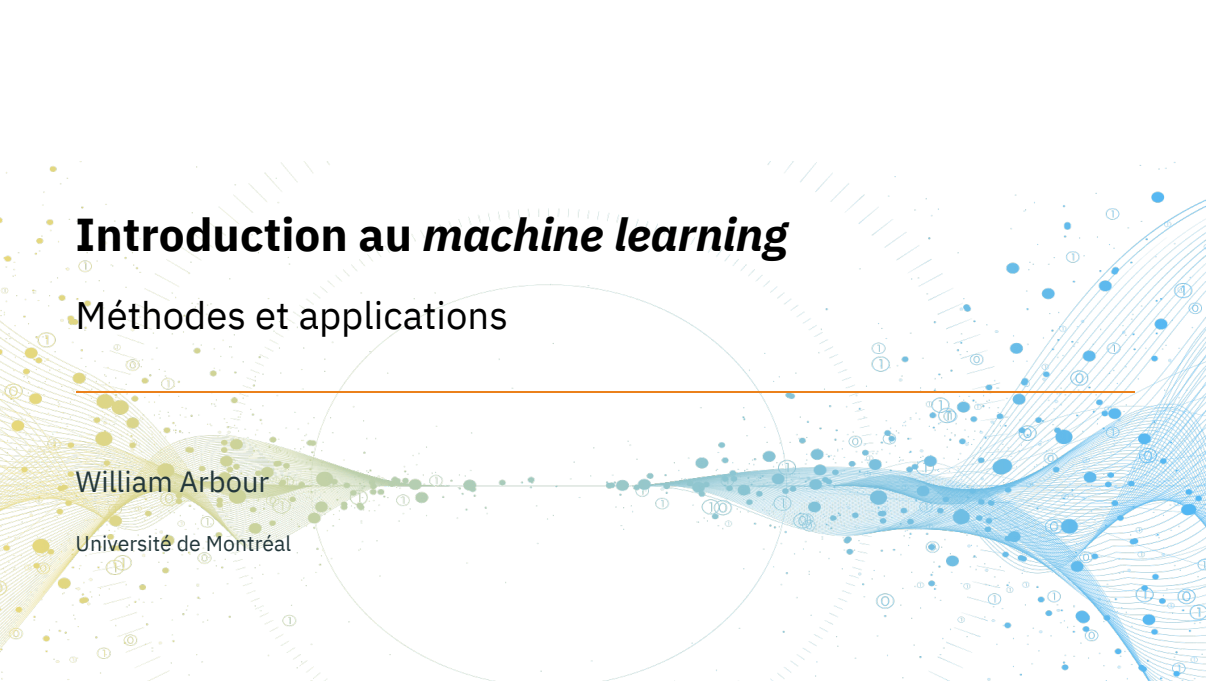
Introduction au *machine learning*

Méthodes et applications

William Arbour

Université de Montréal



Plan de la formation

Présentation

Le *machine learning*, c'est quoi?

Méthodes non-supervisées

Méthodes supervisées paramétriques

Méthodes supervisées non-paramétriques

Conclusion

Présentation

Présentation de l’enseignant

2017	Baccalauréat en économie et mathématiques (Laval)
2018	Maîtrise en économie (Laval)
2023	Doctorat en économie (Toronto)
2023–	Professeur adjoint à l’Université de Montréal

Présentation de l’enseignant

Thèmes d’enseignement:

- Économie de la santé, Évaluation de programmes, Économie du crime, Économétrie appliquée, Science des données

Thèmes de recherche:

- Économie du crime, Économie de l’éducation, Économie du travail

Le *machine learning*, c’est quoi?

Définition

Machine learning (apprentissage-machine): séries d'algorithmes qui permettent d'expliquer des relations entre des variables

