TP: SVM et noyaux

5 mai 2008

- *Avant toute chose, récupérer le code et les données :* `wget http ://cbio.ensmp.fr/ jacob/teaching/ensmp/tpsvm.tgz` *dans un terminal.*
- *Ouvrir l'archive :* `tar xvzf tpsvm.tgz`.
- *Aller dans le répertoire tp :* `cd tpsvm`.
- *Ouvrir le code source :* `gedit tpsvm.r &`.
- *Lancer R :* `R`.
- *Il est possible de copier/coller d'une fenêtre à l'autre en sélectionnant du texte puis en utilisant le bouton du milieu.*

1 Exercice 1 : SVM linéaires (comprendre les SVM)

Le but de ce premier exercice est de mieux appréhender le principe et le fonctionnement de l'algorithme SVM.

- Rappeler briévement l'idée de la SVM (objectif, principe géométrique).
- Charger les données `datalin` et la bibliothèque `kernlab`.
- Entrainer et visualiser la SVM linéaire pour plusieurs valeurs de C.
- On rappelle que la SVM linéaire dans $\mathbb{R}^d$ s'écrit comme un problème d'optimisation :

$$\min_{w \in \mathbb{R}^d} C \sum_{i=1}^{n} L(y_i, w.x_i + b) + \|w\|^2 \tag{1}$$

  et que la solution de l'optimisation est de la forme

$$w = \sum_{i=1}^{n} \alpha_i x_i$$

  avec $\alpha_i \neq 0$ si et seulement si x_i est un vecteur de support.
  Interpréter les deux termes de la fonctionnelle. Expliquer (et vérifier expérimentalement) l'effet de C sur la position de la séparatrice et sur le nombre de vecteurs de support. Que représentent les lignes pointillées ?
- Que se passerait-il si les données étaient plus mélangées, par exemple générées par deux gaussiennes de centres plus proches et/ou de plus grande variance ?
- Visualiser les points de test avec le classifieur et calculer les prédictions du classifieur entrainé sur les points de test. Comment est calculée la prédiction sur un point inconnu x étant donné un w^* solution de (1) ? Et étant donnée

la solution en α (certains algorithmes optimisent le problème en w, d'autres en α) ?

- Charger les données `dataC` (données de spam). Entrainer une SVM avec plusieurs valeurs de C et tracer l'erreur de classification sur les données d'entraînement et de test en fonction de C. Expliquer le comportement de ces deux erreurs.
- Soit un problème de classification pour lequel un certain nombre de points d'entraînement est donné. Comment utiliser ces points pour estimer C minimisant l'erreur sur des données inconnues ? Comment utiliser ces points pour estimer l'erreur qui sera faite par le classifieur sur des données inconnues ?

2 Exercice 2 : SVM à noyaux (comprendre les noyaux)

Le but de cet exercice est de mieux appréhender le concept de noyau et son utilisation dans un algorithme d'apprentissage.

- Rappeler briévement la définition d'un noyau d.p. Dans quel cas et de quelle manière peut-on s'en servir pour travailler implicitement dans un autre espace de descripteurs ? Quel en est l'intérêt ?
- Charger et afficher les données `datamix`.
- Entrainer et visualiser une SVM linéaire sur les données. Qu'observe-t-on ?
- Utiliser maintenant un noyau rbf gaussien. On rappelle que ce noyau est défini par :

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

  Qu'observe-t-on ?
- On rappelle que de manière générale, l'estimateur donné par la SVM à noyaux a la forme

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b$$

  avec $\alpha_i \neq 0$ si et seulement si x_i est un vecteur de support.
  Expliquer comment est construite la séparatrice en termes de noyau et (de manière équivalente) d'espace des descripteurs.
- Décrire l'effet de C et σ sur la séparatrice et les vecteurs de support.

3 Exercice 3 : application à un problème réel

Étant donné un ensemble de cibles thérapeutiques, et pour chaque cible un ensemble de molécules dont on sait si elles se lient ou pas au site actif de la cible :

- Donner deux manières de construire une fonction prédisant si une molécule donnée se lie à une des cibles.
- Décrire la marche à suivre pratique pour appliquer une de ces méthodes (penser à la description des données, au choix des paramètres...).
- Comment prédire si une nouvelle molécule inconnue se lie à une nouvelle cible inconnue ?