

Cours 4

Méthode de classification non supervisée

Régression

Méthode des k-moyennes (K-means)

But :

On dispose d'un ensemble d'exemples x non étiquetés que l'on souhaite regrouper en K ensembles homogènes. Le nombre K d'ensembles est connu a priori.

Méthode non supervisée, itérative :

1. Initialisation : affecter chaque exemple x à un des K prototypes aléatoirement
2. Calculer les nouveaux prototypes : moyenne des exemples affectés à ce prototype
3. Affecter chaque exemple x au prototype le plus proche
4. Retour en 2 si pas idempotence, sinon fin

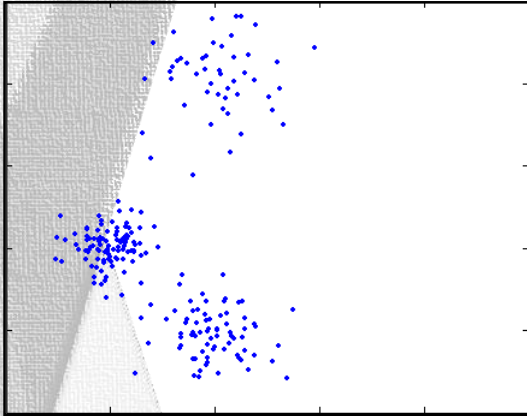
→ Pb : nombre de prototypes optimaux ?

Rq : On peut aussi partir avec une autre initialisation :

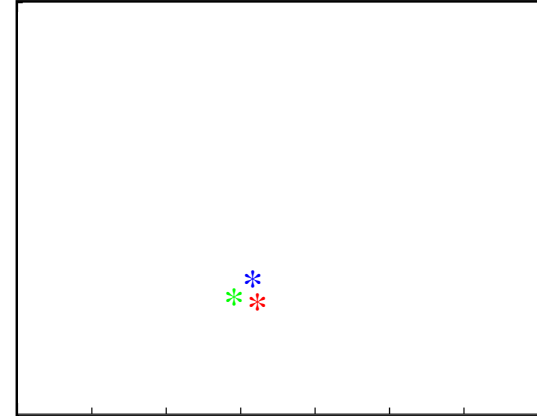
1. Initialisation : on tire K prototypes aléatoirement parmi les exemples
2. Affecter chaque exemple x au prototype le plus proche
3. Calculer les nouveaux prototypes : moyenne des exemples affectés à ce prototype
4. Retour en 2 si pas idempotence, sinon fin

Méthode des K-moyennes (k-means), K=3

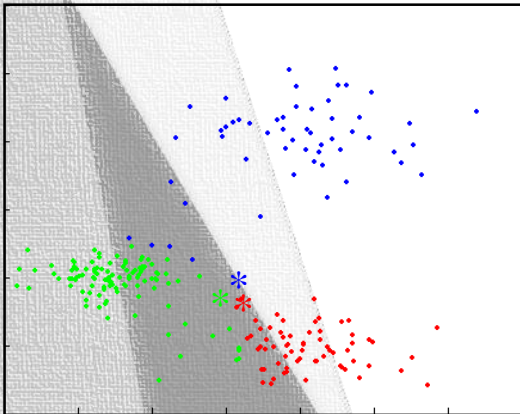
affecter chaque exemple x à un
des K prototypes aléatoirement



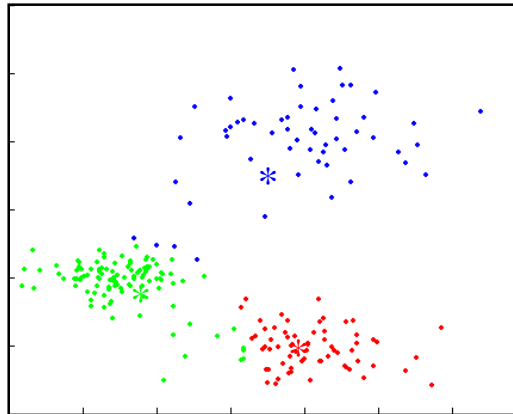
Calcule des prototypes :
moyenne de chaque ensemble



