K-plus proche Voisin (Régression)

L'apprentissage automatique (2016-2017)

UFR MIME

Université Lille 3

7 décembre 2016

Sommaire

1 Types des données

2 k-PPV Régression

Données quantitatives et qualitatives

- Données quantitatives : valeurs numérique soit des scalaires ou vecteurs en $\mathbb{R}^d$
  - L'age, poids, salaires, tailles de maison, couleur (valeur)
- Données qualitatives : valeurs catégorique ou des classes
  - couleur (nom), sexe, nationalité, mode de transport
-

(X) couleur	formes	firmes/pas	(y) Classe
Vert	cylindrique	firme	Légumes
Jaune	cylindrique	mou	Fruit
Blanc	cylindrique	dur	Grain

  - X et y sont catégorique

Classe / $y \in$ { Fruit, Légumes, Grain } , forme $\in$ { rond, cylindrique }

-

(X) # chambres	surface	énergie (grade)	Prix ($)
2	127	1	305000
3	106	3	275000

  - X, y sont numérique, $X \in \mathbb{R}^3$ et $y \in \mathbb{R}$
- Les variables quantitatives / qualitatives peuvent êtres dans les sorties/entrées d'un problème d'apprentissage supervisé

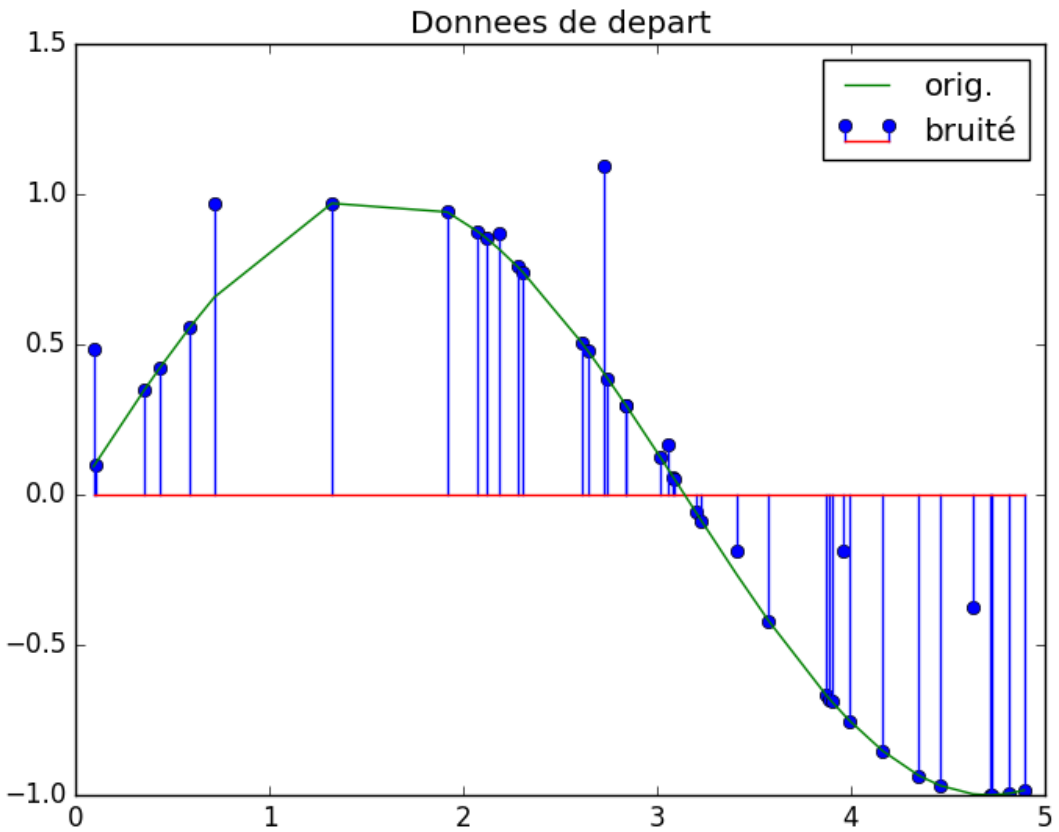
k-PPV [1] : Régression

Algorithme

- Ensemble d'apprentissage (ou Training set) : $X_{\text{Tr}} := \{(\mathbf{x}_1, y_1), \mathbf{x}_2, y_2), \ldots, \mathbf{x}_N, y_N)\}$
- $\mathbf{x}_i \in \mathbb{R}^d$ et $y_i \in \mathbb{R}$
- Pour une nouvelle entrée $\mathbf{z}$ qui vient dans le ensemble test !
- Pour classification
  - $f_\theta(\mathbf{z}) = \text{VoteMajoritaire}\{y_i | i \in \text{k-PPV}(z)\}$ dans X_{Tr}
- Pour Régression
  - $f_\theta(\mathbf{z}) = \text{ValeurMoyenneDe}\{y_i | i \in \text{k-PPV}(z)\}$ dans X_{Tr}
- Pour calculer la k-plus proche voisin de $\mathbf{x}$ utliser la meme algorithme d'avant

