**Méthodes d’évaluation des classifications**

**Objectif** : évaluer des schémas de classification supervisée et non supervisée.

**Trois approches** :

**Approche externe** : on confronte la classification obtenue avec une classification préétablie. Il s’agit de mesurer l’adéquation entre le résultat obtenu et le résultat attendu.

**Approche interne** : on n’utilise pas de connaissances externe mais uniquement les données d’entrée (c’est-à-dire les composantes produites par l’analyseur et la classification générée par le classifieur). Typiquement, il s’agit de mesure la perte d’information entre la matrice de dissimilarité entre les points de données et la classification finale. Une bonne classification, dans cette approche, minimisera la perte d’information.

**Approche relative** : cette dernière approche s’intéresse plus à la structure de la classification dans l’espace ambiant. En effet, elle valorise les classifications présentant des groupes compacts (faible inertie, ou variance, intraclasse) et globalement éloignés les un des autres (forte inertie interclasse). Notamment, on utilise souvent cette approche pour affiner les paramètres des algorithmes de classification.

La diversité des structures des classifications appelle une diversité dans les indices d’évaluation de ces dernières. Notamment, on distingue des indices d’évaluation des partitions strictes, des indices pour les partitions floues et des indices pour les hiérarchies. Comme nous n’avons pas pour l’instant focalisé notre attention sur les hiérarchies produites essentiellement par les classifieurs hiérarchiques, je me focalise pour l’instant sur les deux types de partitions que nous produisons : partition stricte et partition floue. Il est important de noter qu’une partition stricte est toujours une partition floue. **C’est pourquoi je vous demande de bien vous conformer dans les sorties de vos classifieurs à ma demande de matrice stochastique par ligne.**

**Evaluation d’une partition stricte**

**Approche externe** On compare ici une partition obtenue par un classifieur avec une partition connue a priori.

Statistique de Huberts : on regarde si les distance entre les classes de la partition calculée sont proches de celles de la partition connue. Une statistique plus grande indique que la partition calculée est proche de la partition connue.

On appelle liaison une paire de deux points appartenant à la même classe et non-liaison une paire de deux points n’appartenant pas à la même classe.

Indice de Rand : on regarde le taux de liaisons et de non liaisons correctes. Un indice élevé indique une partition calculée conforme à la partition connue.

Indice de Jaccard : on regarde le taux de liaisons correctes. Un indice élevé indique une partition calculée conforme à la partition connue.

Indice de Fowlkes et Mallows : un indice élevé indique une partition calculée conforme à la partition connue.

Indice de pureté (mesure d’entropie) : on s’intéresse à la pureté des classes de la partition calculée vis-à-vis de la partition connue. Un indice proche de 0 indique une grande pureté de la classification calculée.

**Approche interne**Cette approche vise à valoriser les classifications qui portent une information proche de la matrice de dissimilarité.

Statistique de Huberts : on regarde la conformité de la classification avec la matrice de dissimilarité. Une statistique élevée indique que la classification respecte la matrice de dissimilarité, ce qui signifie que la classification est bonne par rapport à la matrice de dissimilarité.

**Approche relative** Cette approche valorise les classifications qui sont conformes avec l’intuition géométrique de la notion de classes.

Statistique de Huberts : cette statistique mesure la compacité et la séparation des classes. Une statistique élevée indique une bonne classification au sens où elle correspond à une classification géométrique intuitive des points dans l’espace ambiant.

Indice de Dunn : cet indice valorise les classifications dont les classes ont des petits diamètres et sont fortement séparées dans l’espace. Un grand indice indique une bonne classification.

Indice de Davies et Bouldin : cet indice valorise les classifications dont les éléments de chaque classe sont peu dispersés et dont les classes sont fortement séparées dans l’espace. Un petit indice indique une bonne classification.

**En résumé :**

* pour tester un classifieur, je vous conseille d’utiliser une méthode externe de votre choix ;
* pour tester un analyseur, je vous conseille d’utiliser une méthode externe de votre choix ;
* pour tester des méthodes de chunking (découpage), je vous conseille une approche interne ou relative ;
* pour faire de l’ajustement de paramètres de classifieurs ou d’analyseurs (renormalisation notamment), je vous conseille une approche relative.

**Attention, tous les indices ne sont pas des indicateurs croissants de la qualité, certains indices sont petits quand la classification est bonne. Lire le détail de chaque indice pour être sûr avant de l’utiliser.**

Afin de pouvoir comparer les algorithmes dont nous disposons, je vous invite à suivre le protocole suivant dès que Guillaume aura mis à disposition les deux corpus d’œuvres à analyser. Ce protocole présente l’intérêt d’une interopérabilité entre les méthodes de clustering et les méthodes de classification avec apprentissage supervisée.

Chaque corpus est découpé en une partie utilisable pour l’apprentissage et une partie à classer.

* Analyser le corpus avec l’analyseur de référence
* Clustering : classer l’ensemble du corpus et retirer de la partition obtenue les textes de la partie apprentissage

Supervisé : entraîner le classifieur et classifier les textes à classer

* Claculer l’indice de Rand de la partition obtenue.

**Evaluation d’une partition floue**

En chantier

**Evaluation d’une hiérarchie**

Me dire s’il y a une autre raison que le nom qui en jette pour que je calcule le coefficient de corrélation cophénétique.

Si vous avez des questions sur les algorithmes d’évaluation ou sur les protocoles d’évaluation à mettre en place, n’hésitez pas à me contacter.