Introduction à l'analyse de données
Sujets de mini-projets

Sujet 1

On dispose d'une séquence de n images $I_1, ..., I_n$ de même dimensions. Il peut s'agir de la base d'image de taxis, de séquences récupérées sur ce site ou de toute autre séquence de votre choix.

1) En général ces séquences sont ordonnées, c'est à dire qu'on peut les afficher rapidement et donner l'impression d'un mouvement. Proposer une méthode pour changer aléatoirement l'ordre, c'est à dire pour récupérer une suite de nombres $i_1, ..., i_n$ obtenus par permutation de $1, ..., n$.

2) Proposer une méthode pour réordonner la séquence d'images $I_{i_1}, ..., I_{i_n}$. Cette méthode pourra être personnelle ou pourra utiliser les techniques de réduction de dimension vues en cours (analyse en composantes principales, Analyse discriminante linéaire , ISOMAP). Dans tous les cas, **justifier** l'utilisation de la méthode dans le problème considéré.

3) Enfin évaluer la qualité de votre approche sur différentes séquences en proposant une distance entre la séquence d'indices obtenue et la séquence correcte $1, ..., n$.

Sujet 2

On considère ici une base d'images de chiffres manuscrits. Cette base est composée d'un ensemble d'apprentissage (environ 7000 images) et d'un ensemble de test (environ 2000 images). Chaque image est de taille 16×16 et codée sur 31 valeurs entre -1 et 1. En outre, à chaque image correspond un entier entre 0 et 9 qui est le chiffre représenté par l'image et qu'on appelera *label*.

Sujet 2 a

On cherche ici à s'initier à la *classification supervisée*. Le but est d'obtenir un algorithme capable de prédire le label d'une image de label inconnu. Bien sur il existe une littérature colossale sur le sujet avec un grand nombre de techniques dont les plus populaires sont les réseaux de neurones ou les SVM. Il ne s'agit pas d'utiliser ces techniques, sauf bien sur si vous êtes très motivés mais d'essayer plutôt des approches plus intuitives ou personnelles.
Par exemple vous pourrez vous documenter sur la *classification par plus proches voisins*. Si vous choisissez une telle approche, il conviendra de tester de façon précise les différents paramètres libres de la méthode comme le nombre de voisins, le choix de la distance comme par exemple celles qui utilisent des normes p :

$$||\mathbf{x}||_p = \sqrt[p]{\sum_{i=1}^{n} x_i^p}$$

Pour résumer :

1) Implémenter/utiliser une méthode de classification et **justifier** le choix de cette méthode. Vous avez le choix d'utiliser les intensités des pixels comme caractéristiques ou bien vous pouvez inventez ou utiliser votre propre méthode.

2) Evaluer la qualité de votre approche sur la base de chiffres (ou une autre si cela vous chante) en proposant une façon personnelle pour mesurer le pourcentage de bonne classification.

3) Tester l'influence des paramètres éventuels présents dans votre méthode.

Sujet 2 b

On cherche ici à découvrir le *clustering* ou *partitionnement de données* (sous-branche de la *classification non supervisée*). Le but est de considérer une base d'images, comme celle des chiffres, et d'essayer d'identifier des groupes d'images ou *classes* qui partagent une similarité. La difficulté du problème général est qu'il s'agit souvent d'un problème d'optimisation de type NP complet. Par ailleurs le nombre de groupes recherchés est a priori inconnu.

1) Implémenter/utiliser une méthode de clustering et **justifier** rapidement le choix de cette méthode. Vous avez le choix d'utiliser les intensités des pixels comme caractéristiques ou bien vous pouvez inventez ou utiliser votre propre méthode.

2) Pour l'évaluation de votre méthode on pourra procéder de la façon suivante : il est a priori raisonnable de rechercher autant de groupes que de chiffres possibles, soit 10. Ensuite, comme on connait malgré tout le label de chaque image on peut essayer de voir quelle est la proportion de chiffres bien reconnus pour chaque groupe. Une façon de faire est de construire un *tableau de contingences* T de taille 10×10 où la case $T(i, j)$ correspond au nombre de fois que le chiffre i apparait dans le groupe j.
Enfin il convient de voir si les regroupements coincident mieux que le hasard avec les 10 classes.

3) Reprenez les analyses précédentes lorsque le nombre de groupes n'est plus fixé à 10. En particulier y aurait il un moyen d'objectiver que 10 est le nombre de groupe optimal ?

