Cours 4
Méthode de classification non supervisée
Régression

Méthode de classification non supervisée

Méthode des k-moyennes (K-means)

But :
On dispose d'un ensemble d'exemples $\boldsymbol{x}$ **non étiqueté** que l'on souhaite regrouper en ensembles homogènes

Méthode des k-moyennes : on connait a priori le nombre K d'ensembles

Méthode non supervisée, itérative :

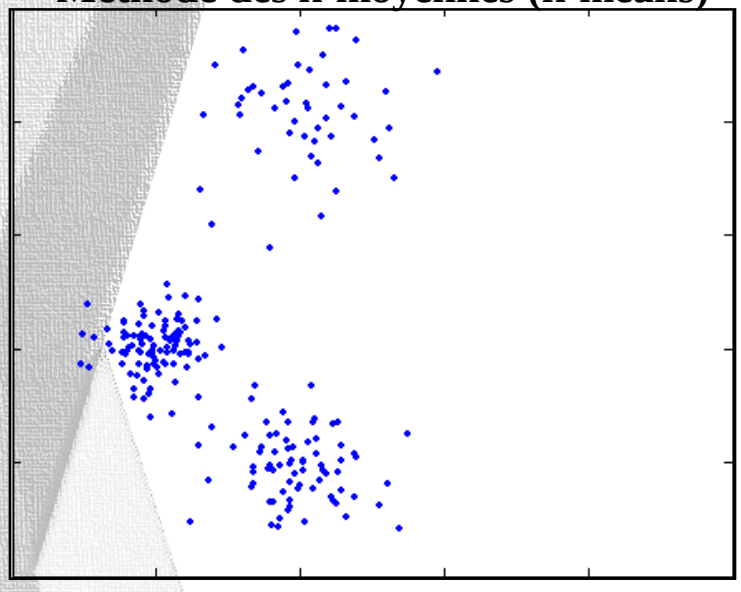
1. Initialisation : affecter chaque exemple $\boldsymbol{x}$ à un des K prototypes aléatoirement
2. Calculer les nouveaux prototypes : moyenne des exemples affectés à ce prototype
3. Affecter chaque exemple $\boldsymbol{x}$ au prototype le plus proche

Retour en 2 si pas idempotence

→ Pb : nombre de prototypes optimaux ?

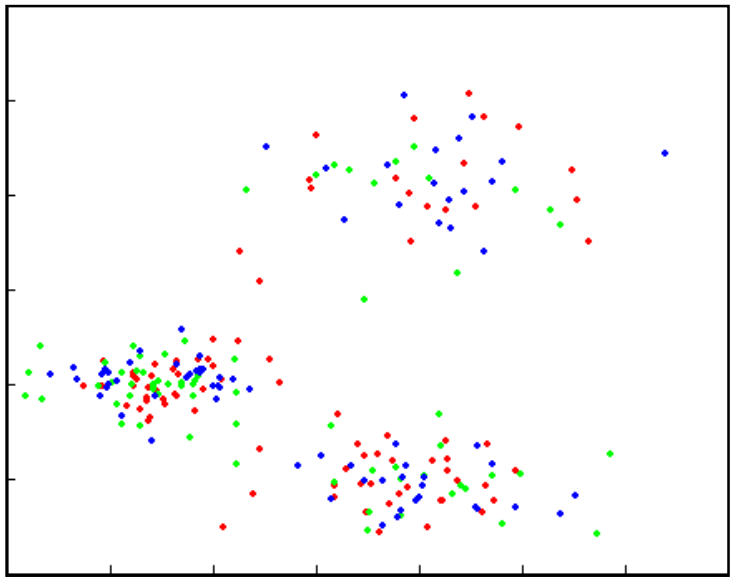
Méthode de classification non supervisée