1. K plus proche voisin

Exemple

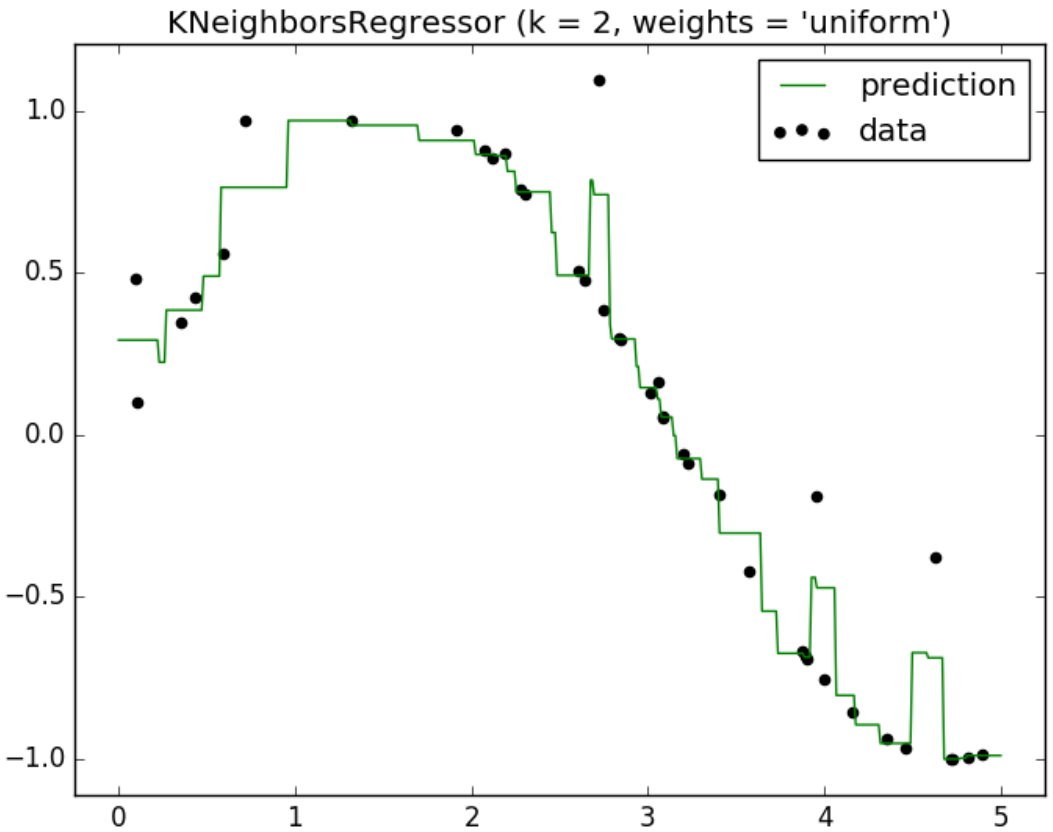


Valeurs Limites

k=1, trop compliqué

