TP5 ET 6 M1 IAA

UNIVERSITÉ TOULOUSE III PAUL SABATIER

Classification et évaluation

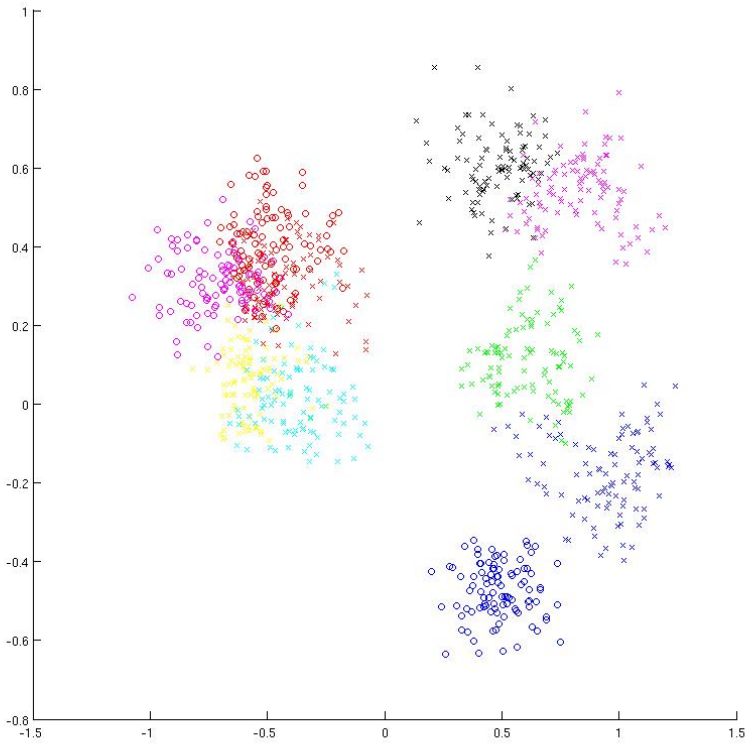
Lors des TPs précédents, vous avez mis en place des fonctions de classification de données :

- K-means
- perceptron
- loi normale
- K-plus proches voisins

La méthode des K-means permet d'obtenir une séparation en K zones homogènes. Le perceptron implémenté permet seulement de trouver un hyperplan séparateur entre deux classes. Les méthodes basées sur la modélisation par loi normale et les K-plus proches voisins permettent de traiter plusieurs classes, mais de manière supervisée (vous disposez d'échantillons représentatifs de vos classes pour mettre en place les classifieurs).

L'objectif de ces deux TP est de pouvoir comparer les méthodes de classification en se basant sur un protocole expérimental rigoureux.

Vous disposez pour réaliser vos expériences de signal audio qui a été enregistré en studio. Le locuteur prononce 10 voyelles. Il y a 100 occurrences de chaque voyelle. Chaque signal possède 1024 échantillons. Une paramétrisation cepstrale a été extraite de ces échantillons, et une Analyse en Composantes Principale a été réalisée afin de réduire la dimension des données à 2. Vous disposez dans le fichier `data2.py` de : 10 classes de voyelles prononcées 100 fois représentées par des échantillons en dimension 2.



1. Classification supervisée

Vous allez comparer les méthodes de modélisation par loi normale et les K-plus proches voisins.

(1) Séparez vos données en sous-ensembles disjoints : les données d'apprentissage et les données de test. Prenez pour commencer 80 % des fichiers pour l'apprentissage et 20 % pour les tests.

(2) Apprenez les paramètres des lois normales et mettez en place la méthode de K-ppv

(3) Évaluez vos systèmes : produisez une matrice de confusion, un taux de reconnaissance et une marge d'erreur (avec un intervalle de confiance à 95%).

(4) Quels sont avantages/inconvénients des deux méthodes (en prenant par exemple comme éléments de comparaison : temps de calcul, prérequis sur les données, efficacité) ?

(5) Faites varier le nombre de données dans les sous-ensembles et comparez les résultats. Vous pouvez produire des courbes sur le taux d'erreur en fonction de la taille des sous-ensembles.

(6) Si l'on souhaite utiliser un maximum de données pour réaliser l'apprentissage, il est possible d'effectuer une évaluation par validation croisée : quel est le principe de cette méthode ? Sous quelles appellations anglaises peut-on trouver cette technique ?

(7) Reprenez la classification supervisée en utilisant les données de dimension supérieure (`data3.py` et `data12.py`). Est-ce que les performances diffèrent en fonction de la dimension ?

2. Classification non supervisée

Vous allez appliquer la méthode des K-means sur trois images. Vous utiliserez uniquement la composante teinte (T) de la représentation TLS (Teinte Saturation Lumière).



Comment évaluer la classification que vous avez produite ? Quelles mesures serait-il possible de mettre en place afin d'évaluer de manière automatique la performance en segmentations en classes ?

Annexes

Matrice de confusion. Permet de présenter les résultats de manière à laisser plus d'informations sur les erreurs (confusions) comises. Cela consiste à rajouter à chaque test +1 dans la matrice à la ligne qui correspond à la classe réelle d'appartenance de la donnée et à la colonne correspondant à la décision.

Ex :

$$\begin{pmatrix} 8 & 0 & 2 & 0 \\ 0 & 9 & 0 & 1 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 10 \end{pmatrix}$$

Ici la classe 1 a été reconnue correctement 8 fois et il y a eu deux confusions avec la classe 3 : le système s'est trompé deux fois et a confondu une observation de la classe réelle 1 avec la classe 3.

Le taux de reconnaissance global est de $37/40 = 92.5\%$

Intervalle de confiance. Permet de quantifier la marge d'erreur des résultats : cela permet d'évaluer la précision de l'estimation d'un paramètres statistique sur une série de données.

$$e = z\sqrt{\frac{P * (1 - P)}{N}}$$

avec :

– z la constante est issue de la loi normale selon un certain seuil de confiance (avec un seuil de confiance de 95% $z = 1,96$)

– P le pourcentage à encadrer
– N la taille de l'échantillon (le nombre de tests réalisés)
– e la marge d'erreur d'échantillonnage

Donc si vous obtenez un résultat P, celui-ci sera considéré, avec une marge d'erreur à 95% égal à [P-e,P+e].

Représentation TLS d'une image. `http://fr.wikipedia.org/wiki/Teinte_saturation_lumiere`