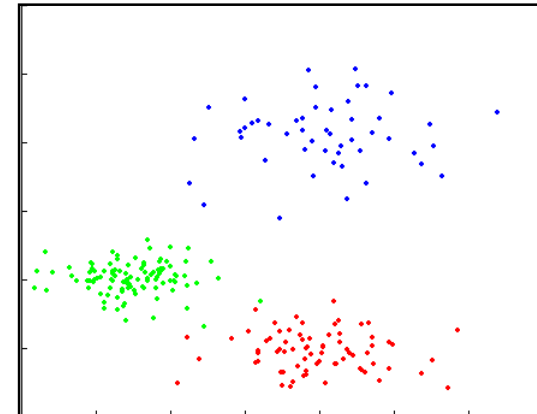
Affecter chaque exemple x au
prototype le plus proche



Calcule des prototypes :
moyenne de chaque ensemble

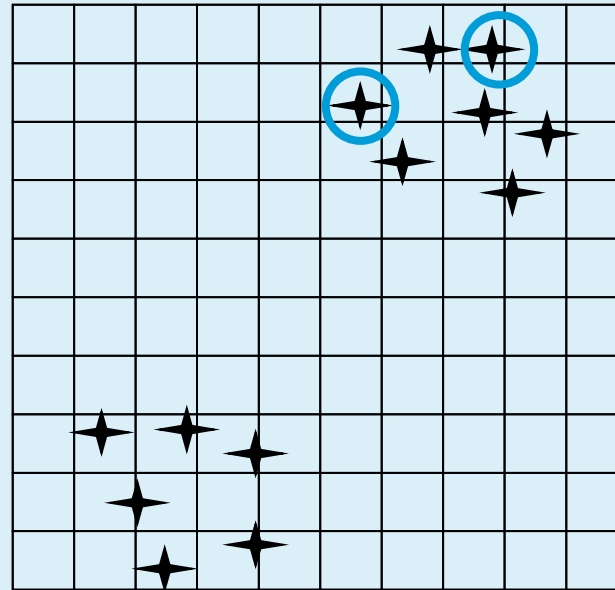


Affecter chaque exemple x au
prototype le plus proche



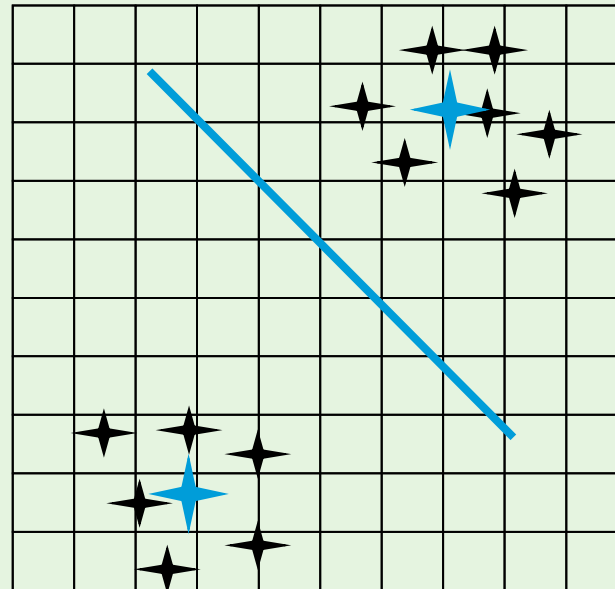
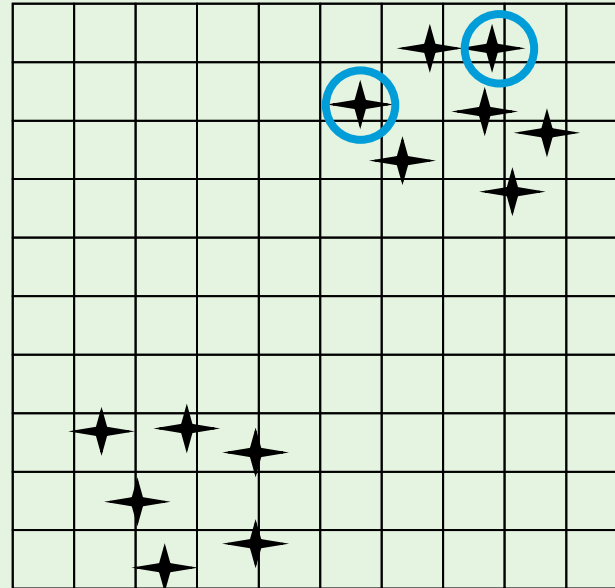
