

Apprentissage statistique : le compromis biais - variance

C. HELBERT

Contexte et notations

- ▶ X_1, \dots, X_p sont des variables explicatives (descripteurs, prédicteurs)
- ▶ Y est la variable à expliquer (variable d'intérêt, réponse).

Objectif du **Supervised learning** : on cherche à modéliser la relation entre Y et (X_1, \dots, X_p) à partir d'un échantillon à n observations.

Different du **Unsupervised learning** : absence de Y .

Contexte et notations

- ▶ X_1, \dots, X_p sont des variables explicatives (descripteurs, prédicteurs)
- ▶ Y est la variable à expliquer (variable d'intérêt, réponse).

Objectif du **Supervised learning** : on cherche à modéliser la relation entre Y et (X_1, \dots, X_p) à partir d'un échantillon à n observations.

Different du **Unsupervised learning** : absence de Y .

Spécificité de l'apprentissage en grande dimension :

- ▶ p grand : $p \approx n$ ou $p > n$
- ▶ diversité des variables explicatives : qualitatives et/ou quantitatives

Dans le contexte supervisé, on distingue deux classes distinctes de problèmes :

- ▶ La régression : la réponse est quantitative
- ▶ La classification : la réponse est qualitative (binaire ou multiclases)

Exemples

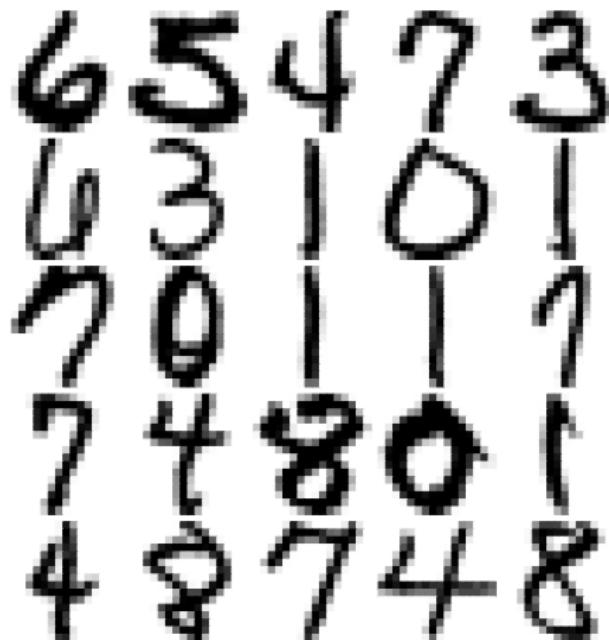
- ▶ Reconnaissance automatique de "spam" (email).

Exemples

- ▶ Reconnaissance automatique de "spam" (email).
- ▶ Reconnaissance automatique de codes postaux

Exemples

- ▶ Reconnaissance automatique de codes postaux



Exemples

- ▶ Reconnaissance automatique de "spam" (email).
- ▶ Reconnaissance automatique de codes postaux
- ▶ Prédiction du taux PSA (Prostat Specific Antigen) en fonction de l'âge, volume de la tumeur, poids de la prostate, etc.
- ▶ Identification de gènes influents dans certaines pathologies.

Plan

Deux approches simples : moindres carrés et k plus proches voisins

Quelques éléments théoriques

Méthodes locales en grande dimension

Compromis biais - variance

On mesure les $p + 1$ variables sur n individus que l'on représente dans \mathbb{R}^n par les vecteurs Y, X_1, \dots, X_p .

On mesure les $p + 1$ variables sur n individus que l'on représente dans \mathbb{R}^n par les vecteurs Y, X_1, \dots, X_p .

On cherche $\hat{\beta}_0, \hat{\beta}_1 \dots \hat{\beta}_p$ tels que $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p = \mathbf{X}\hat{\beta}$ soit le plus proche de Y pour les moindres carrés.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \{ RSS(\beta) \}$$

où $RSS(\beta) = (Y - \mathbf{X}\beta)^T(Y - \mathbf{X}\beta)$.

