Télécom ParisTech, CES Data scientist : Évaluation du module 5
J. Salmon

Classification

Consigne : Rédiger les réponses aux questions suivantes et rendre le travail sous forme d'un document pdf (latex conseillé). Les questions numériques doivent être traitées en Python et le code utilisé pour générer les résultats doit être fourni dans un fichier `.py` à part, avec commentaires facilitant sa lecture. Les résultats doivent être reproductibles, c'est à dire : le code doit être exécutable par le correcteur. Le format ipython notebook (.ipynb) est aussi accepté. Les résultats numériques et graphiques doivent être inclus dans le fichier pdf. Le tout, rassemblé dans une archive, est à envoyer pour le 31 juillet 2015, 23h59 dernier à l'adresse suivante : joseph.salmon@telecom-paristech.fr.

- Comparaisons de différentes méthodes de classification -

On va comparer dans cette partie les différentes méthodes sur la base de donnée obtenue comme dans le premier TP avec les commandes suivantes :

```
from sklearn.datasets import load_digits
digits = load_digits()
X, y = digits.data, digits.target
```

On suivra le protocole expérimental suivant : couper les données en deux parties 80% pour l'apprentissage et 20% pour la validation (donner la taille des deux blocs choisis). Sur la partie d'apprentissage on entraînera les méthodes suivantes :

- Naive Bayes
- LDA
- Régression logistique
- QDA
- KNN (en prenant comme nombre de voisins $k = 1$)
- KNN (en choisissant k par validation croisée (V-fold) avec $V = 8$)
- une autre méthode de votre choix

On validera leur performance en donnant la proportion d'erreurs de classification faite sur la partie des données gardée pour la validation.

1. Pour les méthodes mentionnées faire un tableau avec les renseignements suivants :
   - temps de calcul en seconde pris par chaque méthode pour la partie apprentissage (pour l'entraînement sur les 80% des données)
   - temps de calcul en seconde pris par chaque méthode pour la partie validation (sur les 80% restants)
   - pourcentage d'erreurs de classification de chaque méthode
2. On affichera les matrices de confusion associées : celles de la meilleure et de la pire des méthodes obtenues (au sens du nombre d'erreurs commises) parmi celles étudiées. Commentez vos résultats.
3. Proposer un (cours) paragraphe synthétisant l'ensemble de vos expériences ci-dessus.

- Création de graphe (sous Matplolib par exemple) -

4. Écrire une fonction qui affiche le graphe (tridimensionnel) d'une densité de gaussienne en dimension deux et qui prend en entrée un vecteur de moyenne $mu = [mu_1, mu_2]$ et une matrice de covariance symétrique Σ (Sigma). On veillera à donner un message d'erreur " Matrice Sigma non-symétrique" dans le cas ou la matrice Σ en entrée n'est pas symétrique. Insérer les graphes dans votre document pour :

$$\Sigma = \begin{bmatrix} 1/5 & 1/10 \\ 1/10 & 1/5 \end{bmatrix}$$

et

$$\Sigma = \begin{bmatrix} 1/5 & -2/10 \\ -2/10 & 3/5 \end{bmatrix}$$

Par simplicité on pourra prendre $mu_1 = mu_2 = 0$.

5. Enregistrer par lignes de commande les figures au format pdf. Veiller à ce que la taille de chaque image en pdf soit plus petit que 400ko. On ajoutera ces deux pdf au fichier déposer.