Projet
à envoyer pour le 16 mai 2022 minuit au plus tard

Consignes

Le projet donne lieu à un compte-rendu *rédigé* à effectuer en *binôme*.

- Ne pas oublier de définir un titre, une introduction pour préciser la problématique étudiée, le plan du travail et une conclusion.
- Le document rédigé doit inclure dans le corps du texte les figures qui vous semblent importantes et les résultats nécessaires.
- Tout résultat doit être *justifié* et toute figure *commentée*. La notation prendra en compte la clarté et le soin de la rédaction.
- Deux fichiers sont à télécharger sur l'activité devoir de e-campus: un `pdf` du compte-rendu (qui peut être manuscrit puis photographié ou scanné) et un fichier texte `.R` contenant les commandes. Les fichiers seront nommés avec les noms du binôme: `NOM1-NOM2.pdf` et `NOM1-NOM2.R`.

  Le pdf ne doit pas comporter de commandes R.
- Aucun retard ne sera admis.

Introduction

Le jeu de données `Raisin.xlsx` est fourni par Cinar, Koklu et Tasdemir [1]. Dans cette étude, un système de vision artificielle a été développé afin de distinguer deux variétés différentes de raisins secs (Kecimen et Besni) cultivés en Turquie. 900 grains de raisins secs ont été photographiés, chaque variété à part égale. Ces images ont été soumises à diverses étapes de prétraitement et 7 opérations d'extraction de caractéristiques morphologiques ont été effectuées à l'aide de techniques de traitement d'images. Plusieurs modèles de machine learning ont été testés. Les auteurs affirment que la performance la plus élevée parmi les méthodes qu'ils ont testées a été obtenue avec la méthode SVM.

Il s'agit d'étudier ce jeu de données et d'essayer d'en reproduire les résultats.

Partie I: Analyse non supervisée

Commencer par procéder à une étude non supervisée, en traitant le jeu de données dans son ensemble.

[1] CINAR I., KOKLU M. and TASDEMIR S., (2020), *Classification of Raisin Grains Using Machine Vision and Artificial Intelligence Methods*, Gazi Journal of Engineering Sciences, vol. 6, no. 3, pp. 200-209, December, 2020.

1. Analyses uni- et bi-variée à votre convenance.

   Représenter en particulier le scatter plot avec des attributs de points dépendant du type de raisin.

2. ACP

3. Clustering non supervisé par K-moyennes et par classification hiérachique ascendante.

   Quel nombre de groupes retiendriez-vous ?

   On considère deux groupes: quelle est l'erreur de classification?

   La normalisation des variables a-t-elle une influence sur ces résultats?

4. Pour $k = 1, \ldots, 7$, effectuer un clustering sur les k premières composantes principales. Quel nombre d'axes amène à une erreur de classification la plus faible? Ce résultat est-il sans biais?

Partie II: Des méthodes vues en cours

1. Définir le modèle logistique. Discuter d'un point de vue théorique et pratique les impacts du centrage et de la réduction des variables explicatives.

2. Définir l'échantillon d'apprentissage de la façon suivante:

   ```
   set.seed(1)
   train =  sample(c(TRUE,FALSE),n,rep=TRUE,prob=c(2/3,1/3))
   ```

   puis estimer les modèles suivants de régression logistique:

   - modèle complet, comprenant toutes les composantes
   - en prenant en compte uniquement les deux premières composantes principales
   - obtenu par sélection de variables avec critère AIC
   - obtenu par régression pénalisée lasso.

   Vous décrirez comment choisir les hyper-paramètres quand ces méthodes en ont.

3. Considérer maintenant un modèle SVM linéaire, puis avec un noyau polynomial.

4. Tracer les courbes ROC calculées sur l'échantillon d'apprentissage et sur l'échantillon de test pour le modèle complet; superposer la courbe de la règle aléatoire.

   Superposer les courbes ROC de tous les autres modèles calculées sur l'échantillon de test.

   Calculer l'aire sous la courbe ROC pour chacun des modèles et l'afficher dans la légende.

5. Pour chaque modèle défini, calculer l'erreur sur l'échantillon d'apprentissage et sur l'échantillon de test.

   Motiver le choix du modèle que vous proposez de retenir.

Partie III: Analyse discriminante

L'analyse discriminante fait partie des méthodes de classification supervisée, voir poly p.112 (chap. 9.4). On l'étudie d'abord à partir des *deux premières composantes principales du jeu de données d'apprentissage uniquement*, pour faciliter les représentations. L'étude est élargie à toutes les variables à la fin de cette partie.

1. Calculer les projections des observations du jeu de test sur le premier plan principal associé au jeu d'apprentissage.

   Les retrouver en utilisant la commande `PCA` de façon pertinente.

2. Analyse discriminante linéaire

   (a) Rappeler la définition du modèle.

   (b) Donner l'expression mathématique des coefficients de la droite frontière de décision, puis les calculer avec le logiciel.

   (c) Pour traiter le point précédent, vous avez sans doute inversé la matrice Σ de la variance commune.

   Il n'est pas nécessaire de passer par cette étape qui est numériquement coûteuse. Montrer en effet que la frontière peut s'écrire en fonction des valeurs propres et vecteurs propres de Σ. Vérifier votre calcul théorique en l'appliquant à l'exemple numérique.

   (d) Représenter les points du jeu d'apprentissage sur le premier plan principal, superposer la frontière de décision.

3. Calculer l'erreur de classification sur le jeu de test directement à partir de la question précédente.

   Comparer en utilisant successivement la fonction `lda` du package `MASS`, puis la fonction `predict`.

   Montrer qu'il est possible de définir une courbe ROC associée à cette règle de classification et la tracer sur la figure de la question II-4.

   L'analyse discriminante quadratique peut-elle apporter une amélioration?

4. On considère maintenant toutes les variables initiales du jeu de données. Proposer un modèle d'analyse discriminante. Est-il meilleur que ceux de la Partie II?

Bonus

Data challenge: Tester d'autres modèles (vus en cours ou non) qui pourraient battre les performances de tous ceux déjà étudiés dans ce projet.