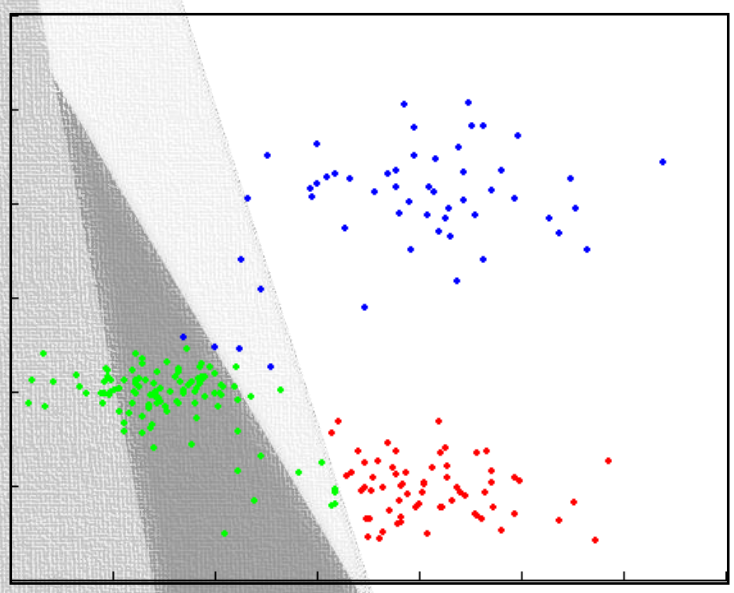
Méthode des k-moyennes (k-means)