But: à partir d'intrants, formuler des prédictions d'extrants

$$extrant = f(intrants) + \epsilon$$

f est une fonction qui change les intrants en un extrant

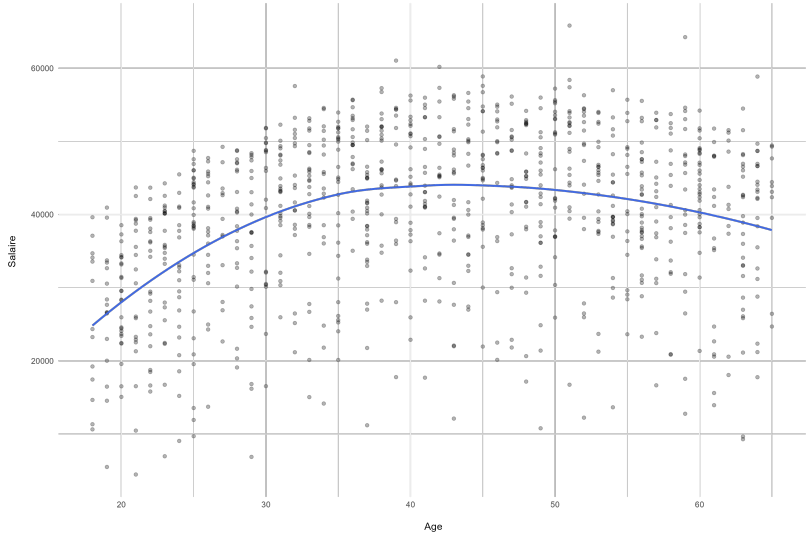
Estimer f

On s'intéresse à la relation entre le salaire, l'âge et l'éducation

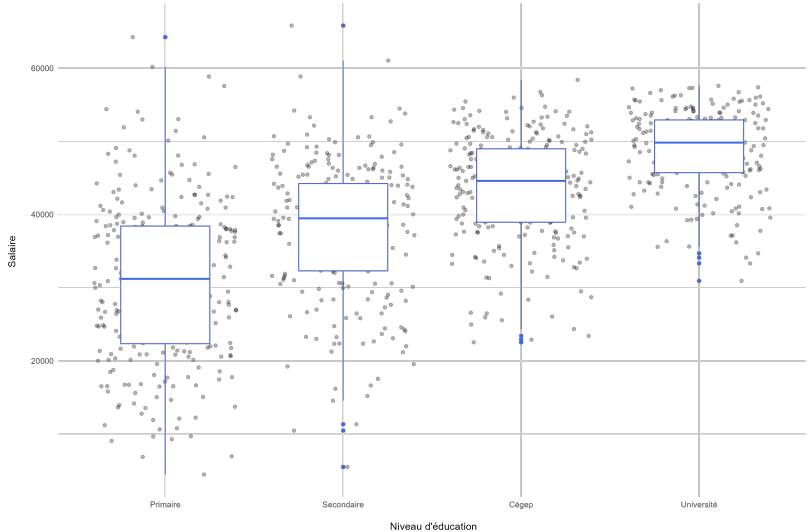
$$salaire = f(age, education) + \epsilon$$

Comment estimer f?

Estimer f



Estimer f


60000
40000
20000
Salaire
Primaire
Secondaire
Cégep
Université
Niveau d'éducation

Estimer f

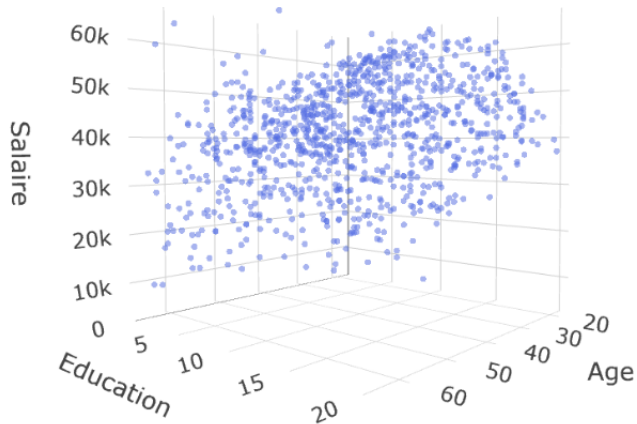
Donc le salaire dépend de l'âge et du niveau d'éducation!

On pourrait estimer

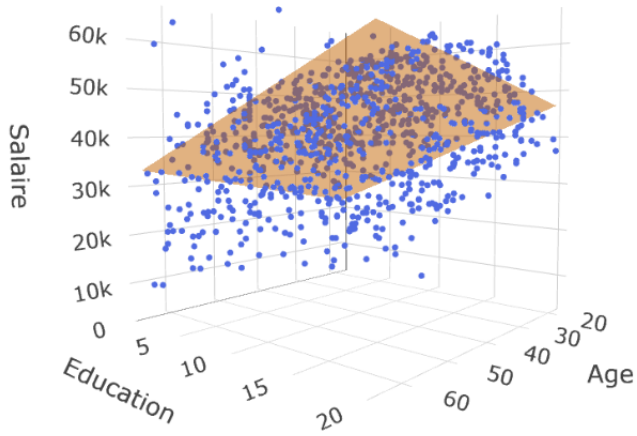
$$\begin{aligned} Salaire &= f(age, education) + \epsilon \\ &= \beta_0 + \beta_1 Age + \beta_2 Education + \epsilon \end{aligned}$$

Quellle est l'interprétation de β_1? Et β_2? Est-ce contraignant? Avons-nous besoin d'un terme d'interaction?