On mesure les $p + 1$ variables sur n individus que l'on représente dans \mathbb{R}^n par les vecteurs Y, X_1, \dots, X_p .

On cherche $\hat{\beta}_0, \hat{\beta}_1 \dots \hat{\beta}_p$ tels que $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p = \mathbf{X}\hat{\beta}$
 soit le plus proche de Y pour les moindres carrés.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \{ RSS(\beta) \}$$

où $RSS(\beta) = (Y - \mathbf{X}\beta)^T(Y - \mathbf{X}\beta)$.

La réponse est : \hat{Y} est la projection de Y sur W le sous-espace vectoriel engendré par $1, X_1, \dots, X_p$:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Y$$

Exemple : le modèle linéaire dans un contexte de classification

$G \in \{\text{bleu}, \text{rouge}\}$.

Sur un échantillon à n observations, on dispose des variables suivantes :

- ▶ $Y \in \{-1(\text{bleu}), 1(\text{rouge})\}$
- ▶ X_1 et X_2 sont quantitatives

Exemple : le modèle linéaire dans un contexte de classification

$G \in \{bleu, rouge\}$.

Sur un échantillon à n observations, on dispose des variables suivantes :

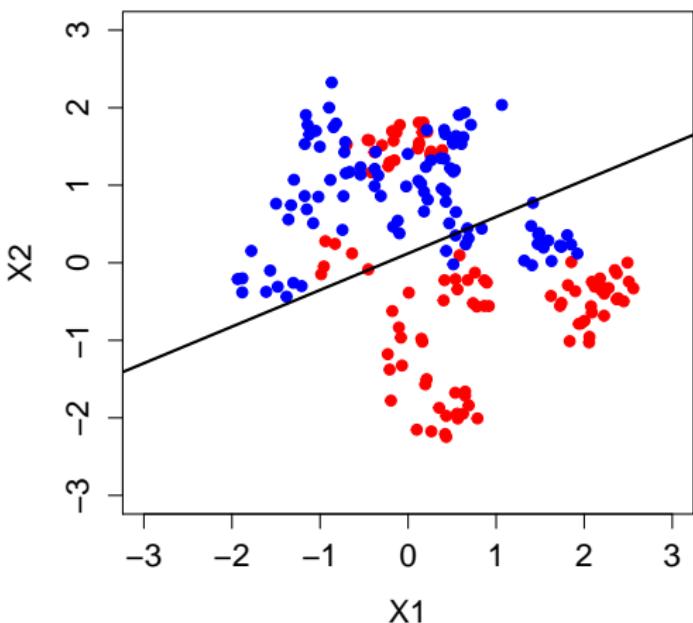
- ▶ $Y \in \{-1(bleu), 1(rouge)\}$
- ▶ X_1 et X_2 sont quantitatives

On cherche $\hat{\beta}$ tels que $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 = \mathbf{X}\hat{\beta}$ soit le plus proche de Y pour les moindres carrés.

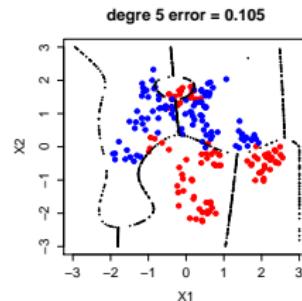
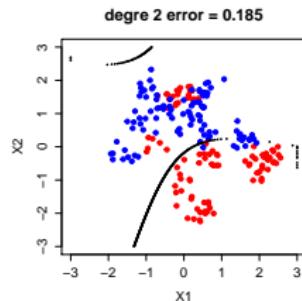
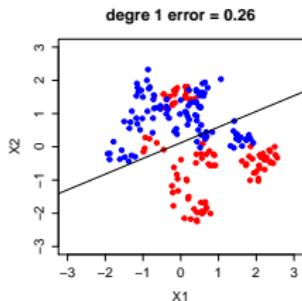
La règle de prédiction est la suivante :

$$\hat{G} = \begin{cases} bleu & \text{si } \hat{Y} \leq 0 \\ rouge & \text{si } \hat{Y} > 0 \end{cases}$$

degre 1 error = 0.26



On peut complexifier ... quel est l'intérêt ?