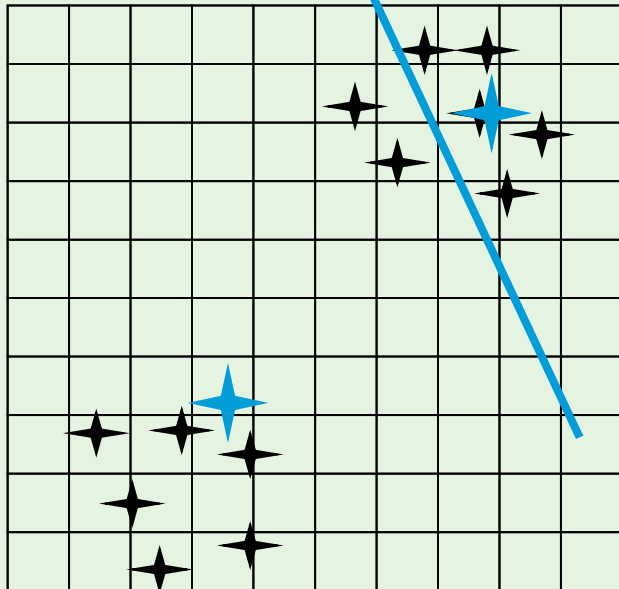
Exercice

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



Exercice

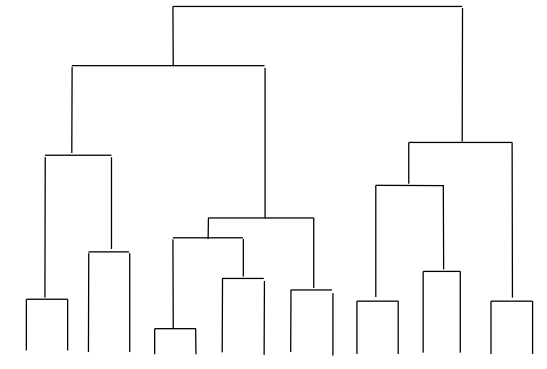
On donne l'ensemble des exemples ci-contre. Représenter les centres obtenus avec la méthode des kmeans en initialisant les centres sur les points ci-contre



