

## Projet 2005-2006

# Construction d'un classifieur homme/femme pour la reconnaissance de visages.

### Introduction

Le but poursuivi à travers le projet est de vous amener à utiliser des concepts vus au cours sur des données relativement complexes. En l'occurrence, les données du projet sont des visages d'hommes et de femmes. On vous demande de construire un modèle qui soit capable de reconnaître si le visage qui lui est présenté est un visage d'homme ou de femme. Pour ce faire, vous utiliserez un classifieur de type Support Vector Machines (SVM).

Précisons que l'objectif principal de ce projet porte plus sur l'analyse des données, la mise en oeuvre des méthodes et l'analyse que vous ferez de vos résultats et de votre démarche que sur les performances en classification du modèle utilisé.

### Description des données

Les données mises à votre disposition sont 750 images de visages en noir et blanc, codés sur 150x150 pixels. Pour notre problème de reconnaissance de visages, les différentes étapes de prétraitement suivantes ont déjà été effectuées :

- une rotation, pour aligner les yeux sur une ligne horizontale,
- une mise à l'échelle pour avoir des images 150x150 pixels,
- une application d'un cache elliptique pour supprimer la partie de l'image qui entoure le visage,
- une harmonisation de la courbe de niveau, pour qu'il y ait autant de pixel en blanc, en gris et en noir.

Toutes ces étapes ont été réalisées pour vous simplifier la tâche, afin que les visages aient une représentation qui soit la même, autant que possible. Il ne vous est donc plus nécessaire de prétraiter les données bien qu'il s'agisse là d'une étape importante de toute analyse de données.

### Visualisation et réduction des données

Les 750 visages qui vous sont fournis sont des images de 150x150 pixels. Vous manipulerez donc initialement des objets dans un espace de dimension 22500.

Une première étape de votre travail sera de vous familiariser avec ces données, notamment en les visualisant. Comme il est très difficile de manipuler des objets dans un espace de grande dimension, il vous est demandé de transformer les images en utilisant une analyse en composantes principales (PCA) pour réduire la dimension des données.

Pour calculer les composantes de la PCA, vous utiliserez la fonction Matlab `svd(., 0)`, où le deuxième argument sert à réduire la mémoire utilisée. Bien entendu, il est conseillé de centrer et réduire les données avant d'appliquer la PCA.

Pour sélectionner le nombre de composantes qui seront conservées, vous devez analyser la proportion de variance conservée en fonction du nombre de composantes. En général, on considère qu'en conservant 95% de la variance on ne perd pas trop d'information. Bien entendu, vous pouvez considérer un seuil nettement inférieur à 95%. Il est utile de comparer les images initiales et les images réduites par PCA puis reconstruites par la projection inverse. Vous pouvez ainsi d'observer la quantité d'information perdue lors des manipulations par PCA.

## Les Support Vector Machines

Nous ne reviendrons pas sur les développements qui mènent aux modèles SVM, puisqu'ils ont été vus au cours. Nous ne reprenons ici que les équations principales qui seront celles que vous devrez utiliser pour implémenter une version simple des SVM. Il s'agit de la version des SVM intitulée "1-norm" dans le cours.

La formulation duale nous dit qu'il faut résoudre le problème suivant :

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \alpha^T Q \alpha - 1^T \alpha \\ \text{s.c.} \quad & 0 \leq \alpha_i \leq C \\ & t^T \alpha = C \end{aligned} \quad (1)$$

avec  $Q_{ij} = t_i t_j K(x_i, x_j)$ , et où  $t$  est le vecteur des étiquettes de classe, valant -1 ou 1, pour les différentes observations. La notation  $K(x_i, x_j)$  désigne le noyau qui est utilisé. Par définition, le noyau est un produit scalaire entre les données dans un espace transformé.

Dans le cadre de ce projet, il vous est demandé de comparer les performances obtenues avec deux types de noyaux différents. Le premier noyau est appelé linéaire et correspond au produit scalaire classique :

$$K(x_i, x_j) = x_i^T \cdot x_j \quad (2)$$

Le second noyau à utiliser est le noyau gaussien :

$$K(x_i, x_j) = e^{-\left(\frac{1}{2} \frac{\|x_i - x_j\|^2}{\sigma^2}\right)} \quad (3)$$

Après apprentissage, pour obtenir la classe d'une nouvelle donnée, il suffit alors d'appliquer la formule :

$$y(x) = \text{sign}(\sum t_i \alpha_i K(x_i, x) + b), \quad (4)$$

où la formulation primale permet d'obtenir le biais  $b$ . En effet, pour les vecteurs de support  $sv$  bornés, c'est-à-dire tels que  $0 < \alpha < C$ , on a

$$\sum \alpha_i K(x_i, x_{sv}) + b = t_{sv} \quad (5)$$

Remarque : dans cette formulation, nous n'avons pas explicitement défini le critère d'erreur à utiliser. Vous êtes donc libre de choisir le critère qui vous semble le plus opportun étant donné ce contexte de classification.

## Sélection de modèles et test de performance

Pour rappel, le but de la sélection de modèle est d'optimiser les valeurs des hyperparamètres d'un modèle, comme par exemple le nombre d'unités d'un RBF, d'un MLP ou d'un SOM ; la largeur des gaussiennes d'un RBF ; etc. Différentes techniques de sélection de modèles ont été vues au cours, le projet sera l'occasion d'appliquer l'une de ces méthodes.

