MLEA - TP 2

Florent D'Halluin

EPITA 2010

8 Juillet 2009

Plan

K-Fold cross-validation

- Divise les données en k sections
- k itérations
- $k - 1$ sections d'apprentissage
- 1 section de test
- Résultat: moyenne des taux de reconnaissances et écart-type
- Ici, $k = 10$

Courbe ROC

- Se base sur le taux de certitude de chaque point reconnu
- Fonction du taux de vrais positifs et de faux positifs

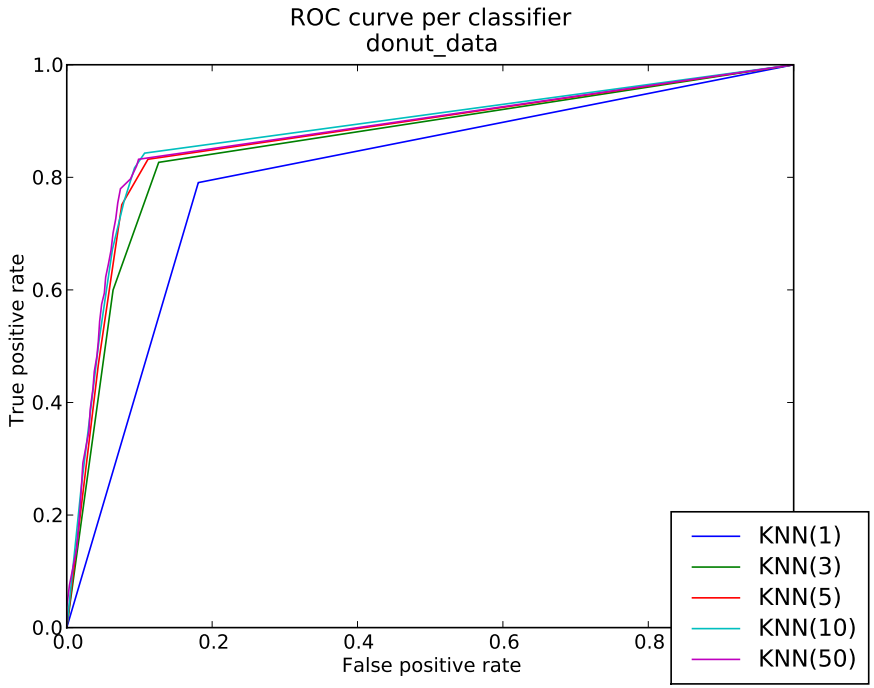

Figure: Courbe ROC pour plusieurs classifieurs, sur donut.

KNN

- On peut calculer la distance entre deux points du dataset
- Trouver les K plus proches voisins

Normal distribution

- Modéliser les données d'apprentissage par une distribution normale (moyenne, écart-type)
- Maximiser la probabilité à postériori