Question essentielle : d'où viennent les données ? Deux scenarii :

- ▶ scenario1 : les individus de chaque classe sont issus d'une gaussienne bivariée
- ▶ scenario2 : les individus viennent d'un mélange de 10 gaussiennes bivariées avec des écarts-types petits

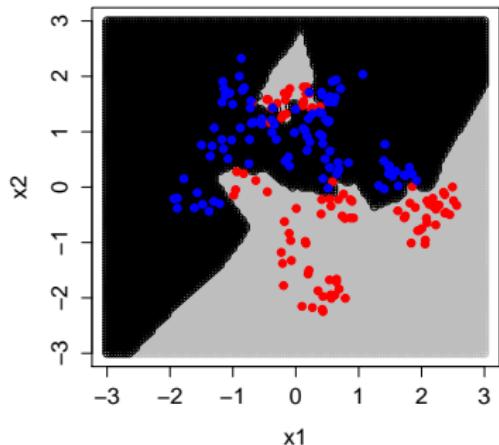
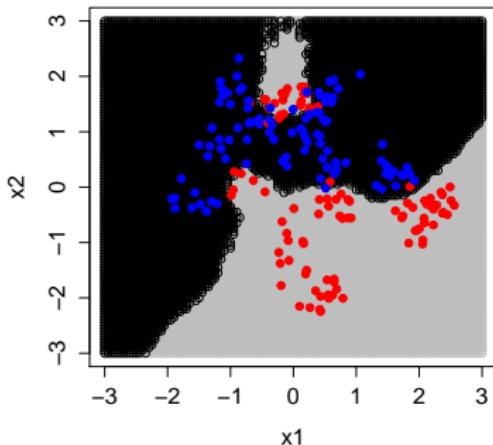
La prédiction par k plus proches voisins (kNN - k Nearest Neighbor) est la suivante :

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

où $N_k(x)$ est constitué des k plus proches voisins de x.

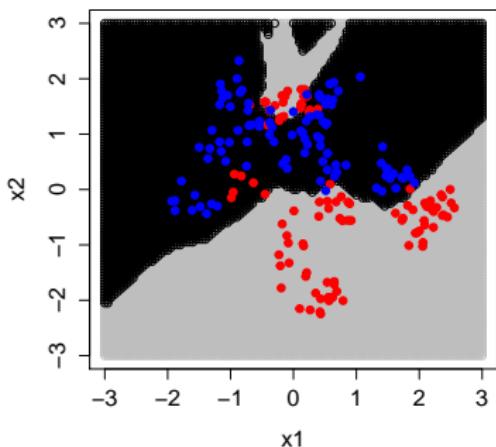
La règle de prédiction est la même :

$$\hat{G} = \begin{cases} \text{bleu} & \text{si } \hat{Y} \leq 0 \\ \text{rouge} & \text{si } \hat{Y} > 0 \end{cases}$$

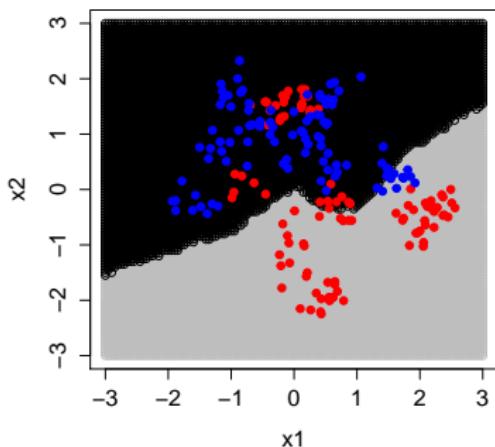
kNN $k = 1$ kNN $k = 10$ 

Méthode très souple...