it=0

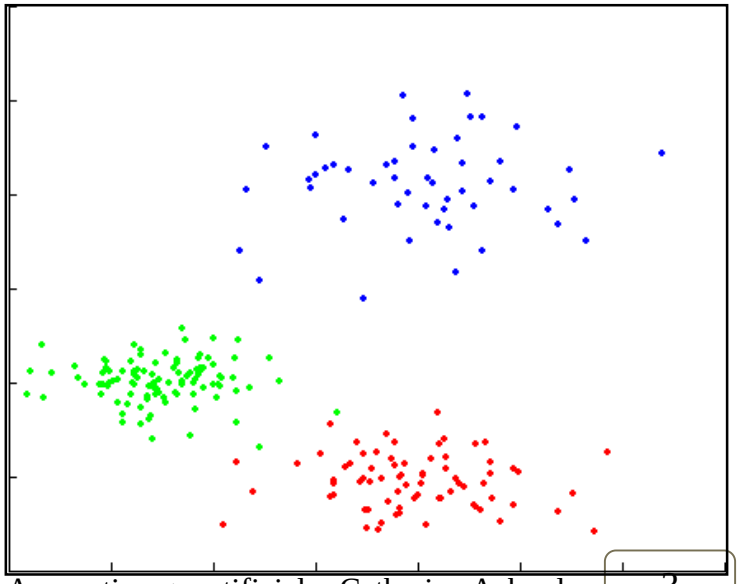
it=1





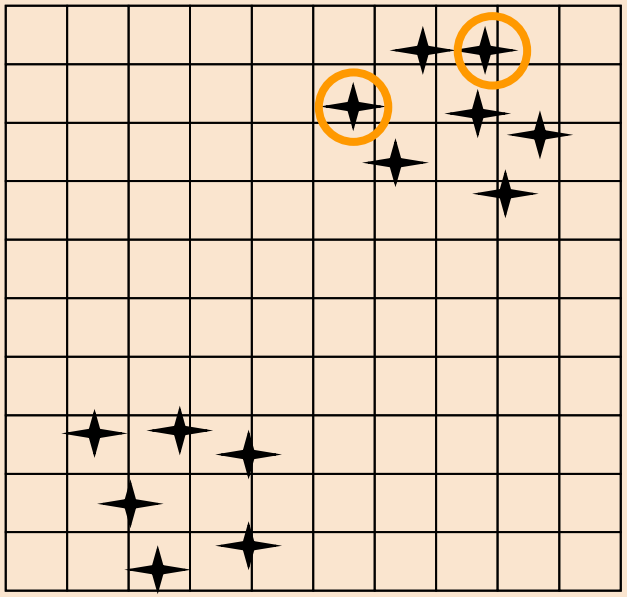
it=2

it=3



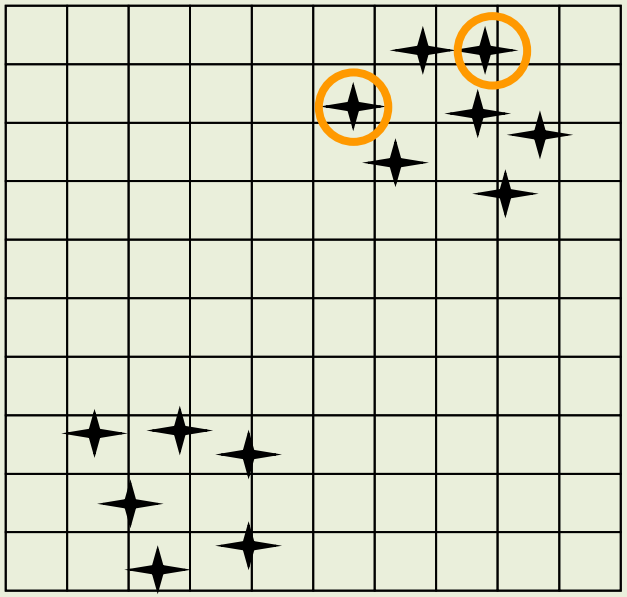
Méthode de classification non supervisée

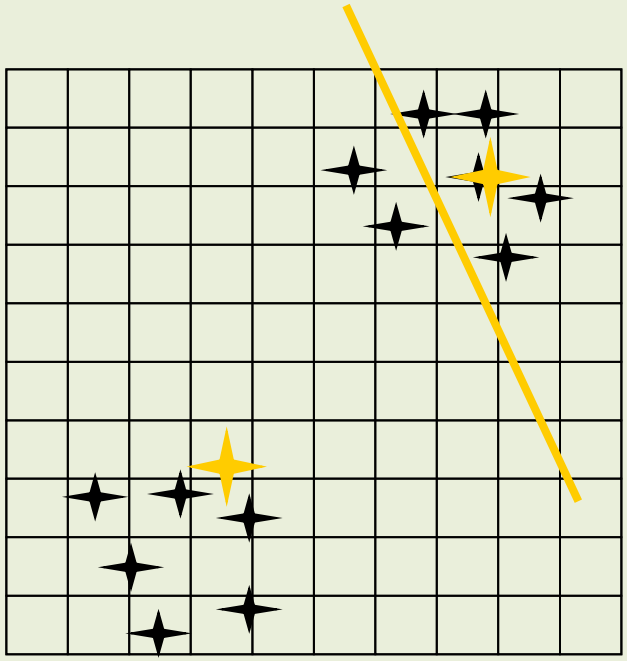
Exercice
On donne l'ensemble des exemples ci-contre. Représenter les centres obtenus avec la méthode des kmeans en initialisant les centres sur les points ci-contre

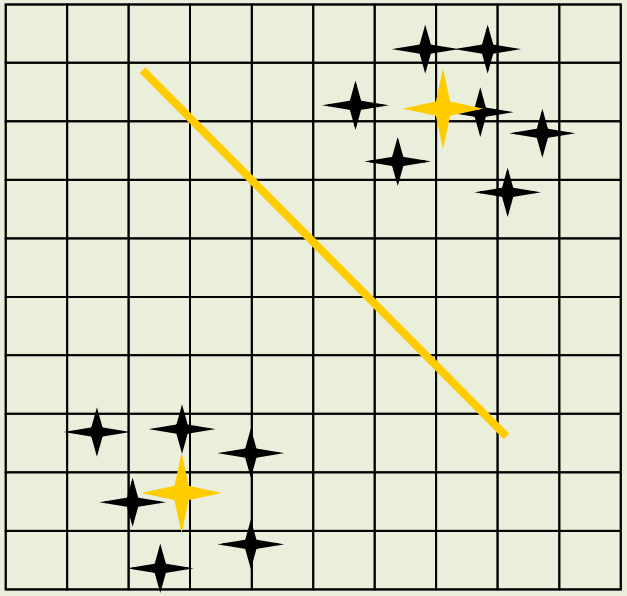


Méthode de classification non supervisée

Exercice
On donne l'ensemble des exemples ci-contre. Représenter les centres obtenus avec la méthode des kmeans en initialisant les centres sur les points ci-contre