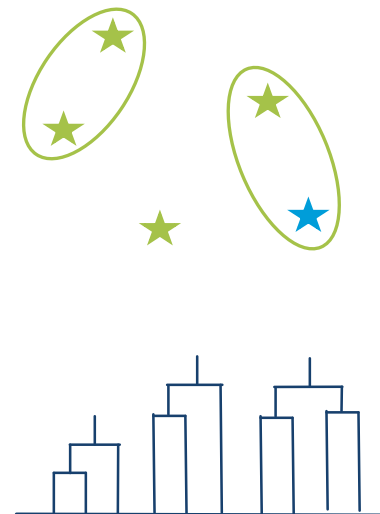
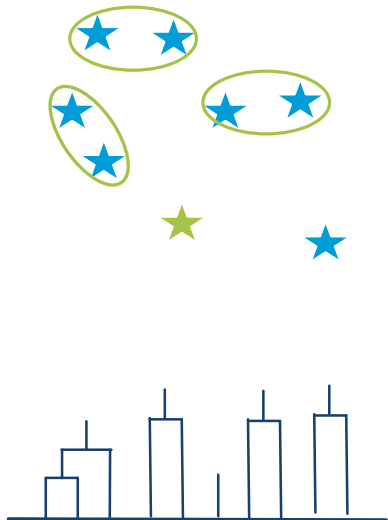
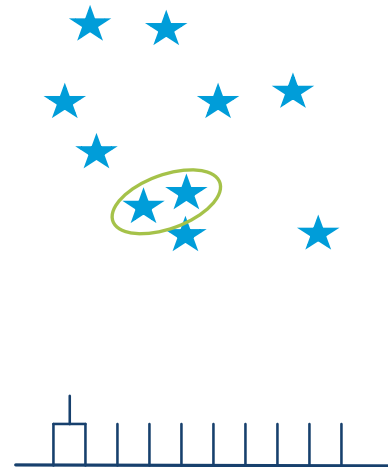
Dendrogramme

Classification ascendante hiérarchique

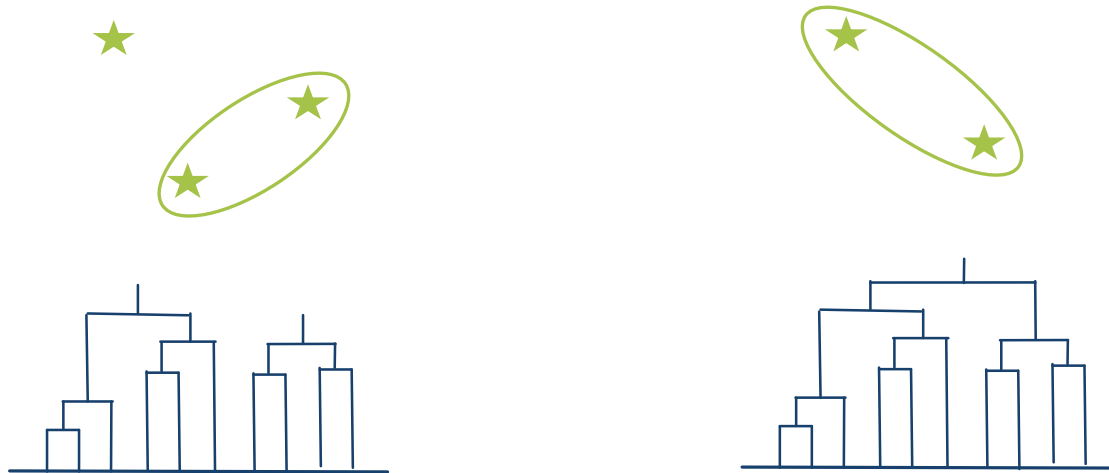
→ Regroupement des données suivant un critère de distance



Dendrogramme



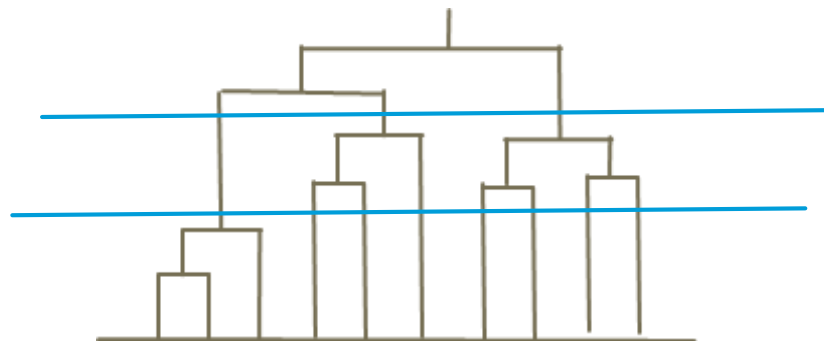
Dendrogramme



→ Détermination des prototypes : coupure dans la hiérarchie

3 prototypes

8 prototypes



**ESPACE DE
REPRESENTATION
OU
ESPACE DES PARAMETRES**

**ESPACE DE DECISION
CONTINU**



Régression

Avec les kppV

Entrée :

Une base d'exemple x avec vérité de terrain y (vecteur de codage+ variable continue y)

Sortie :

Pour un exemple x inconnu, estimer la variable y .