Comparaison

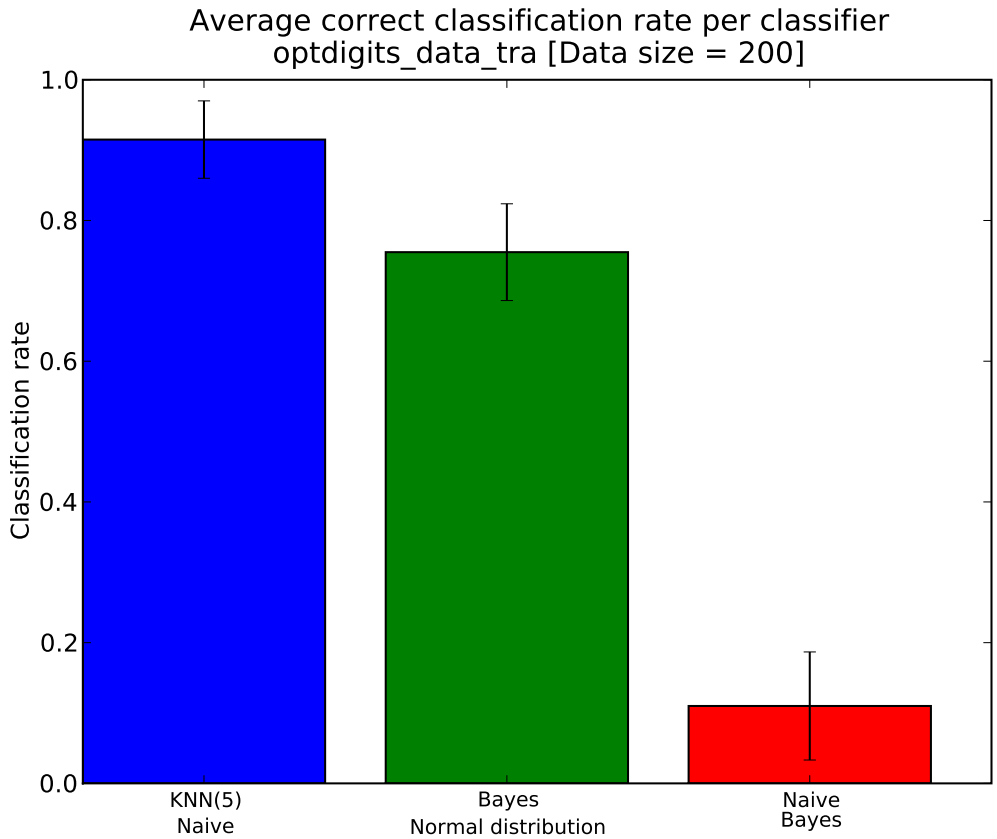

Figure: Comparaison de classifieurs sur une partie du dataset optdigits (training).

Naïve Bayes classifier

- Données à caractéristiques indépendantes
- Estimer certaines probabilités à partir des données d'apprentissage ($P(C = c) et P(Xi = xi|C)$)
- Maximiser $P(C = c|X = x)$

Continuousification

- Rendre des données discrètes continues (pour utiliser KNN)
- Conserver la dépendance entre les variables
- Naïf: Une valeur arbitraire par valeur observée
- NBF: Une dimension par valeur observée
- VDM/MDV: Estimer et utiliser les probabilité conditionnelles