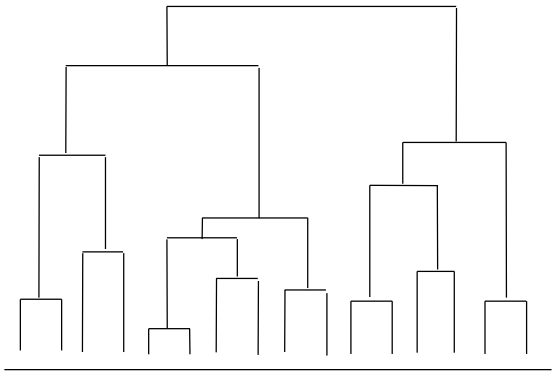


Méthode de classification non supervisée

Dendrogramme

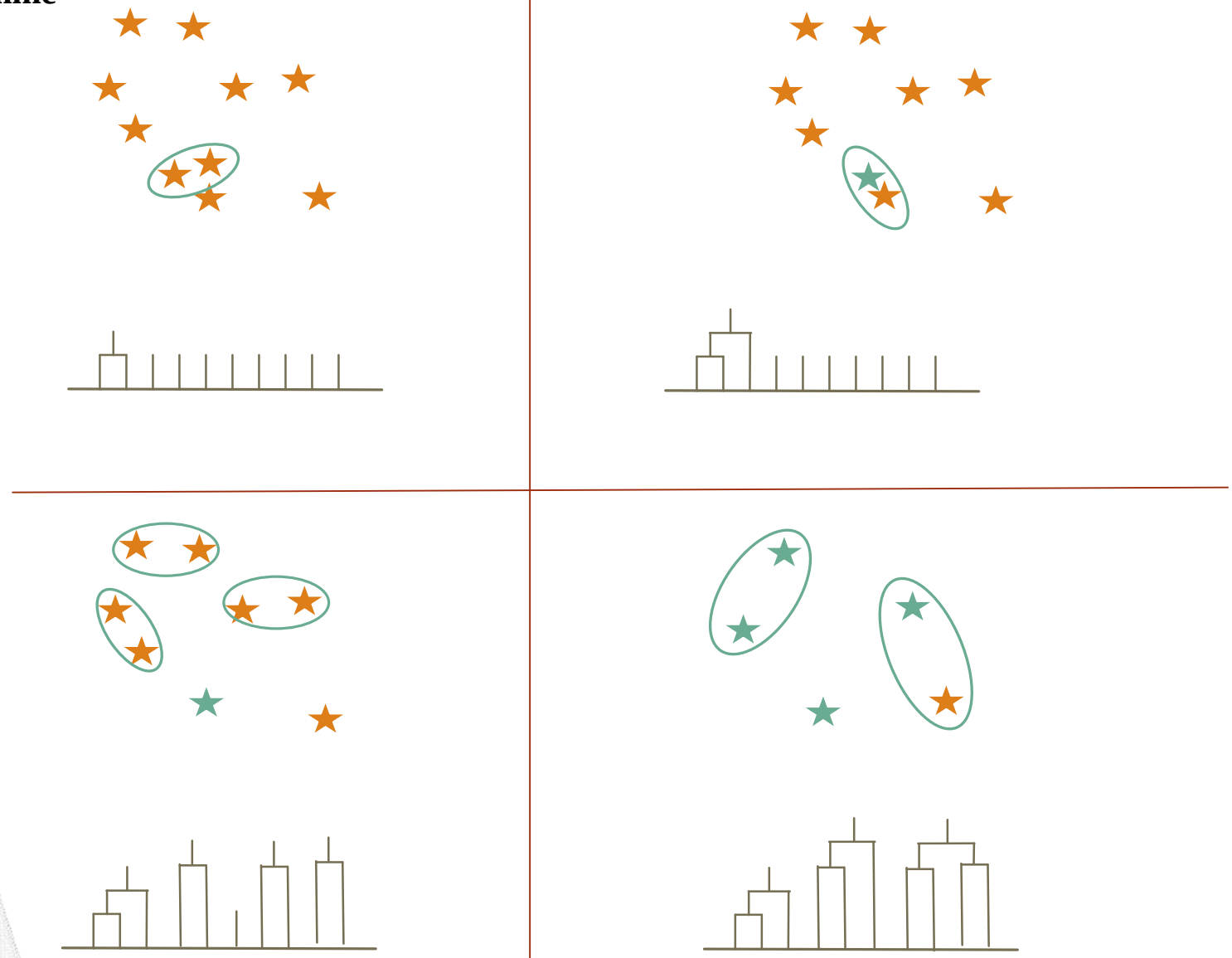
Classification ascendante hiérarchique

→ Regroupement des données suivant un critère de distance



Méthode de classification non supervisée

Dendrogramme



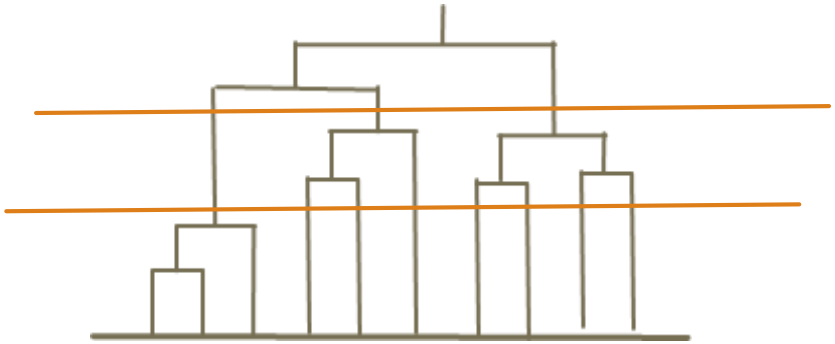
Méthode de classification non supervisée

Dendrogramme

→ Détermination des prototypes : coupure dans la hiérarchie

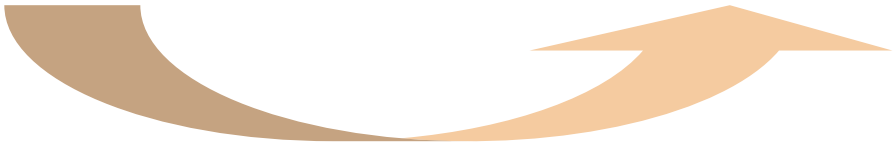
3 prototypes

8 prototypes



ESPACE DE REPRESENTATION OU ESPACE DES PARAMETRES

ESPACE DE DECISION CONTINU



Régression

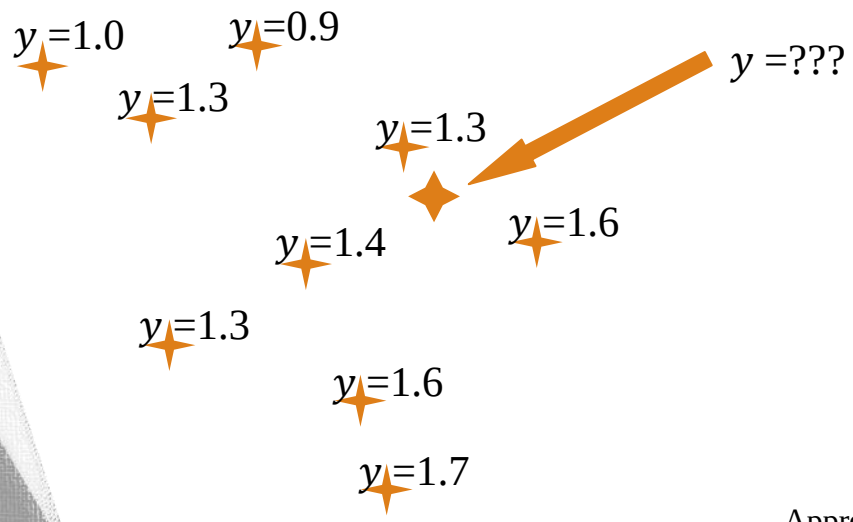
Régression avec les kppv

Entrée :
Une base d'exemple $\boldsymbol{x}$ avec vérité de terrain y (vecteur de codage+ variable continue y)

Sortie :
Pour un exemple $\boldsymbol{x}$ inconnu, estimer la variable y.

Méthode :
On compare $\boldsymbol{x}$ aux exemples de la base.
L'estimation associée à $\boldsymbol{x}$ est la valeur y de l'exemple le plus proche.

Variante :
On garde les k exemples les plus proches
On interpole avec une moyenne ou une médiane des valeurs de y de ces exemples



Régression linéaire

But :
Construire une fonction $y = \boldsymbol{f(x)}$

Données :
Les N exemples d'entrée $\boldsymbol{x_i}$ de dimension n.
Les N variable de sortie y_i continues et de dimension 1