Méthode :

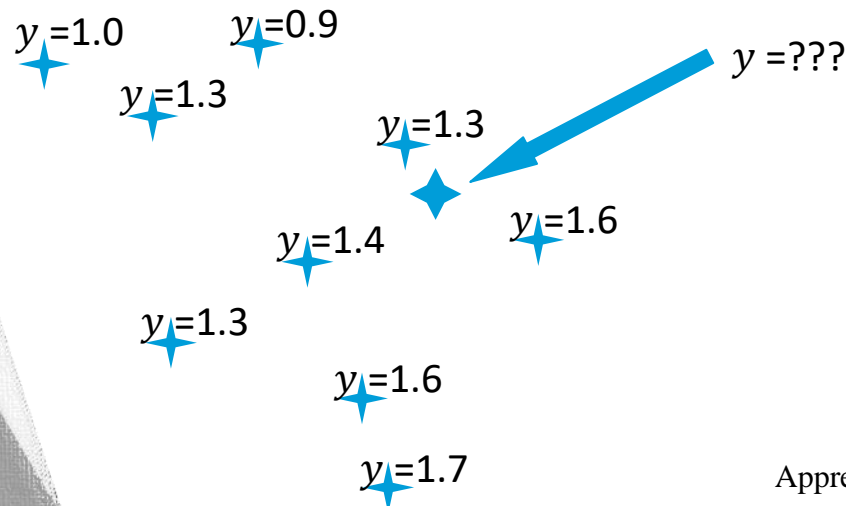
On compare x aux exemples de la base.

L'estimation associée à 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 $y = f(x)$

Données :

Les N exemples d'entrée x_i de dimension n .

Les N variables de sortie y_i continues et de dimension 1

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

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

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

Où les ϵ_i correspondent aux erreurs de modélisation

Avec:

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

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

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

Rappel sur les matrices

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

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

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

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

Exercice : Déterminer l'expression de \mathbf{w}

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

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

Exercice : Déterminer l'expression de \mathbf{w}

Rappel sur les matrices

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

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

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

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

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

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

$$\begin{aligned} \frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} &= -\mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \mathbf{w} = 0 \\ -\mathbf{X}^T \mathbf{y} + \mathbf{X}^T \mathbf{X} \mathbf{w} &= 0 \end{aligned}$$

$$\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{X}^\dagger \mathbf{y}$$

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

Une fois les poids estimés avec $\hat{\mathbf{w}} = \mathbf{X}^{\dagger} \mathbf{y}$, la sortie estimée $\hat{\mathbf{y}}$ pour toute entrée \mathbf{x} s'obtient par:

$$\hat{\mathbf{y}} = \mathbf{x}^T \hat{\mathbf{w}}$$

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

- SSE (sum squared error) = $\sum_i \|y_i - \hat{y}_i\|^2$
- Le coefficient de détermination $r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$ où \bar{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

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 avec chaque variable indépendamment

Commençons par la surface. On modélise le problème par :

$$KW_i = w_0 + w_1 SURFACE_i + \epsilon_i \text{ soit } \mathbf{y} = \mathbf{X} \mathbf{w} + \epsilon_i$$

En prenant le dernier exemple en test (on le supprime lors du calcul des poids)