Comparaison

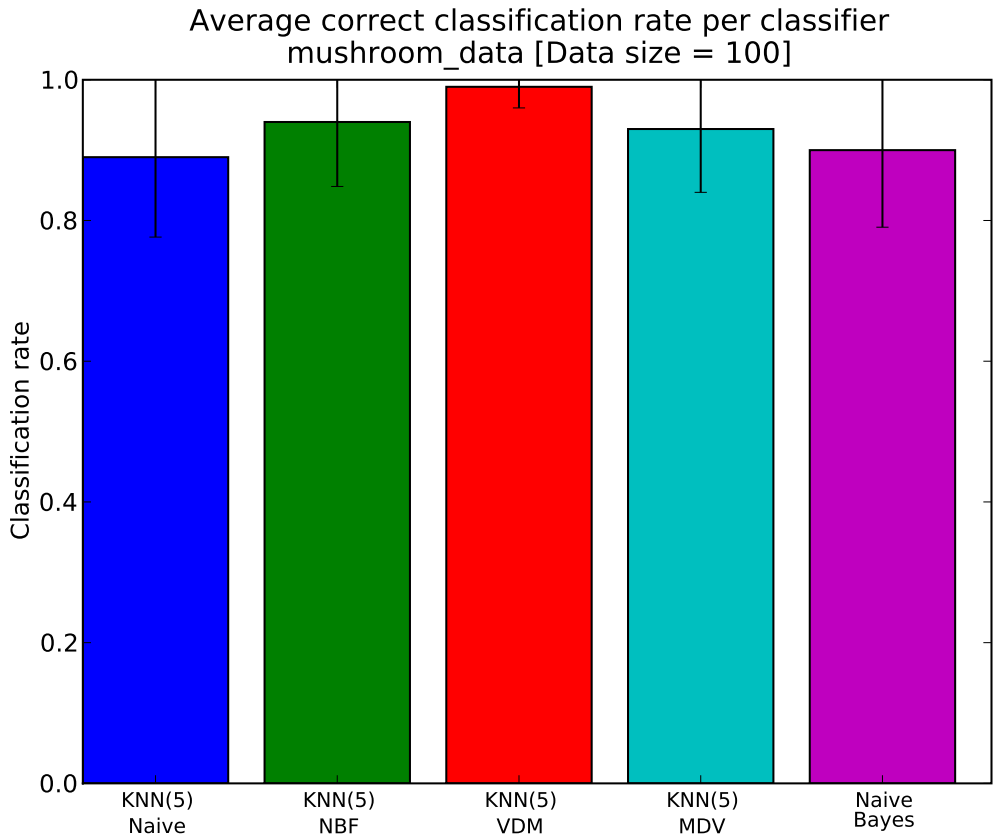

Figure: Comparaison de classifieurs sur une partie du dataset mushroom.

Estimer l'efficacité du clustering

- Détecter la convergence vers un extrémum local
- Somme des distances intra-cluster
- Vitesse de convergence: nombre d'itérations

K-MEANS

- Choisir k centres au hasard
- Calculer les clusters et les nouveaux centres
- Itérer jusqu'à la convergence

K-MEANS - centres initiaux

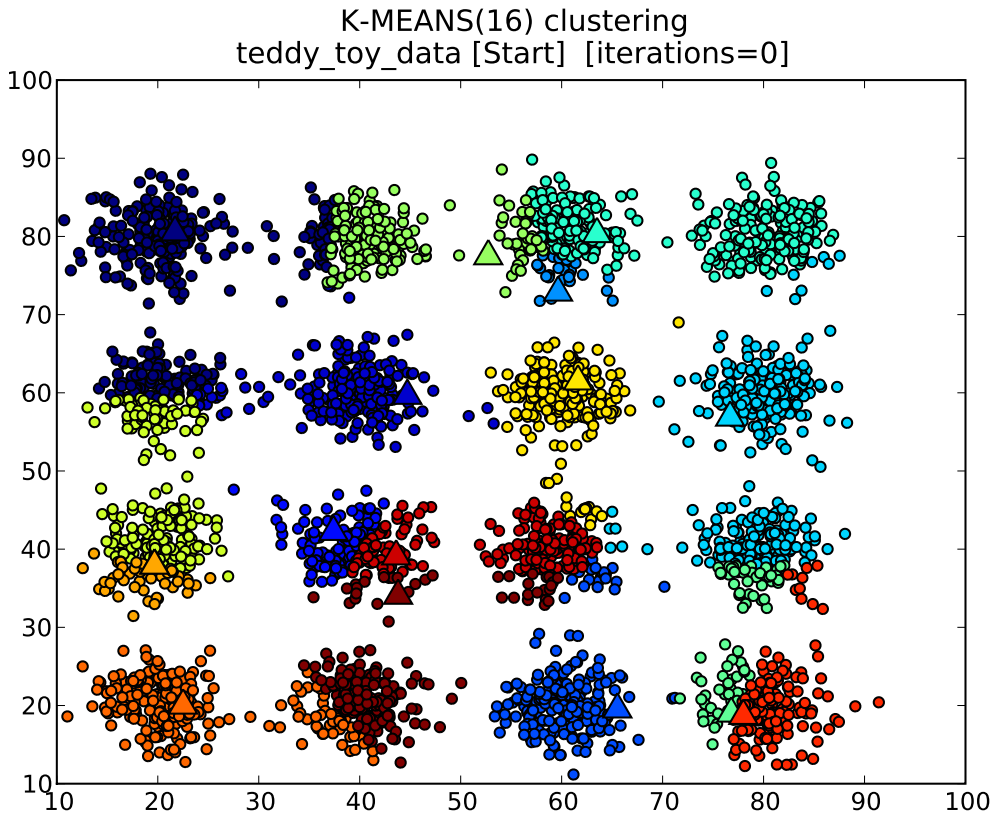

Figure: K-MEANS clustering sur le dataset teddy-toy, centres initiaux.