Estimer f


Salaire
60k
50k
40k
30k
20k
10k
0
Education
5
10
15
20
Age
20
30
40
50
60

Estimer f

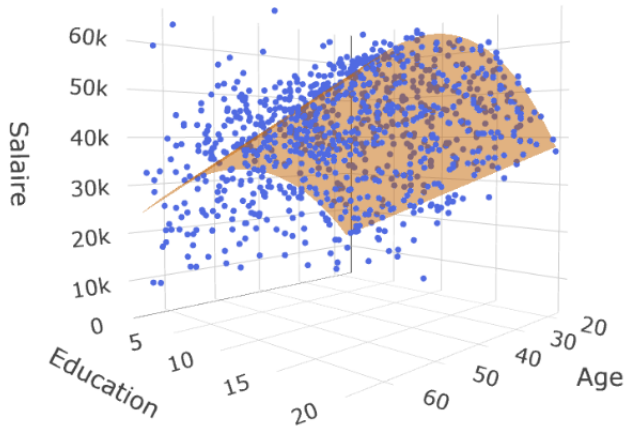

Salaire
60k
50k
40k
30k
20k
10k
0
Education
5
10
15
20
Age
20
30
40
50
60

Estimer f

Je pourrais rendre les prédictions plus *flexibles* en incorporant un terme d'interaction, ou bien en incluant un terme d'âge quadratique:

$$\begin{aligned} Salaire &= f(age, education) + \epsilon \\ &= \beta_0 + \beta_1 Age + \beta_2 Education + \beta_3 Age^2 + \epsilon \end{aligned}$$

Estimer f

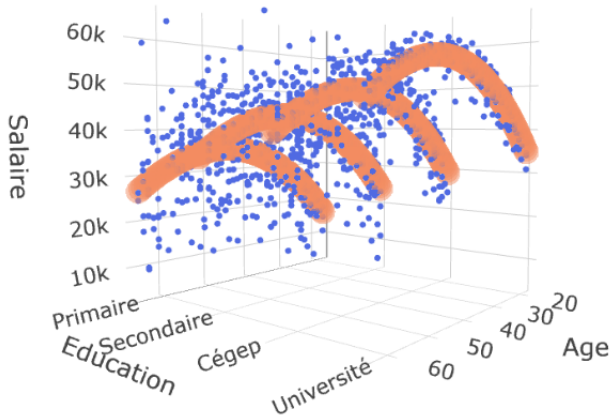

Salaire
60k
50k
40k
30k
20k
10k
0
Education
5
10
15
20
Age
20
30
40
50
60

Estimer f

Je pourrais aussi créer des variables catégoriques pour l'éducation et inclure un effet fixe d'éducation:

$$\begin{aligned} Salaire &= f(age, education) + \epsilon \\ &= \beta_0 + \beta_1 Age + \beta_2 Secondaire + \beta_3 Cegep + \beta_4 Universite + \epsilon \end{aligned}$$

Estimer f

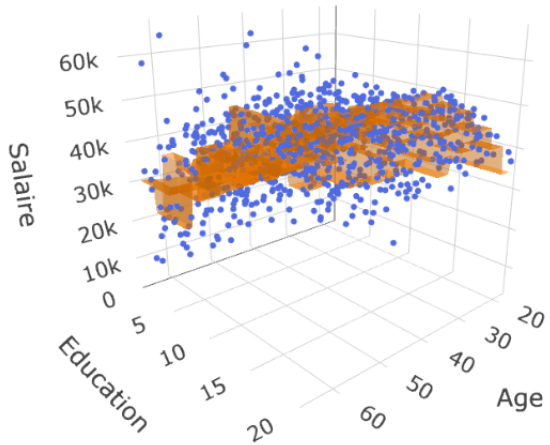

Salaire
60k
50k
40k
30k
20k
10k
Primaire
Secondaire
Cégep
Université
Education
20
30
40
50
60
Age

Estimer f

Ou bien je pourrais estimer une fonction parfaitement flexible (exemple: arbre de régression)

$$Salaire = f(age, education) + \epsilon$$

Estimer f


60k
50k
40k
30k
20k
10k
0
Salaire
5
10
15
20
Education
20
30
40
50
60
Age

Estimer f

Nous venons de voir plusieurs exemples de f:

$$
\begin{aligned}
salaire &= f(age, education) \\
&= \beta_0 + \beta_1 Age + \beta_2 Education + \epsilon \\
&= \beta_0 + \beta_1 Age + \beta_2 Education + \beta_3 Age^2 + \epsilon \\
&= \beta_0 + \beta_1 Age + \beta_2 Secondaire + \beta_3 Cegep + \beta_4 Universite + \epsilon \\
&= \text{25 000 si } education < 10; \quad \text{50 000 sinon}
\end{aligned}
$$

Quelle est la meilleure fonction f?

