ECP, SDM 2014-2015
M. Mougeot, K. Scaman, G. Fäy, N. Vayatis

**SVM et modèles de mélanges**

# Application I : classification

Le but de cette application est de construire un modèle prédictif permettant de classer automatiquement si un courrier électronique est un spam ou non (email). Les données sont constituées de plus de 4500 observations. Chaque observation est décrite par 57 attributs.

## Analyse préliminaire

- Les données sont décrites dans le fichier "spaminfo.txt". Etudier rapidement ce fichier.
- Charger les données SPAM dans l'environnement de travail R à l'aide de l'instruction

  tab=read.table('spam.txt',header=T,sep=';");
- Indiquer le nombre d'observations, le nombre de variables et la nature des variables. Caractériser à l'aide du fichier "names" les variables analysées dans les courriers électroniques ?
- Exécuter l'instruction *head(tab)*. Dans quelle colonne se situe la variable cible ?
- Récupérer dans une variable notée $Y$ les valeurs des lables. Exécuter unes à unes les instructions suivantes et noter le résultat : *levels(Y)*; *nlevels(Y)*; *table(Y)*; *plot(Y)*;

  Dans ce fichier, calculer la proportion d'observations étiquetées spam ? mail ?

  **Fonctions R**: nrow(), ncol(), dim(), names(), table(), help(nomfonction)

## Classification par SVM

- Créer une liste contenant les indices de 75% des observations choisies aléatoirement. Puis, créer deux **matrices**, Xtrain et Ytrain, contenant les valeurs des covariables et de la variable cible pour les observations sélectionnées aléatoirement.
- Calibrer un C-SVM à noyau gaussien sur les données *tabtrain* à l'aide de l'instruction *classif=ksvm(....)*.
- Calculer sur la base d'apprentissage (données utilisées pour la calibration du modèle) l'erreur et la matrice de confusion. Quel est le pourcentage de bonnes détections ? de mauvaises détections ? de données spam prédites comme email ? de données email prédites comme spam ?
- Calculer les mêmes caractéristiques sur la base de test (définies par les 25% restants des observations)?
- On se propose de mesurer empiriquement l'impact du choix aléatoire de la base d'apprentissage, sur la base de test.

  Répeter $K = 20$ fois les opérations suivantes : (1: choix aléatoire de la base; 2: calibration du modèle sur la base d'apprentissage; 3: évaluation de l'erreur sur la base de test).

  Calculer les valeurs minimale, maximale, et moyenne observées sur les différentes bases de test. Calculer l'histogramme des erreurs, en choisissant un nombre opportun de classes pour la visualisation. Visualiser la boite à moustaches des erreurs. Que peut-on conclure ?
- Mener une étude empirique permettant de comparer les performances de trois noyaux différents. Afficher les distributions des erreurs sur la base de test pour les trois noyaux étudiés.

**Fonctions R :** sample(); hist(); table(); plot(); boxplot();
**library kernlab :** library(kernlab), ksvm(), help(ksvm), predict(), table(), as.matrix(), as.factor().

## Application II : régression

- Etudier les fichiers "USCrimeinfo.txt" et "UsCrime.txt". La variable cible ($Y$) est la première variable colonne du fichier.
- Charger le fichier dans l'environnement R en utilisant la fonction `tab=read.table()`. Quel est le nombre d'observations disponibles ? Visualiser les nuages de points entre les variables. Que constate-t-on ? Calculer la matrice de corrélations. Interpréter le résultat.

### Modèles de regression par SVM

Etudier le modèle de régression permettant de prédire le taux de criminalité à l'aide d'un modèle SVM. Quel noyau conseilleriez-vous ? Quelles sont les performances observées ?

## Application III : mélange de classifieurs, algorithme EM

L'algorithme EM permet de réaliser une approche de type maximum de vraisemblance sur des modèles de mélanges. Nous allons ici l'appliquer à la segmentation d'images, puis à la régression lorsque plusieurs comportements linéaires sont présents dans un jeu de données.

### Mélange de gaussiennes

- Charger le fichier `irm_thorax.txt` dans l'environnement R à l'aide de la fonction `irm=as.matrix(read.table("irm_thorax.txt",header=F,sep=';'))`.
- Afficher l'image à l'aide de la fonction `image()`, puis l'histogramme des couleurs. Qu'observez-vous ?
- Implémenter l'algorithme Expectation-Maximization sur ce jeu de données afin de réaliser une segmentation de l'histogramme des couleurs (cf. cours pour les formules d'itération des paramètres).
- Afficher le résultat de la classification pour 2, 3, puis 5 gaussiennes. La segmentation vous semble-t-elle pertinente ?

### Mixture de régressions par l'algorithme EM

- Télécharger, puis charger la librairie `mixtools`.
- Charger le fichier `regression_double.txt`. Afficher les données. Qu'observez-vous ? Quels raisons peuvent expliquer ce type de comportement ? Une régression linéaire simple vous semble-t-elle adaptée ?
- Réaliser une mixture de deux régressions linéaires à l'aide de la fonction `regmixEM`.
- Afficher le résultat de la mixture de régressions, et calculer les résidus.
- Réaliser une mixture de deux régressions linéaires en limitant le nombre d'itérations à 1, 3 puis 5 (`regmixEM(y, x, maxit = k)`). Afficher l'erreur de prédiction en fonction du nombre d'itérations et visualiser les classes calculées dans chaque cas. Qu'observez-vous ?