Modélisation linéaire : **la sortie est un combinaison des n variables d'entrée :**

$$y_i = w_0 + w_1 x_i(1) + w_2 x_i(2) + \ldots + w_n x_i(n) + \epsilon_i$$

Où les ϵ_i correspondent aux erreurs de modélisation

En construisant
- la matrice $\boldsymbol{X}$, de dimension Nx(n+1) tq la première colonne contient le vecteur $\mathbf{1}$ et les exemples $\boldsymbol{x}_i$ sont rangés en ligne
- Le vecteur $\boldsymbol{y}$ composé des N valeurs y_i
- Le vecteur des poids $\boldsymbol{w}$

Le système se met sous la forme :

$$\boldsymbol{y = Xw + \epsilon}$$

Où $\boldsymbol{\epsilon}$ représente le bruit

Régression linéaire

On cherche le vecteur des poids $\boldsymbol{w}$ qui minimise l'erreur :

$$E(\boldsymbol{w}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2$$

Rappel sur les matrices

$(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$

$(\boldsymbol{A} + \boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T$

$\frac{\partial a^T b}{\partial a} = \frac{\partial b^T a}{\partial a}$ = b

$\frac{\partial a^T \boldsymbol{B} a}{\partial a}$=2**B**a

Exercice : Déterminer l'expression de $\boldsymbol{w}$

Régression linéaire

Rappel sur les matrices
$(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$
$(\boldsymbol{A}+\boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T$
$\frac{\partial a^T b}{\partial a} = \frac{\partial b^T a}{\partial a} = \mathrm{b}$
$\frac{\partial a^T \boldsymbol{B} a}{\partial a} = 2\mathbf{B}\mathrm{a}$

On cherche le vecteur des poids $\boldsymbol{w}$ qui minimise l'erreur :

$$E(\boldsymbol{w}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|$$

Exercice : Déterminer l'expression de $\boldsymbol{w}$

$$E(w) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\| = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}$$

En dérivant l'erreur par rapport à w et annulant les dérivées, on a:

$$\frac{\partial E(\boldsymbol{w})}{\partial \boldsymbol{w}} = -\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = 0$$
$$-\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} = 0$$

$$\widehat{\boldsymbol{w}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^{\dagger}\,\boldsymbol{y}$$

Où $\boldsymbol{X}^{\dagger}$ est la pseudo inverse de $\boldsymbol{X}$

Régression linéaire

Une fois les poids estimés $\widehat{\boldsymbol{w}}$, la sortie estimée $\widehat{\boldsymbol{y}}$ pour toute entrée $\boldsymbol{x}$ s'obtient par:

$$\widehat{\boldsymbol{y}} = \boldsymbol{x}^T \widehat{\boldsymbol{w}}$$

La **qualité de la régression** peut se mesurer par:

- SSE (sum squared error) = $\sum_i \|y_i - \widehat{y_i}\|^2$
- Le coefficient de détermination $r^2 = \frac{\sum_i (y_i - \widehat{y_i})^2}{\sum_i (y_i - \overline{y})^2} = \frac{\sum_i (\widehat{y_i} - \overline{y})^2}{\sum_i (y_i - \overline{y})^2}$ où $\overline{\boldsymbol{y}}$ est la moyenne des y_i

  $r^2 \in [0,1]$ *et* r^2 proche de 1 quand la modélisation est de bonne qualité

Régression linéaire

Exemple:
On souhaite prédire la consommation électrique au mois de janvier pour des logements équipés du tout électrique.
Réaliser une régression avec chaque variable indépendamment et estimer le SSE (leave one out). Conclusion
Réaliser une régression avec toutes les variables et calculer le SSE. Conclusion
Réaliser une régression en utilisant Surface et volume puis une autre avec surface et personne. Conclusion

A l'aide du modèle utilisant toutes les variables donner une estimation pour la consommation d'un pavillon d'une surface de 150 m2, habité par 4 personnes, construit il y a 18 ans, comprenant deux salles de bains et dont le volume intérieur est de 405 m3

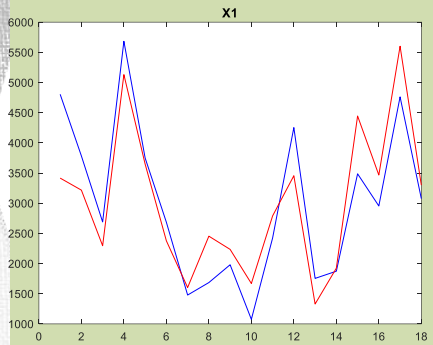
- KW: Nombre de KWH consommés pendant le mois de janvier
- SURFACE: Surface du logement en m2
- PERS: Nombre de personnes habitant le logement
- PAVILLON: Pavillon codé 1; Appartement codé 0
- AGE: Age du logement
- VOL: Volume intérieur du logement en m3
- SBAINS: Nombre de salles de bains

