TD Data Mining
Master 2 BioInformatique

February 3, 2020

1 Partie Clustering

Datamining et dynamique moléculaire...

Le logiciel R est un environnement mathématique utilisé pour l'analyse statistique. Il est installé sur les machines du cremi. On trouve dans ce logiciel, plusieurs bibliothéques bien utiles pour le datamining comme, par exemple les bibliothéques `stats` et `cluster`.Celle-ci contiennent les fonctions les plus courantes en clustering (kmeans,agnes,diana, etc...).

R va donc vous permettre de faire du datamining dans un environnement dédié. Maintenant, il vous faut des données...

Vous trouverez à l'adresse habituelle[1], une archive `PDB.zip`. Celle-ci contient :

- deux fichiers PDB contenant des structures normalisées (i.e. un même nombre d'atomes) du complexe III de la Chaîne Respiratoire. Ces fichiers représentent le complexe III dans **deux conformations différentes**.
- un script `generate-dat.sh` permettant de crér un fichier `.dat` à partir d'un fichier PDB en ne gardant que les positions des atomes,
- un fichier distance.py calculant des distances euclidiennes entre deux fichiers `.dat`.

Votre mission, si vous l'acceptez[2], est de **découvrir les mouvements internes du complexe III** en vous basant sur les fichiers PDB et une suite de clustering...

- Donnez la procédure qui vous permet d'obtenir un clustering.
- Donnez une visualisation de ce résultat par exemple en nuages de points.
- Donnez vos conclusions.

2 Partie Deep Learning

A partir du résultat du clustering, ajouter dans votre fichier .dat une colonne avec le numéro de la classe affectée à chaque atome.
Maintenant votre objectif est de créer un modèle basé sur les réseaux de neuronnes pour prédire la classification d'autres molécules.
Pour cela vous allez utiliser vos résultats précédents pour mettre en oeuvre votre réseau. Vous utiliserez votre fichier ,dat comme vérité terrain. Vous découperez comme il se doit ce fichier pour obtenir un jeu d'entrainement, un jeu de validation et un de test.

- Définir les paramètres du réseau qui permettent de répondre à la question.

[1] `http://dept-info.labri.u-bordeaux.fr/~beurton/Enseignement/DataMining/projet`
[2] mais en fait vous n'avez pas le choix, donc la réponse est oui

- Mettre en oeuvre le modèle dans un environnement réseau de neuronnes au choix : tensorflow ou pytorch et le faire tourner.
- Donnez la qualité, accuracy et loss, de vos résultats.
- Quelles sont vos conclusions sur les capacités de votre réseau avec ce jeu de données, proposer des améliorations.

Pour conclure cet exercice, proposez des ajouts d'analyse, par exemple une analyse multi-dimmensionnelle pour améliorer le résultat de cette analyse.