Distance maximale

- Optimiser le choix des centres initiaux
- Maximiser la distance avec les centres déjà choisis
- Sensible au bruit

Distance maximale - centres initiaux

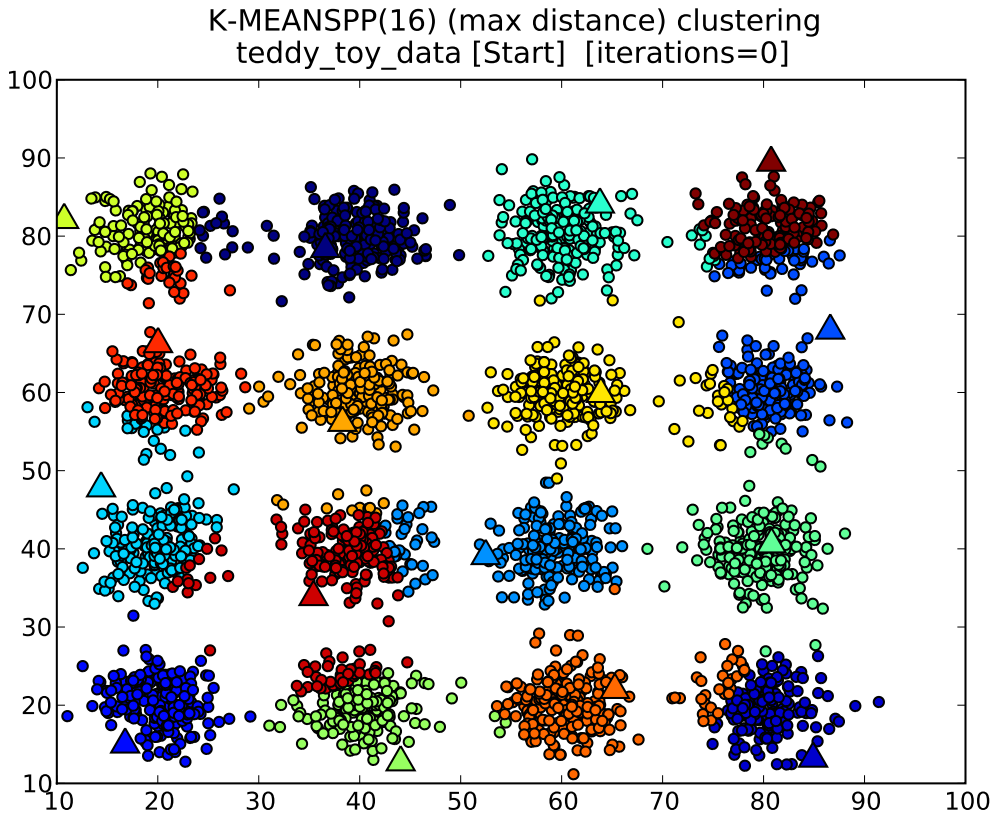

Figure: K-MEANSPP (max distance) clustering sur le dataset teddy-toy, centres initiaux.

K-MEANS++

- Optimiser le choix des centres initiaux
- Choisir en fonction de la distance aux centres déjà choisis
- Résistant au bruit
- Efficace