Erreur quadratique moyenne: $E(Y - \hat{Y})^2$

$$
\begin{aligned}
E(Y - \hat{Y})^2 &= E(f(X) + \epsilon - \hat{f}(X))^2 \\
&= \underbrace{(f(x) - \hat{f}(x))^2}_{\text{erreur réductible}} + \underbrace{var(\epsilon)}_{\text{erreur irréductible}}
\end{aligned}
$$

But des techniques de machine learning: réduire l'erreur réductible

Il y aura toujours des erreurs de prédiction dues à ϵ

Méthodes paramétriques

Il y a plusieurs façons d'estimer f: les méthodes paramétriques et non-paramétriques

Méthodes paramétriques: on fait une hypothèse sur la forme de f

Par exemple, on suppose la forme linéaire suivante:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p,$$

où on a $p + 1$ coefficients à estimer

Méthodes paramétriques

Une fois que la forme fonctionnelle est choisie, on doit trouver les valeurs des paramètres, par moindres carrés ordinaires (MCO) par exemple

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Méthodes non-paramétriques

Méthodes non-paramétriques: on ne fait pas d'hypothèse sur la forme de f

Plutôt, on cherche une fonction f qui se rapproche le plus de tous les points

Désavantage: ces méthodes ne réduisent pas le nombre de paramètres à estimer (contrairement aux méthodes paramétriques); elles nécessitent donc un grand nombre d'observations

Comment estimer f?

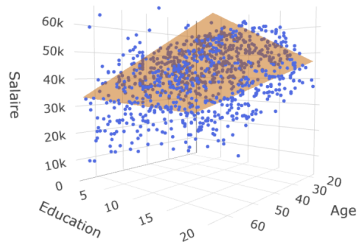


Figure 1: Paramétrique

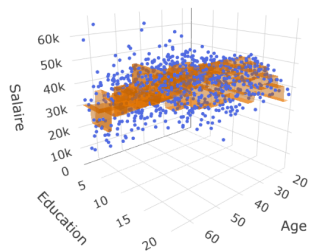


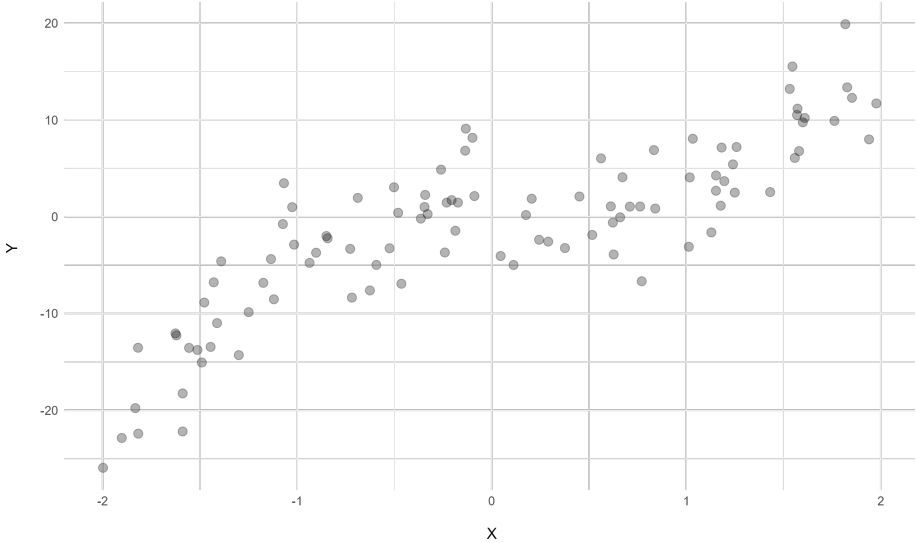
Figure 2: Non-paramétrique

Comment estimer f?