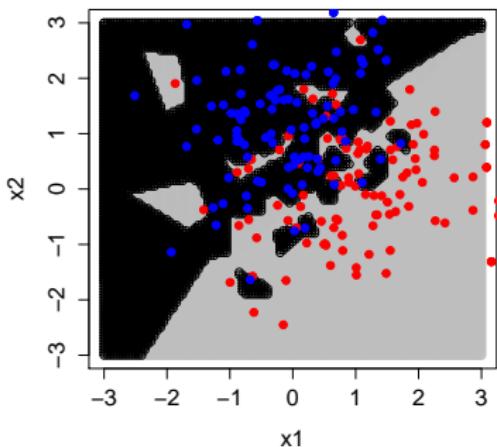
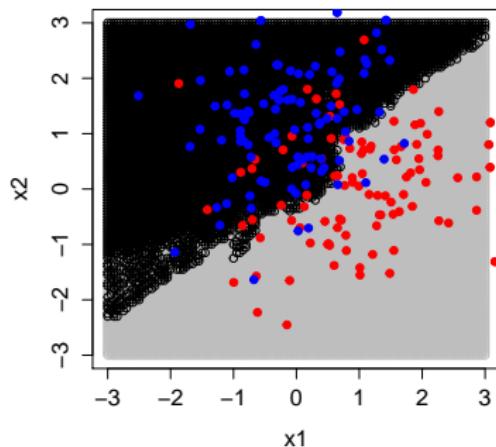
kNN k = 25



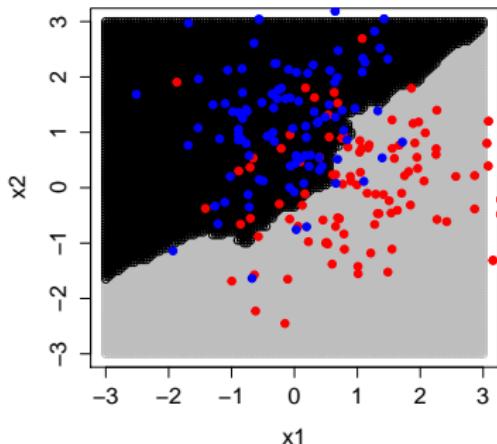
kNN k = 50



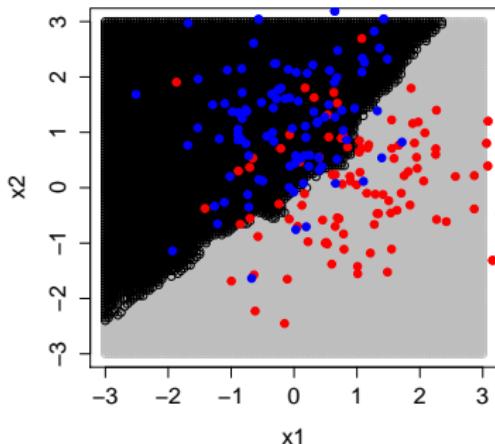
Méthode très souple : un k petit permet de bien prédire les clusters, un k trop grand entraîne une prédiction trop lisse.
Cette méthode convient-elle au scenario 1 ?

kNN $k = 1$ kNN $k = 10$ 

kNN k = 25



kNN k = 50



Bon choix de k : ici il faut prendre k suffisamment grand pour retrouver la séparation linéaire.

- ▶ Régression linéaire = robuste mais repose sur **une hypothèse de modélisation** forte (séparation linéaire convient).
Biais potentiellement élevé, faible variance.
Convient au scenario 1
- ▶ kNN = Pas d'hypothèse, elle peut s'adapter à toutes les situations. Attention, cette méthode dépend directement de la position des points les uns par rapport aux autres. Méthode peu lisse et instable.
Biais faible, forte variance.
Convient au scenario 2

Nombreuses améliorations de ces méthodes (degrés, kNN poids lisse, kNN linéaire locale...).

Plan

Deux approches simples : moindres carrés et k plus proches voisins

Quelques éléments théoriques

Méthodes locales en grande dimension

Compromis biais - variance

On considère une réponse quantitative. On suppose connue la loi jointe (X, Y) .

