SVM et classification multi-classe

La base de données étudiée pour ce TP est un ensemble d'informations collectées sur les élèves de différentes écoles : l'objectif est de prédire le niveau des élèves (parmi trois classes, "High","Medium","Low"), à partir de facteurs extérieurs. Une description précise des différents variables est disponible ici : `https://www.kaggle.com/aljarah/xAPI-Edu-Data`

1 Importation des données

Importer la base de données, la séparer en une base d'apprentissage et une base de test.

Combien de variables qualitatives et quantitatives sont contenues dans la base de données ? A l'aide de la fonction `pairs()` visualiser sur l'ensemble d'apprentissage les nuages de points croisés des variables quantitatives, en appliquant une coloration qui dépend de la classe. Les variables quantitatives vous paraissent-elles suffisantes pour prédire le niveau des élèves ?

2 Support Vector Machine

La fonction `svm` du package `e1071` permet de créer un objet de type "svm" doté d'une méthode `predict` (comme c'était le cas pour des objets construits par la fonction `glm` pour la régression logistique).

```
library(e1071)
svm.fit <- svm(Class ~ ., data=train, kernel="radial",gamma
    =0.5,cost=1)
```

A l'aide de la fonction `predict`, calculer l'erreur de test (et afficher la table de confusion associée). Essayer différents choix de noyaux et différents paramètres associés (voir dans l'aide). Quel noyau semble le plus adapté pour ce problème ?

3 Ajustement des paramètres

Les paramètres du noyau et le paramètre de coût C doivent être optimisé à l'aide d'un ensemble de validation ou par validation croisée. Plutôt que de faire cette optimisation "à la main" comme lors des TP précédents, vous apprendrez à vous servir de la fonction `tune` (ici `tune.svm()`) du package `e1071`. Ci-dessous un exemple d'utilisation.

```
obj <- tune.svm(Class ~ ., data = train, kernel = "
    polynomial", degree = 1:10, gamma=seq(0.1,2,0.1),cost =
    2^(0:4),tunecontrol=tune.control(sampling="fix",fix=2/3))
summary(obj)
Best <- obj$best.parameters
svm.optimized <- svm(Class ~ ., data=train,kernel="polynomial
    ",degree=Best$degree,gamma=Best$gamma,cost=Best$cost)
```

Que font les paramètre actuels ? Ajuster les paramètres de `tune.control()` de sorte à effectuer une validation croisée V-fold avec $V = 8$. Il est également possible de spécifier un ensemble de validation choisi soi-même. Regarder dans l'aide et le faire. `https://cran.r-project.org/web/packages/e1071/e1071.pdf`.

Evaluer les performances du classifier "optimisé" obtenu.

4 Avec les variables quantitatives

On se propose d'essayer de prédire les résultats des élèves à l'aide des variables quantitatives uniquement. Parmi les méthodes vues en classes, lesquelles peuvent être utilisées ? On cherchera en particulier le meilleur classifieur des k-plus proches voisins. On pourra se servir de la fonction `tune.knn()` pour le choix de k.

Que dire de la qualité du classifieur obtenu ?