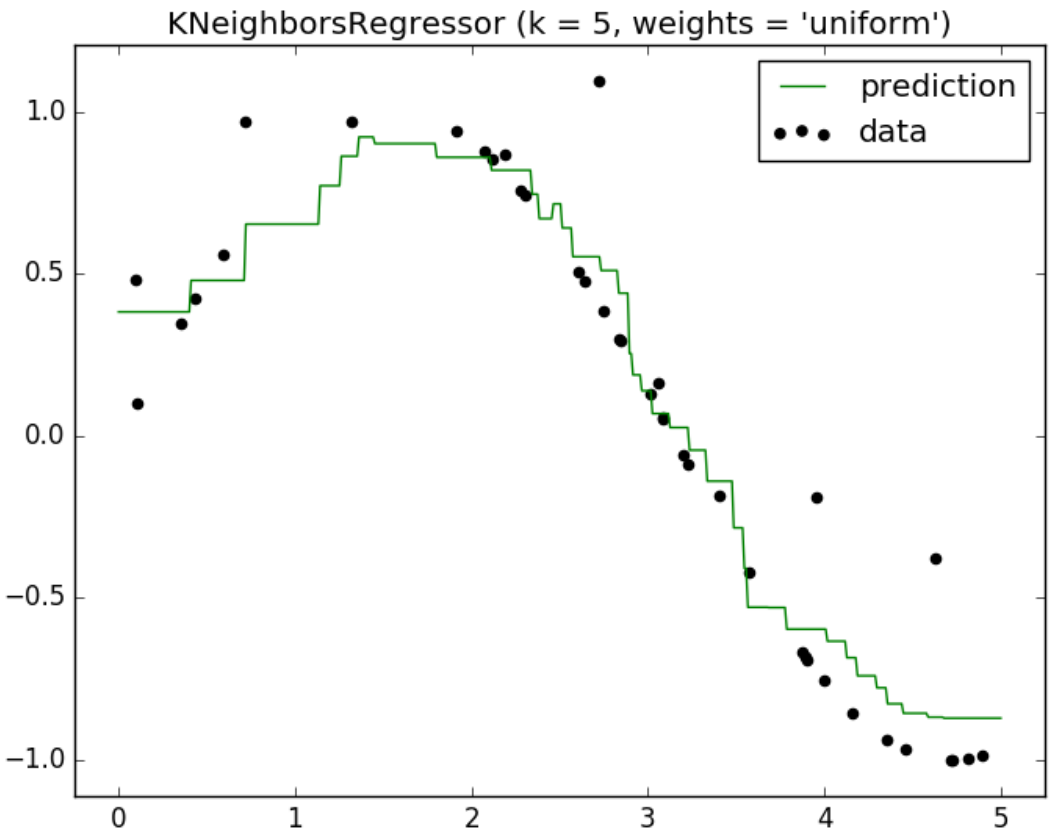
Valeurs Limites

k=2, compliqué



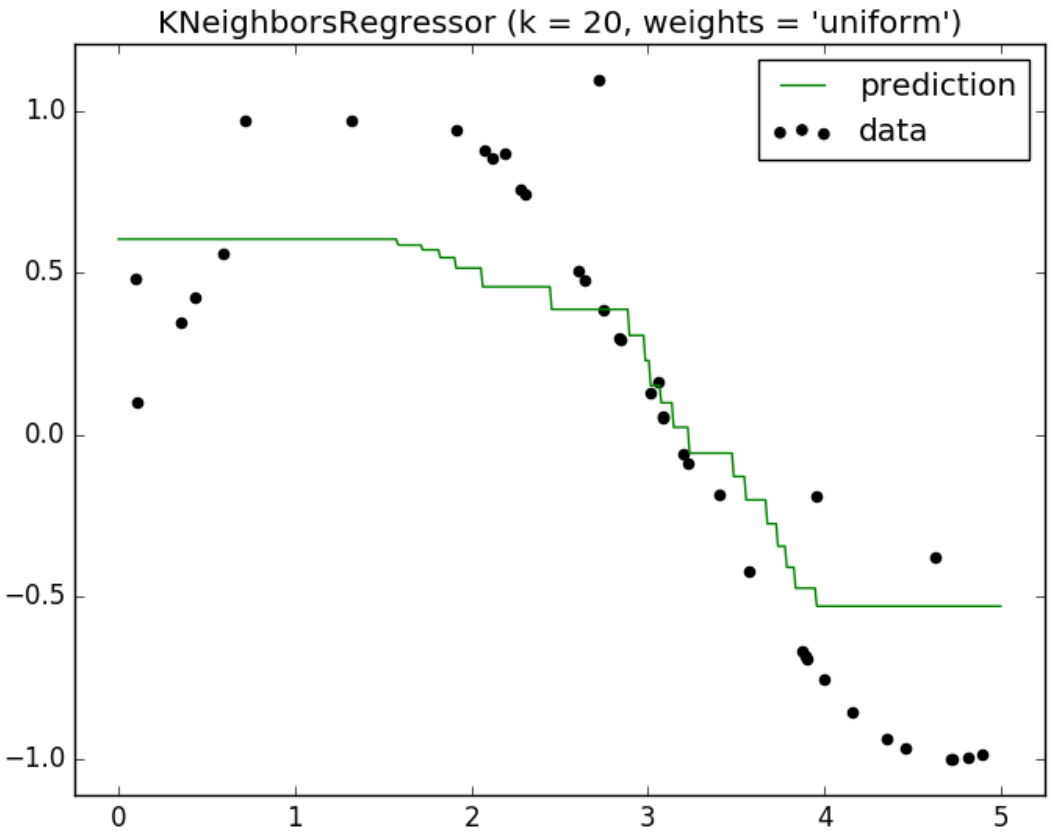
Valeurs Limites

k=5, bien !