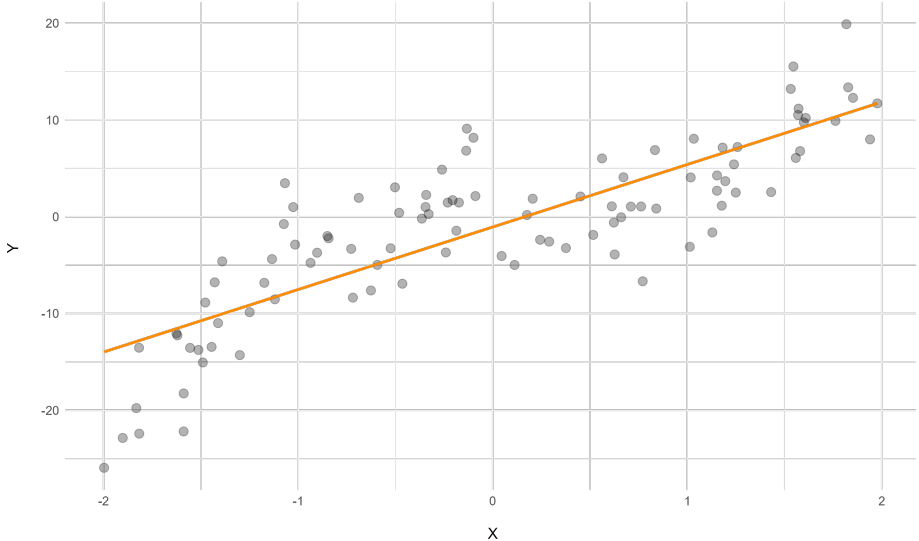
Les résultats d'une méthode paramétrique sont facilement interprétables (exemple: il est facile d'interpréter les coefficients d'une régression)

Les méthodes non-paramétriques, quoique plus flexibles, ne sont pas toujours préférables

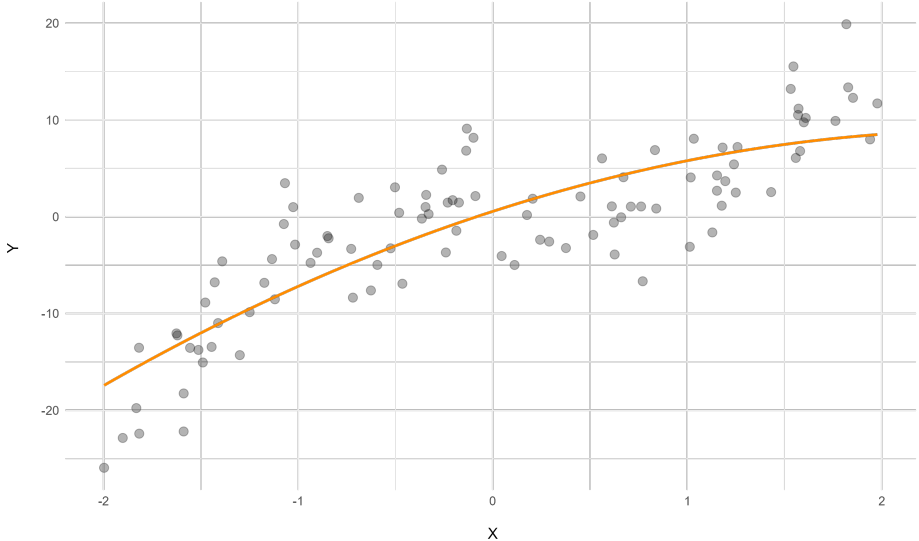
Comment estimer f?



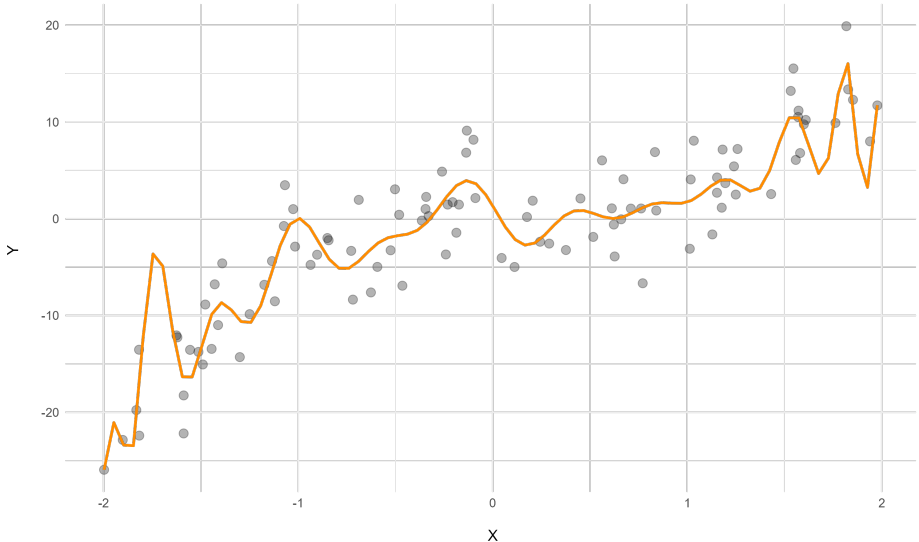
Comment estimer f?