$\mathbf{y} =$	4805	$\mathbf{X} =$	1	130
	3783		1	123
	2689		1	98
	5683		1	178
	3750		1	134
	2684		1	100
	1478		1	78
	1685		1	100
	1980		1	95
	1075		1	78
	2423		1	110
	4253		1	130
	1754		1	73
	1873		1	87
	3487		1	152
	2954		1	128
	4762		1	180

$$\mathbf{w} = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix}$$

Et on prédit les poids avec

$$\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{X}^\dagger \mathbf{y}$$

On trouve :

$$\hat{\mathbf{w}} = \begin{pmatrix} -1.2988 \\ 0.037 \end{pmatrix} 10^3 \quad 1$$

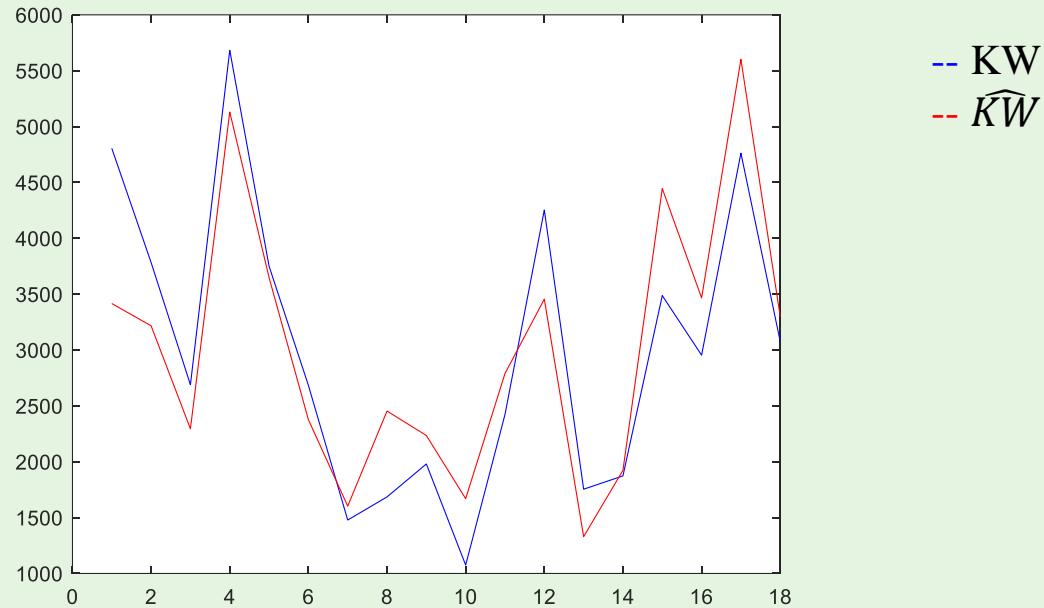
Et en testant sur le dernier exemple, on a :

$$\widehat{KW} = (1 \ 124) \begin{pmatrix} -1.2988 \\ 0.0371 \end{pmatrix} 10^3 = 3299$$

Alors que la vraie valeur est 3076

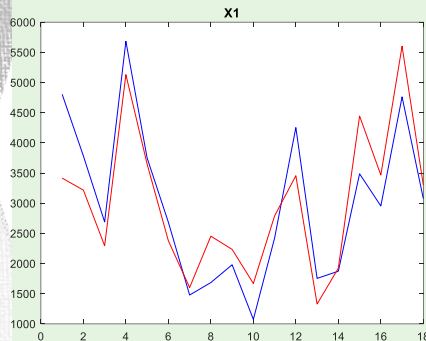
On passe chaque exemple en test à tour de rôle

Courbe obtenue en passant chacun des 18 exemples en test :