KW	SURFACE	PERS	PAVILLON	AGE	VOL	SBAINS
4805	130	4	1	65	410	1
3783	123	4	1	5	307	2
2689	98	3	0	18	254	1
5683	178	6	1	77	570	3
3750	134	4	1	5	335	2
2684	100	4	0	34	280	1
1478	78	3	0	7	180	1
1685	100	4	0	10	250	1
1980	95	3	0	8	237	1
1075	78	4	0	5	180	1
2423	110	5	1	12	286	1
4253	130	4	1	25	351	1
1754	73	2	0	56	220	1
1873	87	4	1	2	217	2
3487	152	5	1	12	400	2
2954	128	5	1	20	356	1
4762	180	7	1	27	520	2
3076	124	4	0	22	330	1

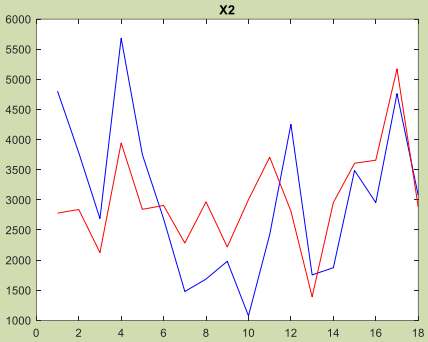
Régression linéaire

Régression avec chaque variable indépendamment

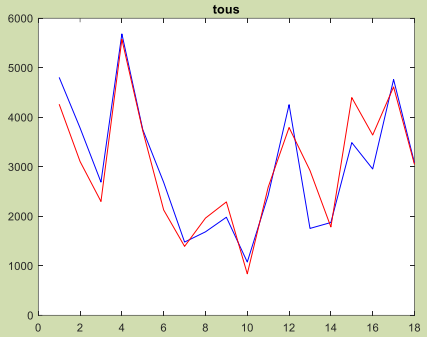
Résultat de prédiction avec **la surface** (rouge)



Résultat de prédiction avec le **nb de personnes**



Résultat de prédiction avec **toutes les variables**



	surface	personnes	pavillon	age	volume	sdbain	toutes
SSE (*10^3)	2,59	4,58	4,39	5,12	2,19	4,83	2,10

Régression avec Surface et Volume, SSE=2,38 10^3
Régression avec Volume et personne, SSE=2,18 10^3

Pour le modèle avec toutes les variables,
$\widehat{\boldsymbol{w}} = 1.0e+03 * (3.10, \quad 0.88, \quad -0.52, \quad 0.31, \quad 0.18, \quad 0.50, \quad -0.05)^T$
Avec les paramètres donnés, conso= $(1\ 150\ 4\ 1\ 18\ 405\ 2)^T * \widehat{\boldsymbol{w}}$= 4.6106e+03

Régression linéaire

Compromis biais/variance

➔ Ces critères (SSE et r) ne permettent pas une bonne sélection de modèles

Exemple en 1D :

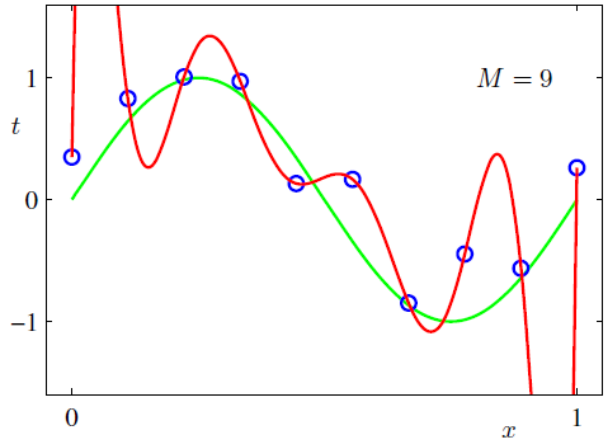


Image issue de Pattern Recognition and machine learning – M. Bishop – 2007

Le polynôme d'ordre 9 passe exactement par les 9 points de données

En vert, l'approximation idéale

Plus le modèle sera complexe (d'ordre élevé) et plus SSE et r^2 seront bons

2 solutions pour éviter « l'over-fitting » :

