SPAD Niveau 2

TD

Exercice 1 : Apprentissage non supervisé – ACP (Analyse en Composantes Principales

1. Créer un nouveau projet de nom « SPAD niveau 2 ».
2. Importer la base de données « Eaux.sda » fournies par SPAD (C:\Program Files\SPAD9\data\) avec le composant d'importation approprié. L'extension « .sda » se traduit par « SPAD Data Archive »
3. Réaliser une Analyse en Composantes Principales (ACP) sur les données (ACP – Analyse en Composantes Principales).
   Paramétrer l'ACP en choisissant les variables actives (7 à 22) et illustratives (2 à 6) pour l'analyse.
   a. Afficher le rapport Excel :
      i. Identifier s'il y a des variables corrélées significativement entre elles (t de student > 1.96 ou t de student < -1.96). Si c'est le cas, identifier les couples de variables les plus corrélées entre elles (3 premiers ?).
      ii. Combien d'axes retenez-vous pour l'analyse (critère du coude, critère de Kaiser…) ?
      iii. Quel est le pourcentage de variance expliquée (d'informations) sur ces axes ?
   b. Générer les graphiques d'analyse factorielle :
      i. Afficher les variables illustratives sur le graphique des variables (si vous en avez choisies).
      ii. Créer les libellés sur le graphique des variables
      iii. Identifier graphiquement s'il existe des groupes de variables corrélées plus significativement entre elles. Si c'est le cas, colorier chaque groupe d'une couleur différente (habillage par groupe de modalités).
      iv. Créer un nouveau plan factoriel en croisant des axes différents que les axes 1 et 2 choisis par défaut (onglet graphique).
4. Pour approfondir l'analyse, utiliser le composant « Description des axes factoriels ».
   Paramétrer le composant de sorte à afficher les résultats pour l'ensemble des variables et l'ensemble des individus sur les axes 1 et 2.
   a. Quelles sont les variables qui contribuent le plus aux deux premiers axes factoriels ?
   b. Quels sont les individus qui contribuent le plus aux deux premiers axes factoriels ? Quel peut être le problème posé par ces individus au profil très atypique ? Mettez-les en valeurs sur le graphique du premier plan factoriel précédemment créé.

Exercice 2 : Apprentissage non supervisé – Clustering

1. A partir de l'analyse factorielle, réaliser une classification automatique des individus (Classification (CAH, K-means, Mixte))
   Paramétrer la classification selon la méthode qui vous semble appropriée (ne retenez que les deux premiers axes).
2. Afficher le dendrogramme. Quelles sont les partitions proposées et quel est le nombre de classe dans chaque partition ?
3. Visualiser les graphiques d'analyse factorielle:
   Afficher alternativement les différentes partitions sur le graphique des individus. A partir de l'analyse graphique du dendrogramme et des graphiques de l'analyse factorielle, choisissez la partition qui vous semble la plus intéressante[1].
4. Suite au choix de la partition, paramétrer de nouveau le composant de la classification en sélectionnant manuellement le nombre de classes voulu.
5. Pour approfondir l'analyse, utiliser le composant « Caractérisation des classes de la typologie ».
   a. Quelles sont les variables les plus caractéristiques de chaque classe en termes de surreprésentation (V-test positif) et de sous-représentation (V-test négatif) ?
   b. Afficher les sorties graphiques
6. Ajouter dans la base de données une colonne contenant le nom de la classe auquel appartient chaque individu (Archivage (prédictions, axes, partitions)).

Exercice 3 : Apprentissage supervisé – Arbre de décision

1. Importer la base de données « Attrition assurance auto.sda » fournies par SPAD.
2. Réaliser un arbre de décision sur les données (Arbres de décision).
   a. Paramétrer l'arbre de décision en choisissant la variable « Attrition Client » comme variable à expliquer et l'ensemble des variables qualitatives comme variables explicatives.
   b. Vérifier que la population a été divisée en un échantillon d'apprentissage (70%) et un échantillon test (30%).
3. Afficher le rapport Excel de l'arbre de décision
   a. Quelle est la variable la plus discriminante sur l'ensemble de l'arbre ?
   b. Quel est le nombre de règles de l'arbre ?
   c. Quelle est la sensibilité du modèle portant sur l'échantillon test ?
   d. Quelle est la précision du modèle portant sur l'échantillon test ?
4. Visualiser l'arbre interactif
   a. Citer plusieurs règles de décision grâce à l'arbre et la conclusion pour chacune d'entre elle (modalité prépondérante de la variable à prédire).
   b. Créer votre propre arbre de décision en prenant l'ensemble des variables cette fois-ci. Élaguer ou segmenter l'arbre si nécessaire. La matrice de confusion obtenue est-elle meilleure que la précédente ?

[1] Le choix de la partition dépend de différents critères selon les situations : pourcentages d'individus dans chaque classe, variance intra et variance inter, analyse graphique etc mais surtout d'un aspect « connaissance métier » très important. Ce dernier n'existant pas ici, votre choix sera donc personnel.

   c. Exporter l'arbre nouvellement créé au format image dans vos documents personnels.
5. Ajouter dans la base de données une colonne contenant le nom de la classe auquel appartient chaque individu (Archivage (prédictions, axes, partitions)).

Exercice 4 : Apprentissage supervisé – Confrontation de plusieurs modèles

1. Ajouter un modèle de Réseau Bayésien sur les données et archiver les prédictions.
2. Ajouter un modèle de Réseau de neurones sur les données et archiver les prédictions.
3. Comparer maintenant les trois modèles statistiques (Courbes de lift). Quel est le meilleur modèle ?