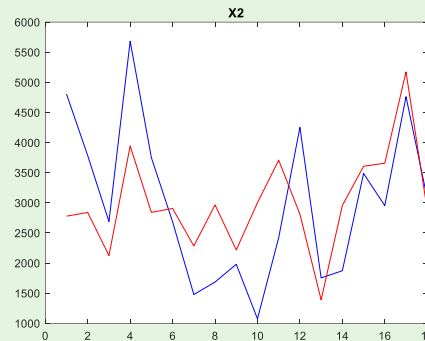


Régression avec chaque variable indépendamment

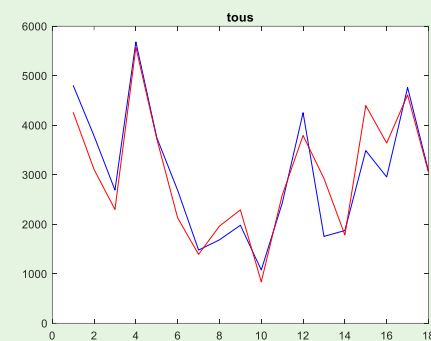
Résultat de prédiction
avec la **surface**



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



Résultat de prédiction
avec **toutes les variables**



	surface	personnes	pavillon	age	volume	sdbain	toutes
SSE (*10 ⁷)	0.67	2.10	1.93	2.62	0.48	2.34	0.44

Régression avec Surface et Volume, SSE= 0.57 10⁶

Régression avec Volume et personne, SSE=0.4 10⁶

Pour le modèle avec toutes les variables,

$$\hat{\mathbf{w}} = 1.0e + 03 * (3.10, \quad 0.88, \quad -0.52, \quad 0.31, \quad 0.18, \quad 0.50, \quad -0.05)^T$$

Avec les paramètres donnés, conso= (1 150 4 1 18 405 2)^T * $\hat{\mathbf{w}}$ = 4.6106e+03

Compromis biais/variance

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

Exemple en 1D :

On souhaite régresser une courbe passant par les points 1D suivants (en vert, la courbe idéale).
Pour cela, on formalise le problème avec :

$$y = w_0 + w_1x + w_2x^2 + \dots + w_Nx^N$$

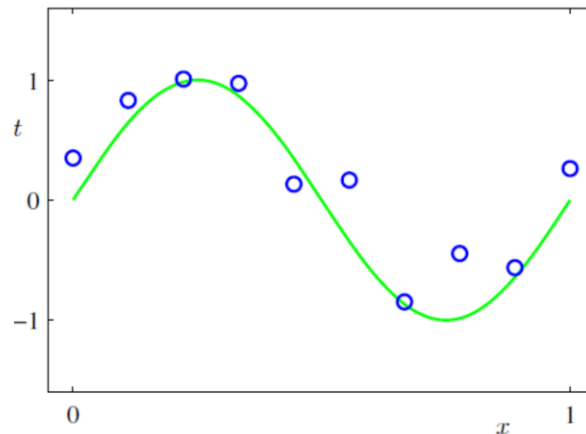
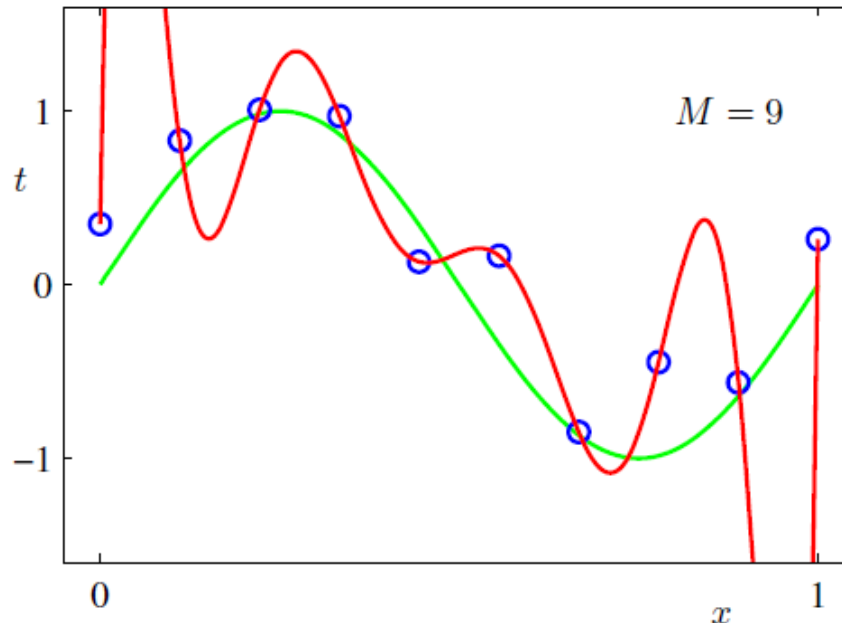