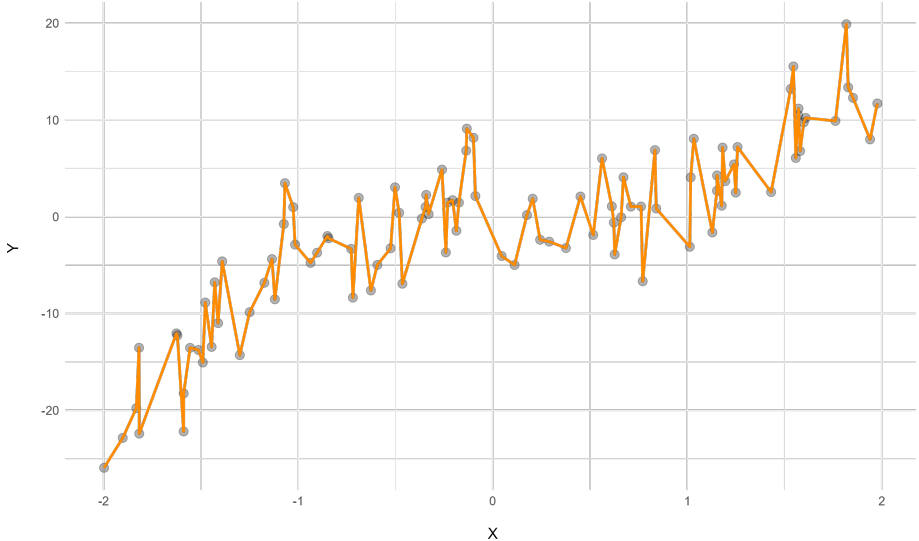
Comment estimer f?



Comment estimer f?

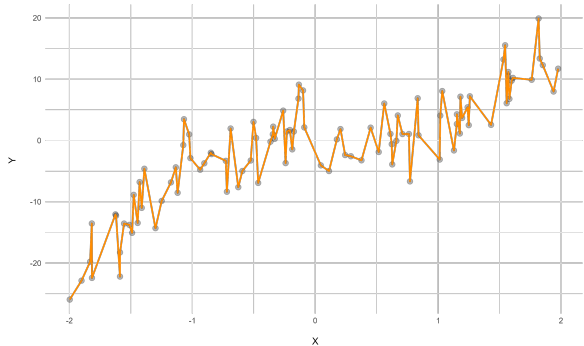


Comment estimer f?



Comment estimer f?

Les méthodes flexibles fonctionnent bien dans l'échantillon qui a servi à estimer f

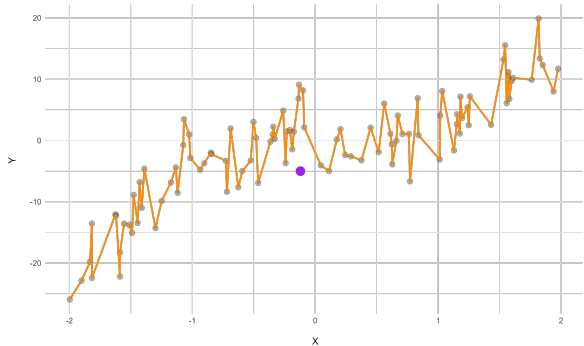


Comment estimer f?

Par contre, sur de nouvelles données...

$\Rightarrow$ problème de surapprentissage

$\Rightarrow$ *overfitting*

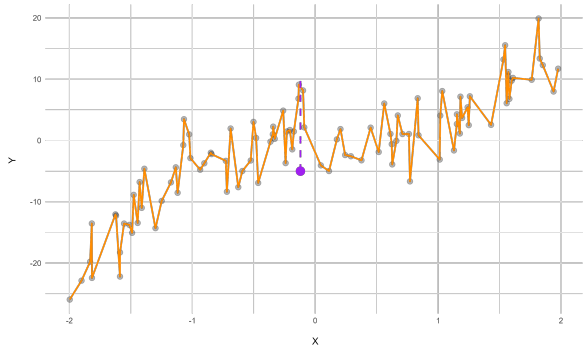


Comment estimer f?

Par contre, sur de nouvelles données...

$\Rightarrow$ problème de surapprentissage

$\Rightarrow$ grande erreur quadratique moyenne



Comment estimer f?

On veut s'assurer que les prédictions soient bonnes sur de nouvelles données

On sépare l'échantillon aléatoirement en deux parties:

- Échantillon d'entraînement: $\sim 70\%$
- Échantillon de test: $\sim 30\%$

Comment estimer f?