Dans ce contexte de classification des images, il vous est demandé d'appliquer la cross-validation pour optimiser d'une part le paramètre de régularisation  $C$  du SVM, et d'autre part la largeur  $\sigma$  des noyaux gaussiens. Pour rappel, la cross-validation consiste en une découpe du fichier de données disponibles en deux ensembles distincts, l'ensemble d'apprentissage et l'ensemble de validation. Cette découpe est répétée un certain nombre de fois, en faisant varier le contenu des deux ensembles.

Lors de votre utilisation de la cross-validation, nous vous conseillons de découper le fichier de données en un ensemble d'apprentissage et un ensemble de validation. Une proportion habituelle est de l'ordre de  $2/3$  pour l'ensemble d'apprentissage,  $1/3$  pour l'ensemble de validation.

**Indication** : ces opérations sont relativement lourdes en temps calcul, surtout de par la dimension des données à manipuler et donc des matrices que Matlab va générer. Pensez à sauver les ensembles d'apprentissage et de validation pour ne pas devoir les recréer lors de chaque nouvelle simulation. Sauvegardez également un maximum de résultats intermédiaires pour ne pas devoir tout recommencer à chaque fois.

Pour vous permettre de tester les performances en classification de votre meilleur modèle, nous mettrons à disposition un nouvel ensemble de données comportant cette fois 300 visages. Cet ensemble sera rendu disponible sur la page web du cours à partir du 15 décembre.

## Programmes mis à disposition

Les visages vous sont fournis au format « JPEG » pour vous permettre de les visualiser plus aisément et de réduire la taille du dossier. La fonction **load\_dataProjet2005** vous permet de charger et de sauvegarder les images dans le format de Matlab. Vous disposerez alors de deux matrices contenant respectivement les images correspondant aux hommes et aux femmes. Après chargement des données, chaque image occupe une colonne (de longueur 22500), qui correspond en fait à une matrice  $150 \times 150$  dont toutes les colonnes ont été mises bout à bout.

Pour visualiser dans Matlab un visage, qui serait par exemple sur la colonne  $X(:,10)$ , nous vous conseillons d'utiliser les fonctions Matlab **imagesc** et **colormap**, ce qui donne la ligne de code suivante :

```
imagesc(reshape(X(:,10), 150, 150)) ; colormap gray ;
```

La fonction **quadprog** de MATLAB permet de résoudre des problèmes quadratiques tels que ceux apparaissant dans la formulation du SVM. Cependant, en pratique, l'optimisation d'un SVM avec cette fonction prend un certain temps. La fonction **quadprog\_svmProjet2005** devrait vous permettre de gagner du temps.

## Rapport

Ce projet devra être réalisé par groupe de deux. Nous vous laissons vous arranger entre vous pour la composition des groupes, mais nous vous demandons qu'un de vous deux nous envoie un e-mail aux adresse ci-dessous dès que la composition est fixée (nom + e-mail des deux membres du groupe).

Nous vous demandons de rédiger un rapport, qui servira de base à la discussion qui suivra l'examen du cours. Votre rapport (trois copies sur papier) doit nous être rendu au plus tard le 23 décembre à 17h. Le rapport comptera un maximum de 10 pages, avec votre code éventuellement en annexe. Vous pouvez nous envoyer votre code par e-mail.

Dans ce rapport mettez l'accent sur votre démarche, vos choix, les résultats obtenus et l'analyse de ces résultats. Inutile de nous présenter le modèle SVM, de nous expliquer la sélection de modèle par cross-validation ou encore de nous rappeler le principe de la PCA. Autant que possible, illustrez vos résultats aux moyens de graphiques et commentez-les. Ces graphes doivent supporter votre propos, soyez donc cohérents.

Il est possible, en fonction de certains de vos choix, que certains résultats ne vous semblent pas logiques. Essayez néanmoins d'analyser les raisons de ces résultats. Dans ce genre de situation, les graphes sont parfois une aide précieuse.

Enfin, nous nous permettons d'insister : vous êtes volontairement limités en place, ne mentionnez donc que l'essentiel.

## Pour ouvrir la réflexion

Ci-dessous, nous avons listé quelques questions subsidiaires qui peuvent vous aider à approfondir la réflexion sur les méthodes utilisées. La réponse à ces questions ne doit pas nécessairement figurer dans votre rapport, mais il est quand même conseillé d'y réfléchir.

- Comment pourrait-on faire pour repérer les zones de l'image les plus utiles pour la classification ?
- Le noyau gaussien, tel qu'il vous est demandé de l'utiliser, donne autant d'importance à chaque dimension. Que pourrait-on faire pour donner plus d'importance aux dimensions qui sont les plus pertinentes pour la classification ?
- Les deux classes homme/femme reçoivent a priori la même importance dans la classification. Comment pourrait-on faire pour ajuster cette importance ?

## Détails pratiques

Tout comme cet énoncé, les données se trouvent sur la page web du cours :

<http://www.dice.ucl.ac.be/~verleyse/lectures/elec2870/elec2870.htm>

En cas de question, il vous est demandé de nous envoyer un e-mail ([nicolas.delannay@uclouvain.be](mailto:nicolas.delannay@uclouvain.be) ou [geoffroy.simon@uclouvain.be](mailto:geoffroy.simon@uclouvain.be)). Si nécessaire, il est possible de fixer un rendez-vous pour discuter plus en détails de vos éventuels problèmes. Normalement, ces rendez-vous auront lieu pendant les horaires des séances, à savoir le mercredi de 14h à 16h et le vendredi de 10h45 à 12h45.