Image issue de Pattern Recognition and machine learning – M.
Bishop – 2007

Le polynôme d'ordre 10 passe exactement par les 10 points de données (courbe rouge)

→ il conduit à $SSE=0$ et $r=1$



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

Solutions pour éviter « l'over-fitting » :

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

Rappelons que nous avons formalisé le problème par :

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

La ridge régression minimise

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

- On ajoute une **pénalité** ($\lambda\|\mathbf{w}\|^2$) 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

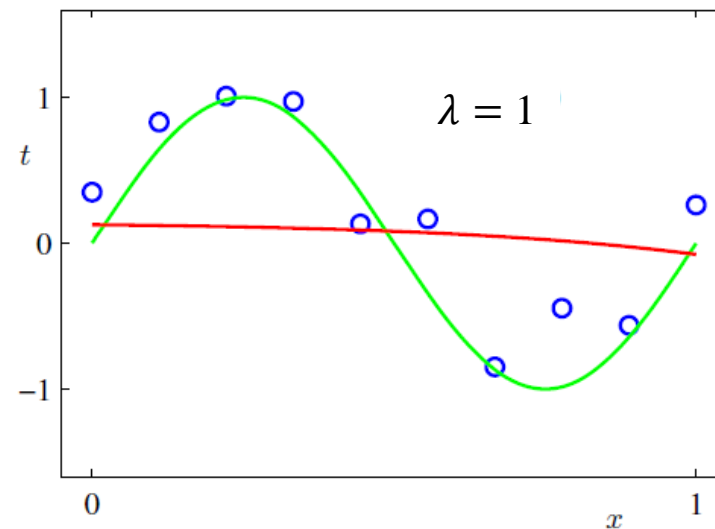
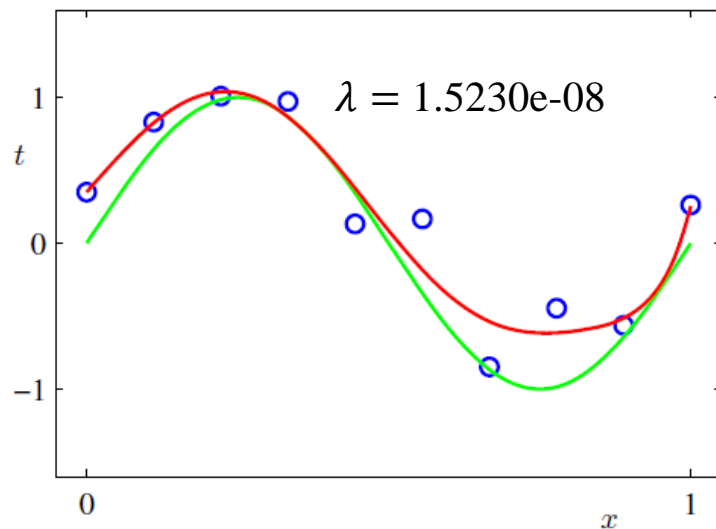


Image issue de Pattern Recognition and machine learning – M. Bishop – 2007

La régression LASSO minimise :

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

Où

- L'opérateur $\|\mathbf{w}\|_1$ représente la norme ℓ_1 : $\|\mathbf{w}\|_1 = \sum_{i=1}^n |w_i|$
- On ajoute une **pénalité** ($\lambda\|\mathbf{w}\|_1$) 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 ℓ_1 n'est pas différentiable, pas de solution analytique, algorithme de minimisation
- λ est optimisé par validation croisée