L'objectif est de trouver une prédiction de Y de type $f(X)$ la plus proche de Y pour un certain critère.

Par exemple, trouver f d'erreur quadratique moyenne minimale :

$$\begin{aligned} EPE(f) &= E(Y - f(X))^2 \\ &= \int [y - f(x)]^2 Pr(dx, dy) \\ &= \int \int [y - f(x)]^2 Pr(dy|dx) Pr(dx) \end{aligned}$$

Il suffit de résoudre le problème point par point :

$$f(x) = \underset{c}{\operatorname{argmin}} E_{Y|X}((Y - c)^2 | X = x)$$

La solution est $f(x) = E(Y|X = x)$

La méthode kNN a pour objectif direct de calculer cette quantité :

$$\hat{f}(x) = \text{Ave} (y_i | x_i \in N_k(x))$$

Deux hypothèses sont faites ici :

- ▶ l'espérance est remplacée par une moyenne sur l'ensemble d'apprentissage
- ▶ le conditionnement en un point est approché par un voisinage "proche" du point en question.

Et la régression ? quelle est sa légitimité ?

Si on fait l'hypothèse supplémentaire :

$$f(x) \approx x^T \beta$$

On est alors dans le cas d'une approche basée sur un modèle. La quantité à estimer est le vecteur β . L'expression minimisant l'erreur quadratique moyenne est :

$$\beta = [E(XX^T)]^{-1}E(XY)$$

On utilise ici toute les valeurs de X (pas de conditionnement). En pratique, l'espérance est remplacée par une moyenne sur les données.

Comment faire un choix entre ces deux approches optimales dans un certain contexte ?

Deux approches simples : moindres carrés et k plus proches voisins

Quelques éléments théoriques

Méthodes locales en grande dimension

Compromis biais - variance

Plan

Deux approches simples : moindres carrés et k plus proches voisins

Quelques éléments théoriques

Méthodes locales en grande dimension

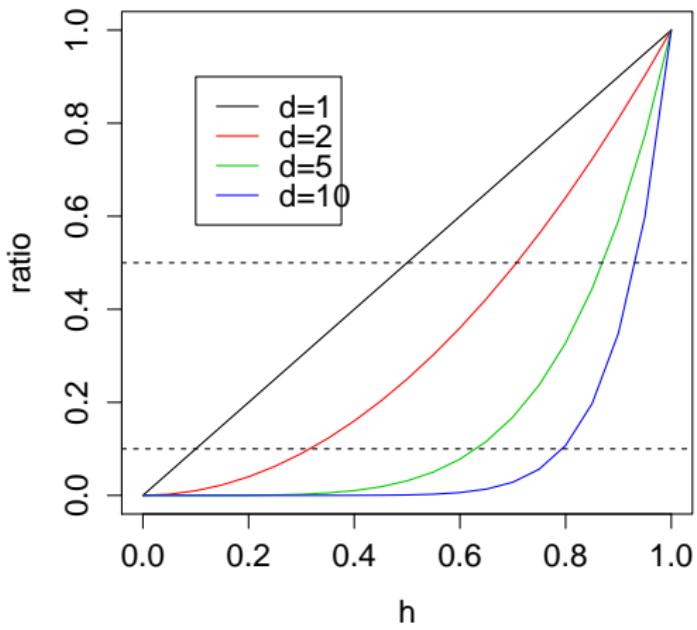
Compromis biais - variance

Le fléau de la dimension (cruse of dimensionality).

Considérons l'hypercube de côté 1 et de dimension d : $D = [0, 1]^d$. Considérons un sous-hypercube inclus dans D de côté $h < 1$ centré sur le centre du domaine et qui contienne $x\%$ du volume du domaine (par exemple 5%, 10%, 25%...).