Observation	x	y	Échantillon	$\hat{f}(x)$	Erreur
1	23	34	Entrainement	$\hat{f}(23) = 30$	$34 - 30$
2	10	36	Entrainement	30	6
3	12	50	Test	45	5
4	8	10	Test	12	-2
5	22	20	Entrainement	22	-2

EQM dans l'échantillon d'entraînement: $\frac{1}{3}(4^2 + 6^2 + (-2)^2)$
EQM dans l'échantillon de test: $\frac{1}{2}(5^2 + (-2)^2)$

Compromis entre le biais et la variance

Biais: algorithme qui manque les relations pertinentes
(sous-apprentissage, avoir un modèle trop simple)

Variance: sensibilité aux variations dans l'échantillon d'entraînement
(surapprentissage, avoir un modèle trop complexe)

Compromis entre le biais et la variance

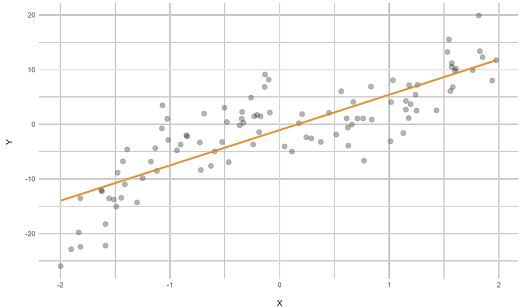


Figure 3: Biais élevé, variance faible

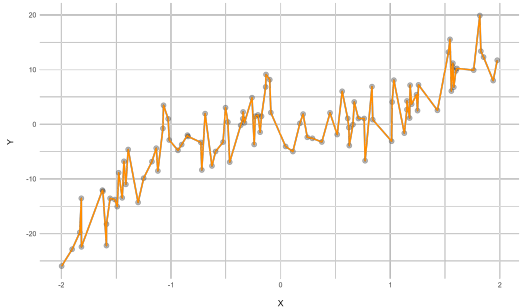


Figure 4: Biais faible, variance élevé

Compromis entre le biais et la variance

Pour une nouvelle donnée x_0, l'EQM peut être décomposée ainsi:

$$E(y_0 - \hat{f}(x_0))^2 = var(\hat{f}(x_0)) + (Biais(\hat{f}(x_0)))^2 + var(\epsilon)$$

On cherche à minimiser l'EQM

$\Rightarrow$ minimiser la variance

$\Rightarrow$ minimiser le biais

Il s'agit d'un compromis parce qu'il est facile de trouver une méthode qui minimise un des deux critères

Méthodes supervisées et non-supervisées

Pour la grande partie de la formation, nous verrons les méthodes supervisées

ML supervisé: nous avons des prédicteurs x_i et une réponse y_i

En l'*absence* de réponse, on tombe dans le non-supervisé

Exemple d'application: à partir des prédicteurs x_i, on cherche à former n groupes distincts

Classification et régression

Dernière chose!

Classification: la variable réponse y_i détermine l'appartenance à une classe (marié ou non; diagnostic; achat A, B ou C)

Régression: la variable réponse y_i prend des valeurs numériques (âge, revenu, grandeur)

Quiz!

Un chercheur veut prédire la récidive criminelle à partir de déterminants individuels (antécédents judicaires, âge, sexe, etc.).

1. Quelle est la variable réponse y_i?
2. Quels sont les prédicteurs x_i?
3. S'agit-il d'un problème supervisé ou non-supervisé?
4. S'agit-il d'un problème de classification ou de régression?

Méthodes non-supervisées

Méthodes non-supervisées

Puisqu'il n'y a pas de variable réponse, les techniques sont souvent utilisées pour faire une analyse exploratoire des données

- Aucun moyen d'entraîner un modèle
- Aucun moyen de vérifier sa validité

Mais elles restent très utiles!