K-MEANS++ - centres initiaux

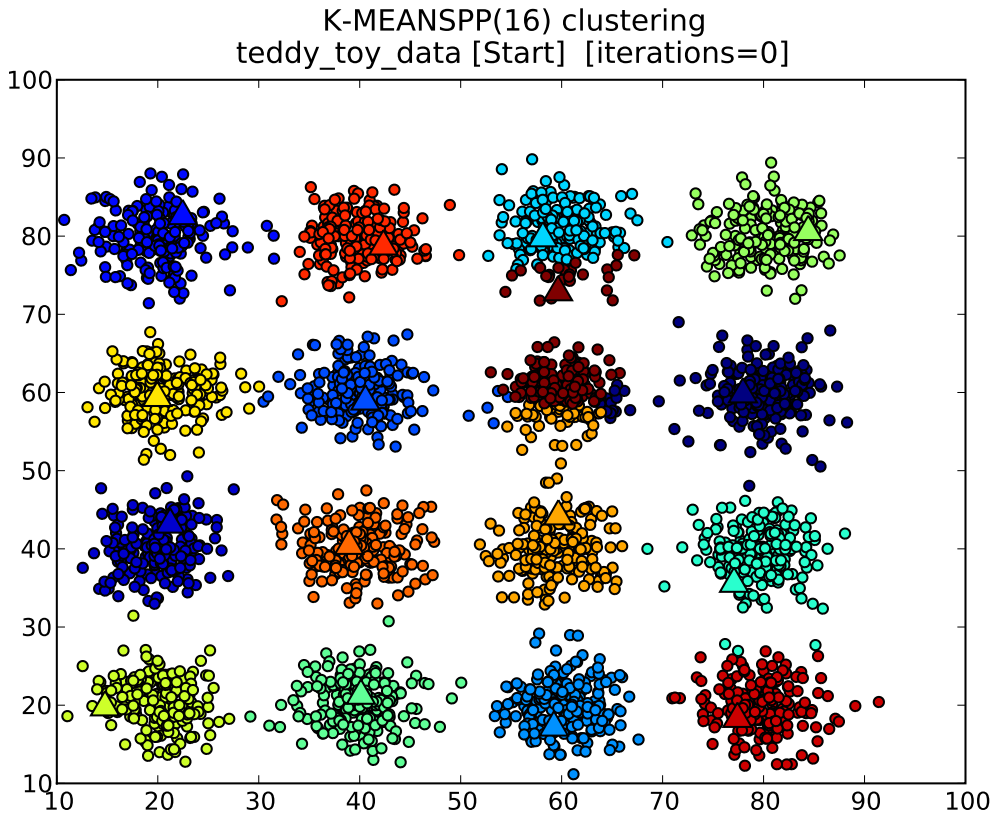

Figure: K-MEANSPP clustering sur le dataset teddy-toy, centres initiaux.