Quelle est la valeur de h en fonction de la dimension d ?

curse of dimensionality



Plan

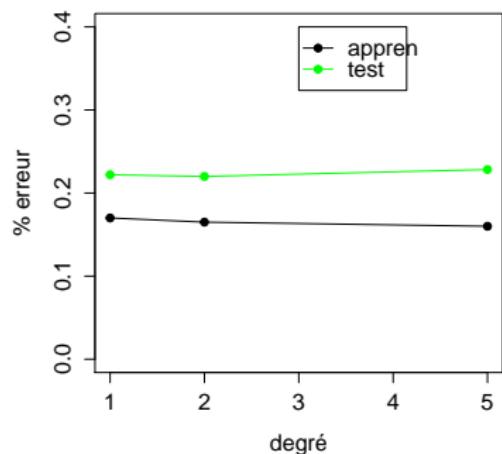
Deux approches simples : moindres carrés et k plus proches voisins

Quelques éléments théoriques

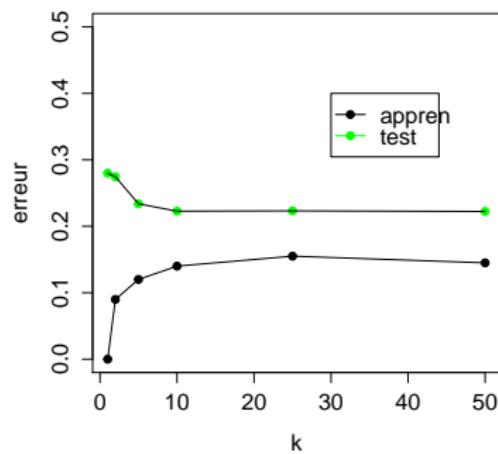
Méthodes locales en grande dimension

Compromis biais - variance

Scenario 1 : Régression

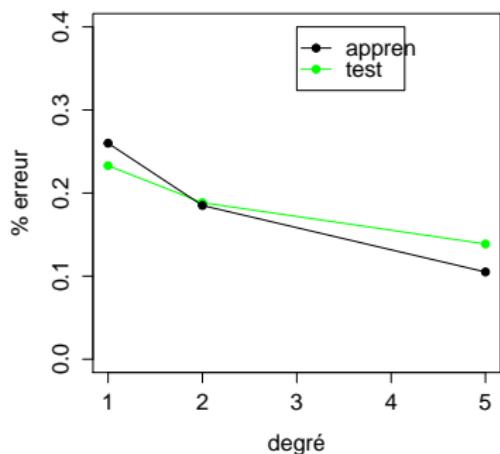


KNN

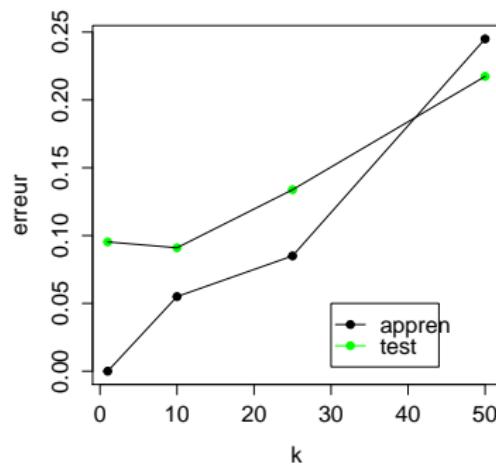


Scenario 2 :

Régression



KNN



Ainsi les méthodes de type k plus proches voisins qui sont des estimateurs directs de la quantité $f(x) = E(Y|X = x)$ posent des problèmes :

- ▶ si la dimension de l'espace des entrées est trop grande, les voisins sont alors très éloignés du point en question,
- ▶ si une structure particulière existe, cette information doit être utilisée pour réduire le biais et la variance.

L'objectif du cours est :

- ▶ de faire un tour d'horizon de diverses méthodes d'apprentissage pour la régression ou pour la classification :
 - ▶ basées sur des hypothèses de modèle (structure) ou non
 - ▶ adaptées à la grande dimension.
- ▶ de donner des outils pour sélectionner le bon paramètre de régularisation (degré pour la régression, k pour les plus proches voisins, ...) => meilleur compromis biais - variance.