Travail dirigé n^o 4

Reconnaissance de formes : approches statistiques

Instructions

- Les travaux pratiques sont effectués par équipes de deux.
- Le compte rendu doit comporter une réponse concise mais complète à chacune des questions, accompagnée au besoin des courbes, figures et images appropriées ;
- Le compte rendu doit être rédigé à l'aide des fonctionnalités de publication de `matlab` (menu "File / Publish" de l'éditeur `matlab`), en format html ou pdf. L'ensemble des fichiers doit être placé dans une unique archive `zip`.
- Le compte rendu doit être remis au plus tard à minuit le jour de la séance, en utilisant l'outil approprié disponible sur le site web du cours.
- Le travail doit être remis par un seul des membres du groupe. Si tel n'est pas le cas, la version la plus récente du travail remis sera prise en compte.

1 Introduction

L'objet de cette séance de travaux pratiques est de mettre en œuvre certaines des techniques de reconnaissance de formes présentées en cours et d'évaluer leur comportement. Nous allons nous intéresser ici plus particulièrement à la classification automatique par espèces de spécimens d'iris. Il s'agit d'un sous-ensemble des données rassemblées par Edgar Anderson dans son étude géobotanique de la Gaspésie.

Pour chaque fleur, on mesure la longueur et la largeur des sépales et des pétales. Ces informations suffisent pour distinguer les spécimens des espèces *setosa*, *versicolor* et *virginica*.

Téléchargement des données

L'archive `TD4.zip`, qui contient l'unique fichier de données `iris.mat`, est disponible sur le site web du cours. Récupérez-le et chargez le fichier de données dans Matlab. Vous aurez ainsi accès aux données d'entraînement (variables `setosa`, `versicolor` et `virginica`), ainsi qu'aux spécimens à classifier (variables `fleurs` et `classes`). Les spécimens pour l'entraînement et la classification sont stockés dans des matrices de dimension $N \times 4$, chaque ligne correspondant à un spécimen. Finalement, la classification véritable des spécimens de la matrice `fleurs` stockée dans le vecteur

colonne `classes` : l'élément k de ce vecteur, correspondant à la ligne k de `fleurs`, indique s'il s'agit d'un spécimen de *setosa* (0), *versicolor* (1) ou de *virginica* (2).

2 Classification automatique de spécimens d'iris

Comme indiqué plus haut, l'objectif général du travail dirigé est de comparer deux approches de la classification automatique en les appliquant à la reconnaissance d'espèces d'iris : l'approche empirique (distance minimale) et l'approche bayésienne. Nous donnons ci-dessous quelques précisions sur ces approches de la classification.

2.1 Précisions sur les méthodes à employer

Une première approche très simple à mettre en œuvre est une approche empirique basée sur la distance minimale. Si nous choisissons une norme euclidienne, nous avons :

$$d_i(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}_i\|$$

où $\boldsymbol{x}_i$ représente la moyenne empirique de la classe i. Cette approche peut être mise en œuvre très simplement en utilisant la fonction `norm` de `matlab`.

La seconde approche est celle de type bayésien. Pour chaque classe ω_i (dans notre cas, $i = 1$, 2 ou 3), le risque bayésien a pour expression :

$$E[C(\widehat{\omega}_i, \omega_j)|\boldsymbol{x}] = r_i(\boldsymbol{x}) = \sum_{j=1}^{3} C(\widehat{\omega}_i, \omega_j) f(\boldsymbol{x}|\omega_j) f(\omega_j)$$

où $r_i(\boldsymbol{x})$ représente la nouvelle fonction de décision, $C(\widehat{\omega}_i, \omega_j)$ le coût, soit l'importance que l'on accorde au fait de décider que $\boldsymbol{x}$ appartient à la classe i alors qu'il appartient en réalité à la classe j, $f(\omega_i)$ la probabilité d'être dans la classe i. Nous allons ici travailler dans le cas particulier où $f(\boldsymbol{x}|\omega_i)$ est une gaussienne, soit $f(\boldsymbol{x}|\omega_i) = N(\boldsymbol{m}_i, \boldsymbol{R}_i)$. Comme pour la première approche, nous trouvons la classification en minimisant par rapport à i la fonction de décision, qui est ici égale à $r_i(\boldsymbol{x})$.

2.2 Travail à effectuer

Utilisez les spécimens des matrices `setosa`, `versicolor` et `virginica` comme ensembles d'apprentissage pour déterminer les moyennes et matrices de covariance utiles pour la suite. Programmez et testez les deux approches présentées précédemment afin de classifier les spécimens de la matrice `fleurs`. Comparez vos résultats aux vraies occurrences stockées dans le vecteur `classes` (que l'on pourra afficher en utilisant la fonction `stem` de `Matlab`).

Travail à remettre Code permettant d'effectuer les deux approches. Les valeurs choisies pour le coût et $f(\omega_i)$. Graphiques représentant les occurrences obtenues pour chacune des deux approches. Commentaires sur le comportement respectif de chacune des méthodes. Brève critique sur les approches utilisées et sur le choix des attributs. Suggestion pour optimiser la détection des attributs.

Annexe - Commentaire sur les signaux utilisés

Les données utilisées pour ce laboratoire sont issues de la banque de données de l'University of California-Irvine (disponible à l'adresse : http://archive.ics.uci.edu/ml/machine-learning-databases/iris/. Comme mentionné en introduction, il s'agit de données géobotaniques acquises par le botaniste Edgar Anderson et utilisées d'abord par Sir Ronald Aylmer Fisher pour la mise au point de la classification par discriminant linéaire. Référence : http://en.wikipedia.org/wiki/Iris_flower_data_set