliqués à de nombreux problèmes. Voici en complément quelques références :

1. Reconnaissance faciale par SVM (et ACP)
2. *20NewsGroups* et SVM : *20NewsGroups* fait partie des bases "textuelles" célèbres (collection d'environ 20000 articles de journaux). Le but est alors de les classer de manière automatique par thème (voir par exemple ce lien).
   pour un développement sur le sujet.