Sujet 2 c

Dans ce sous-projet on cherche à étudier les *modes de variation* des chiffres en utilisant les techniques vues en cours. L'idée principale est qu'on cherche à capturer les différentes façon d'écrire un chiffre par des techniques de réduction de dimension.
1) Comparer différentes méthodes telles que l'Analyse en Composantes Principales ou l'ISOMAP pour représenter la façon dont varient les chiffres "4" et "9", par exemple. Se contenter de projections en deux dimensions et essayer d'interpréter les axes.

2) Toujours sur l'exemple des "4" et des "9" (ou peut-être plus facile des "6" et des "7"), proposez une méthode de classification entre ces deux chiffres en utilisant uniquement deux premières composantes principales (d'un certain ensemble d'images obtenues à partir du fichier post_train.txt). La méthode sera essentiellement géométrique et pourra être très simple dans sa mise en oeuvre. Vous validerez la méthode sur les images du fichier post_test.txt .

3) Commentez enfin sur les performances de classification et la question de la variabilité intrinsèque aux données.

Sujet 3

On s'intéresse ici à des questions d'appariements ou de classification pour des bases d'images représentant par exemple une personne sous différents aspects. Pour les deux premiers sujets on propose deux bases d'images, ici et là.

Sujet 3 a

1) Sur l'une ou l'autre des deux bases, calculer une matrice de distance entre les différentes images. Attention, cette étape peut être assez longue aussi il conviendra de sauvegarder la matrice et peut-être

éventuellement de sous-échantillonner les images. Le choix de la distance (ou de la similarité) peut avoir son importance aussi vous pourrez, dans la mesure du possible, essayer des distances obtenues avec la norme p (cf sujet 2a) ou avec des corrélations.

2) A partir de cette matrice, proposer une visualisation des relations entre ces images qui utilise la méthode d'ISOMAP.

3) Proposer enfin une approche personnelle pour réaliser l'appariement de deux images représentant la même personne (couleur vs dessin pour la première base ou neutre vs sourire pour la deuxième base).

Sujet 3 b

1) et 2) Voir les mêmes questions du sujet 3 a.

3) La méthode ISOMAP fournit elle une bonne discrimination des visages d'hommes et de femmes ? Proposer une évaluation quantitative pour **justifier votre réponse**.

Sujet 3 c

On considère ici une base de visages de 13 personnes avec pour chacun 7 émotions différentes (colère, dégoût, peur, joie, douleur, surprise et neutre). Cette base est disponible ici.

1) Calculer une matrice de distance entre les différentes images. Le choix de la distance (ou de la similarité) peut avoir son importance aussi vous pourrez, dans la mesure du possible, essayer des distances obtenues avec la norme p (cf sujet 2a) ou avec des corrélations.

2) A partir de cette matrice, proposer une visualisation des relations entre ces 91 images qui utilise la méthode d'ISOMAP. En particulier vous pourrez générer deux images qui privilégient pour l'une les personnes et pour l'autre les émotions.

3) Est il a priori plus facile de distinguer les personnes ou les émotions ? Proposez les grandes lignes d'une démarche permettant de valider votre réponse. Si vous avez le temps de la mettre en oeuvre ce sera un bonus apprécié.

Sujet 4

Dans les TPs notés, les professeurs reçoivent parfois des soumissions identiques pour des groupes différents, même des fois, ces étudiants oublient de changer leurs noms dans les commentaires. D'autres étudiant(e)s (plus intelligent(e)s) modifient le code copié par :

- Changer les noms de variables, fonction.
- Supprimer/ajouter des espaces et changer les commentaires
- Changer l'ordre quand il est possible.
- Etc.

Cependant, d'autres caractéristiques restent inchangées. Par exemple :

- Le nombre de boucles
- Le nombre et la forme de conditions
- le nombre de variables
- Etc.

Proposer une méthode pour estimer la distance (la dissemblance) entre deux fonctions écrites en Matlab.

1) D'abord, il faut trouver un moyen pour représenter les caractéristiques de la fonction numériquement (Par exemple, une combinaison de : l'histogramme de mots-clés, le nombre de variables, des statistiques des signes utilisés, etc.). Autrement dit, transformer la fonction en descripteur numérique. La méthode proposée pourra utiliser les techniques de réduction de dimensions vues en cours.

2) Enfin, évaluer la qualité de votre approche sur différentes fonctions modifiées. Que se passe-t-il avec des modifications plus intelligentes ?