- Réduire la dimension des données (sélection de caractéristique, début du cours)
- Contraindre les paramètres de la régression

Ridge régression

Rappelons que nous avons formalisé le problème par :

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$$

La ridge régression minimise

$$E(\boldsymbol{w}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|^2$$

➔ On ajoute une **pénalité** ($\lambda\|\boldsymbol{w}\|^2$) qui évite aux coefficient w_i de prendre des valeurs trop grandes

➔ On pénalise plus ou moins en fonction de λ

➔ λ est optimisé par validation croisée

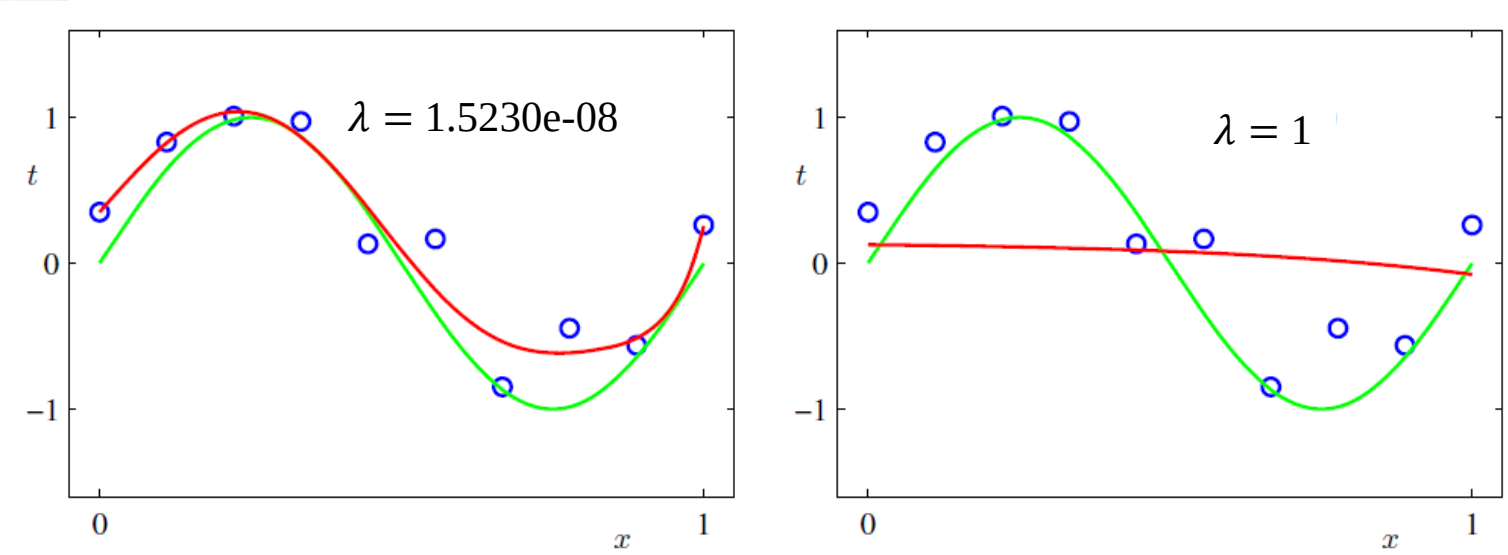


Image issue de Pattern Recognition and machine learning – M. Bishop – 2007

Régression LASSO

La régresion LASSO minimise :

$$E(\boldsymbol{w}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2 + \lambda\|\boldsymbol{w}\|_1$$

Où

- $L'opérateur \|\boldsymbol{w}\|_1$ représente la norme $\ell 1$: $\|\boldsymbol{w}\|_1 = \sum_{i=1}^{n} |w_i|$

➔ On ajoute une **pénalité** ($\lambda\|\boldsymbol{w}\|_1$) qui permet d'avoir une **représentation parcimonieuse** (avec beaucoup de coefficients w_i nuls)

➔ **On fait en même temps la régression et la sélection**

➔ Plus λ est grand, plus il y a de coefficients w_i nuls

➔ Comme la norme $\ell 1$ n'est pas différentiable, pas de solution analytique, algorithme de minimisation

➔ λ est optimisé par validation croisée