Fouille de données – TP 3

M2 Informatique, Université Paris Diderot
Anne-Claire Haury
2017/2018

L'objectif de ce troisième TP est de poursuivre le travail que vous avez initié au TP 2 pour vous essayer au clustering et à la classification supervisée. À partir du jeu de données que vous avez déjà transformé, vous allez pouvoir tester différents modèles et comprendre comment choisir le meilleur.

Apprentissage non supervisé (clustering)

Exercice 1 - Un simple KMeans

Trouvez dans la bibliothèque `sklearn` la fonction qui vous sera utile pour appliquer un algorithme du KMeans. Lancez sur l'ensemble `X` du TP précédent un KMeans avec K = 10. Vous verrez un exemple de script sur la page d'aide de l'algorithme. Inspirez-vous en !

Évaluez cet algorithme en regardant la valeur de la fonction `sklearn.metrics.silhouette_score`. Expliquez ce que représente cette valeur.
Bonus : écrivez une fonction qui trouve les 10 mots les plus utilisés dans chaque cluster.

Exercice 2 - Choisir le meilleur "K"

Cette fois, vous ne connaissez pas la valeur de K. Comment trouver la meilleure valeur de K ? Implémentez cette solution.

Exercice 3 - Comparer deux algorithmes de clustering

Vous n'avez presque rien à faire pour lancer cette fois un algorithme de classification hiérarchique sur ces données. L'algorithme s'appelle `AgglomerativeClustering` dans la bibliothèque `sklearn`.

Apprentissage supervisé (classification)

Cette fois, on cherche à prédire le type d'un message (spam ou ham). On va donc utiliser y qu'on avait laissé de côté dans la partie précédente.

Exercice 4 - Classification supervisée: les plus proches voisins

1. Utilisez la fonction `sklearn.model_selection.train_test_split` pour obtenir des ensembles d'entraînement et de test.
2. Lancez un algorithme des plus proches voisins (`KNeighborsClassifier`) avec k = 1 sur l'ensemble d'entraînement. Evaluez le sur l'ensemble de test.
3. Faire de même avec k = 3. Lequel donne les meilleurs résultats?
4. Vous venez de faire quelque chose d'interdit. Pourquoi les étapes 2 et 3 sont-elles interdites ?
5. Utilisez `sklearn.model_selection.StratifiedKFold` pour choisir le meilleur k entre 1 et 100 (uniquement les nombres impairs) sur l'ensemble d'entraînement. Une fois que vous avez trouvé le meilleur k, relancez l'algorithme avec cette valeur sur la totalité de l'ensemble d'entraînement et testez sur l'ensemble de test.

Exercice 5 - Classification supervisée: arbre ou forêt ?

Reprenez la dernière question de l'exercice 4 pour choisir les meilleurs paramètres d'un arbre de décision, puis d'une forêt aléatoire. Quel algorithme donne le meilleur résultat ?