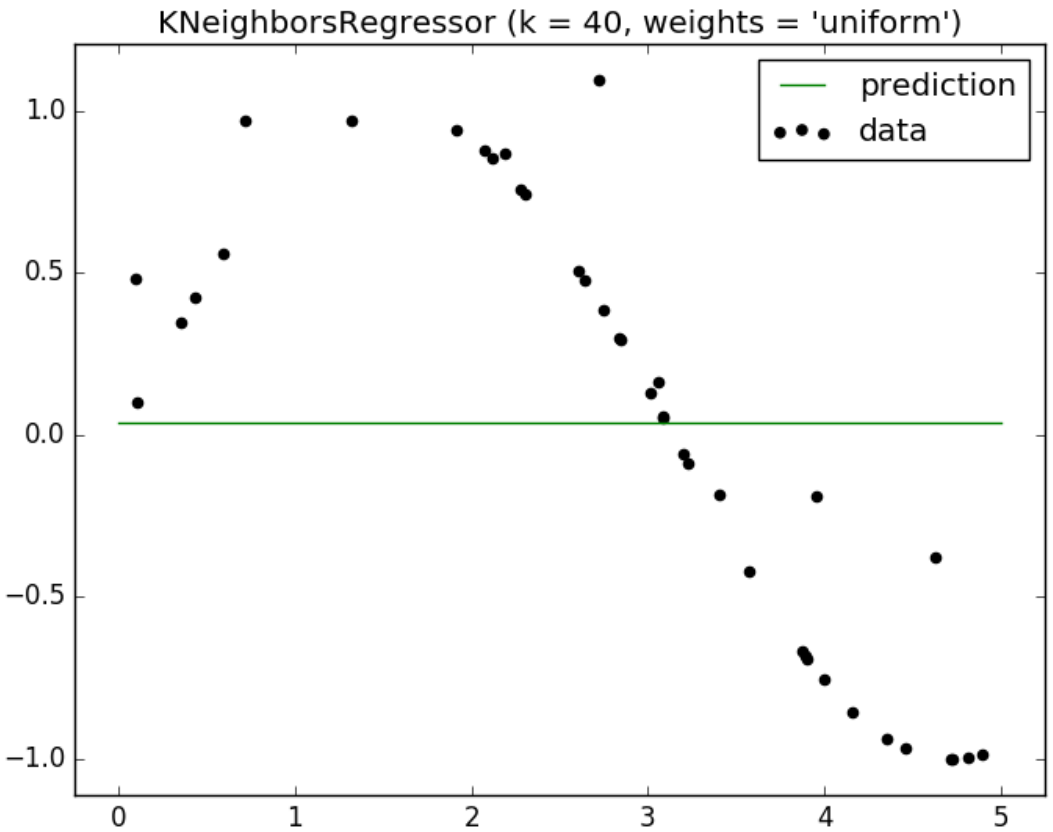
Valeurs Limites

k=20, simplifié



Valeurs Limites

k=40, trop simplifié



Classification Vs Régression

- La régression et utilisé pour prédire les valeurs **continue** en étudiant la relation entres les variables d'entrés, et pour classification on prédit les valeurs **catégorique**
- Pour calculer la distance entres les valeurs catégorique il nous faut une méthode de vectorisation qui associe une valeur numérique a chaque valeur catégorique.
- On peut classifier/régresser les données catégorique avec les arbres de décision. (exemple médecin)
- Évaluation de performance : Taux d'erreurs
  - Classification : $\frac{\#\text{ prédictions incorrecte}}{\#\text{ échantillons dans l'ensemble teste}}$
  - Régression : $\sum_{i=1}^{\text{Ntest}}(\text{cible}_i - \text{prediction}_i)^2 = \sum_{i=1}^{\text{Ntest}}(y_i - \hat{y}_i)$