Comparaison - distances intra-cluster

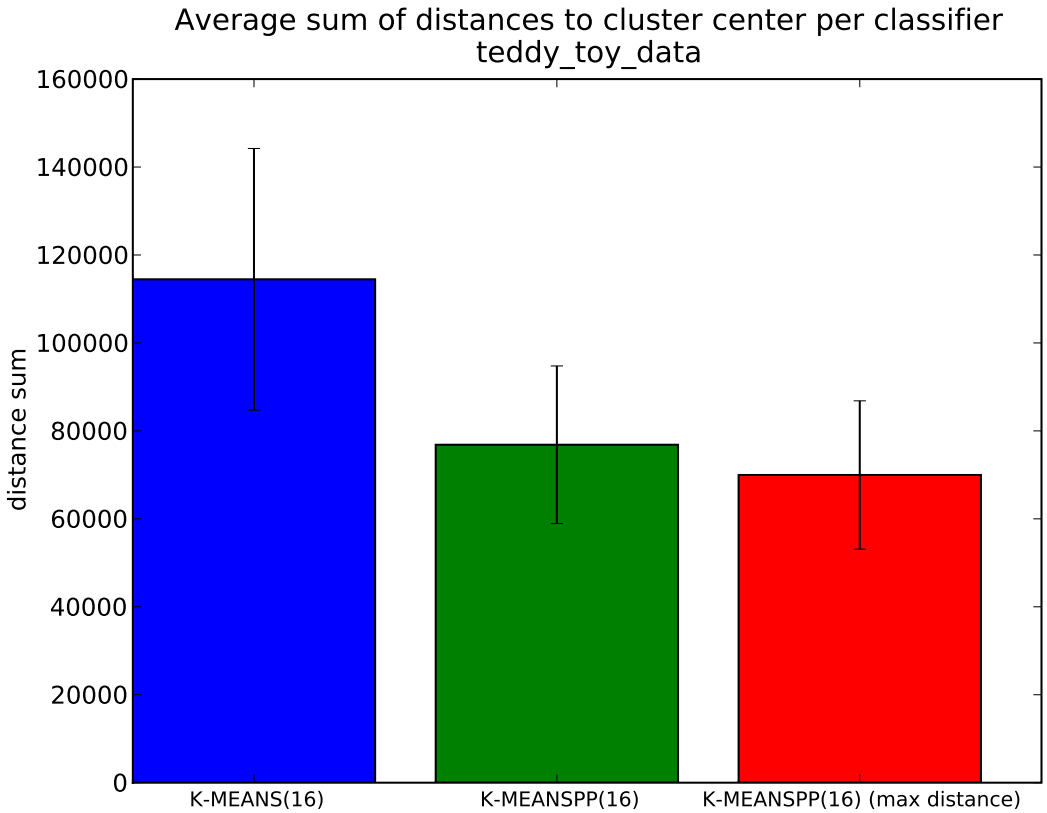

Figure: Somme des distances intra-cluster pour le dataset teddy-toy.

Comparaison - nombre d'itération

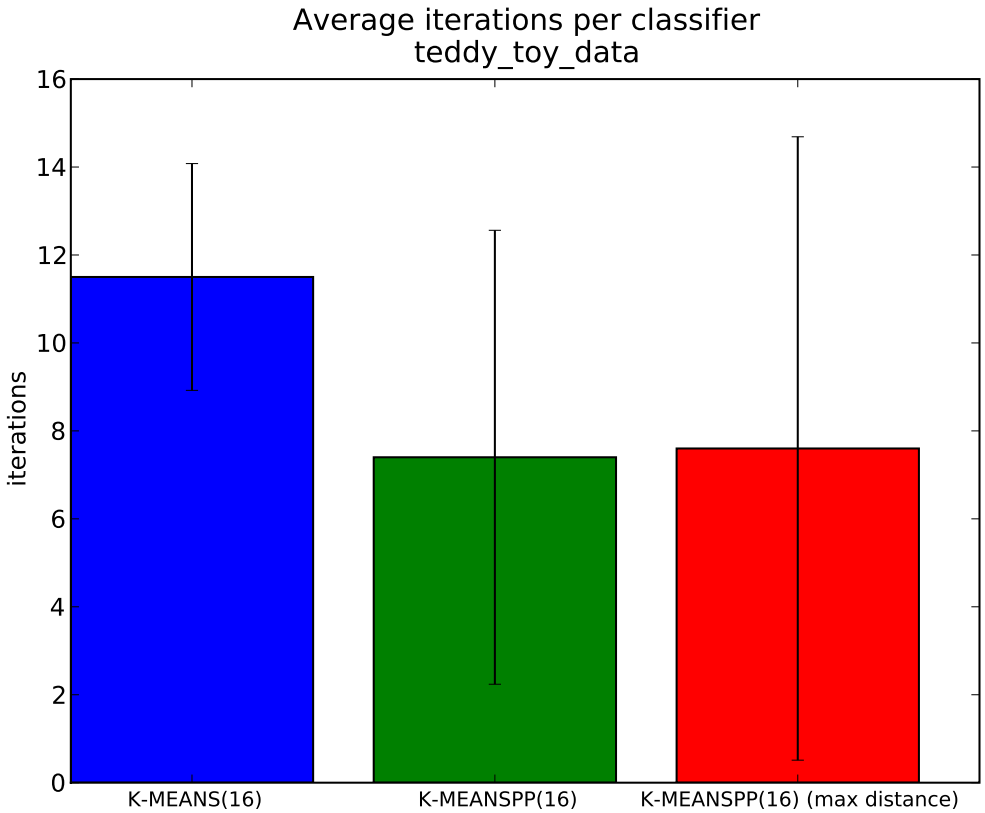

Figure: Nombre d'itérations pour le dataset teddy-toy.

Questions