Model-based learning

Projet sous le logiciel R - 2021-2022

L'objectif du projet est de créer un package R permettant de réaliser une classification par modèle de mélange.

L'algorithme Le modèle considéré sera un modèle de mélange où chaque classe pourra être modélisée par un mélange de gaussiennes. Au sein de chaque classe, l'algorithme EM sera utilisé pour l'inférence du modèle, et le critère BIC sera utilisé pour sélectionner le nombre optimal de composantes. Chaque composante du mélange sera soit un modèle gaussien complet, soit un modèle plus parcimonieux, que vous choisirez vous-même parmi les modèles vus en cours.
Votre fonction de classification devra prendre en entrée :

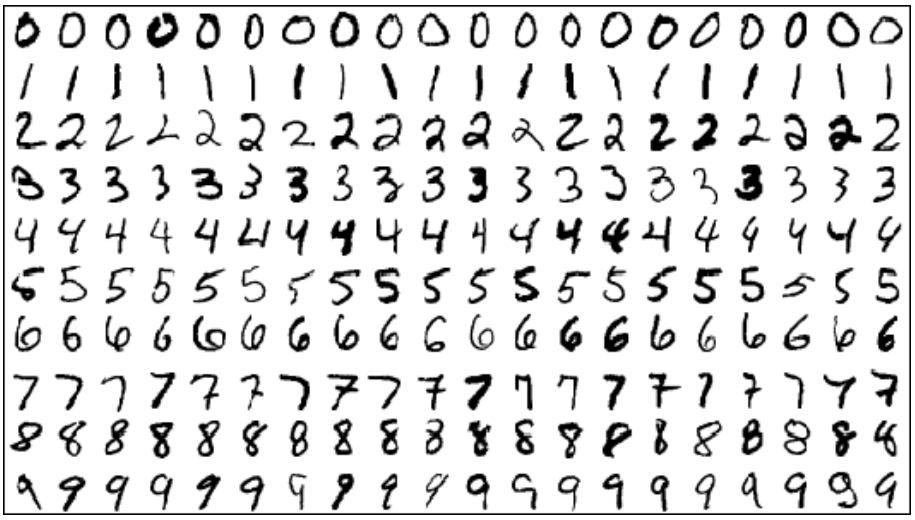
- les variables descriptives des individus de l'échantillon d'apprentissage,
- la variable réponse à prédire pour les individus de l'échantillon d'apprentissage,
- les variables descriptives des individus de l'échantillon test,
- le nombre maximal de composantes par classe,

En sortie, l'algorithme devra retourner :

- la variable réponse à prédire pour les individus de l'échantillon d'apprentissage,
- le modèle choisi, et notamment le nombre de composantes par classe,
- les paramètres du modèle estimé.

Le package Votre fonction R devra être incorporée à un package R.
Comme tout package R il devra comporter une aide pour votre fonction, incluant un exemple d'utilisation (sur les données iris).
Pour la création du package R, vous pourrez suivre ce document pour savoir comment procéder :
http://math.agrocampus-ouest.fr/infoglueDeliverLive/digitalAssets/20596_faire_pkg_R.pdf

Applications Outre le test de votre modèle sur les données iris qui sera présent dans l'aide de votre fonction, vous devrez tester votre modèle sur les données MNIST (http://yann.lecun.com/exdb/mnist/), et vous comparer à quelques méthodes classiques de classification (SVM, Random Forest, ...). Vous pourrez travailler sur un sous-échantillon de taille réduite de ces données pour réduire les temps de calculs.



Livrable

- un package R (en format `.tar.gz` que je puisse installer sans soucis sur mon Mac),
- un mini rapport en format `pdf` décrivant la comparaison de votre modèle avec les méthodes classiques de classification sur les données MNIST.