Exemples d’application

- Une chercheuse en cancérologie pourrait analyser les niveaux d’expression génique chez 100 patients
- Un site de vêtements en ligne veut identifier des groupes homogènes d’acheteurs
- Netflix veut optimiser les choix de séries à mettre de l’avant

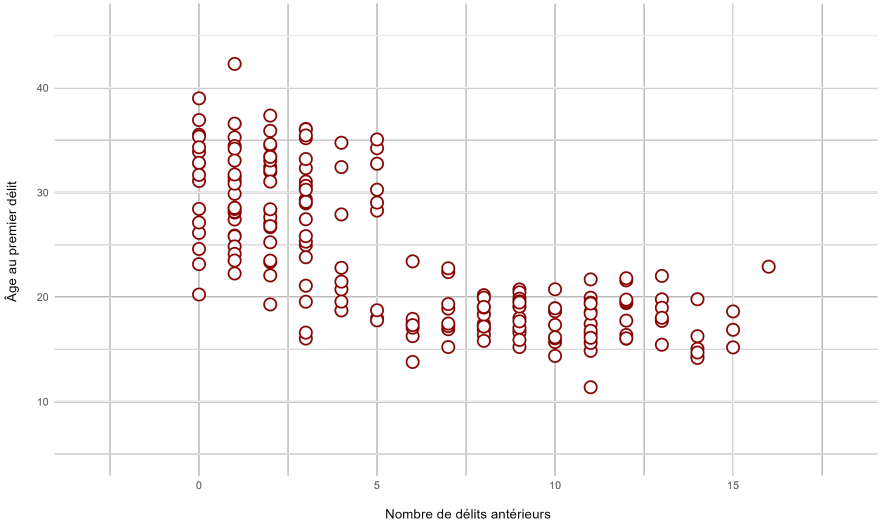
Clustering

Clustering (regroupement): approche simple, mais élégante, pour former des groupes homogènes

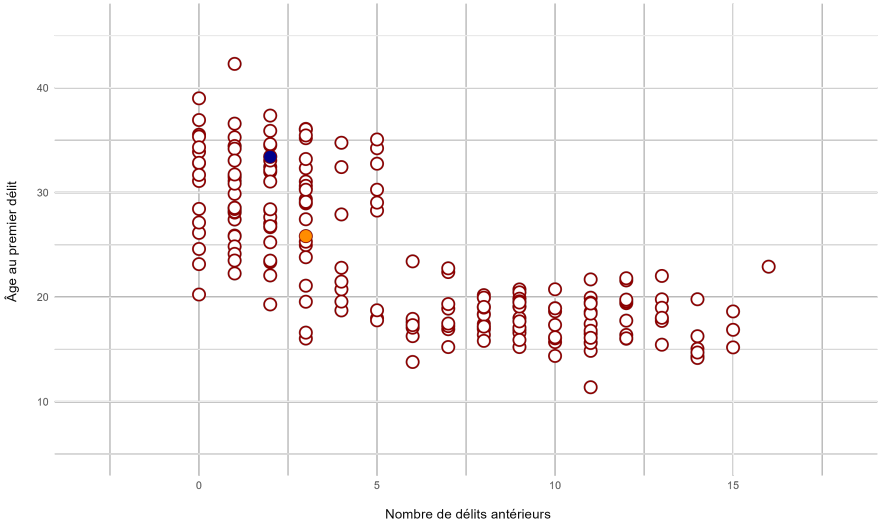
On veut assigner à chaque observation un cluster (groupe, partition) avec comme objectif que les observations au sein d'un cluster se ressemblent

$\Rightarrow$ algorithme du *k-means clustering*

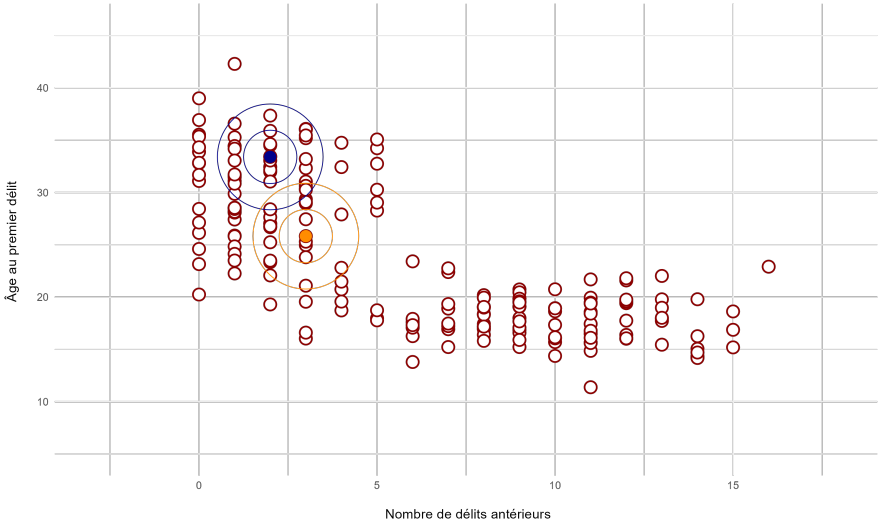
Algorithme


Âge au premier délit
40
30
20
10
0
5
10
15
Nombre de délits antérieurs

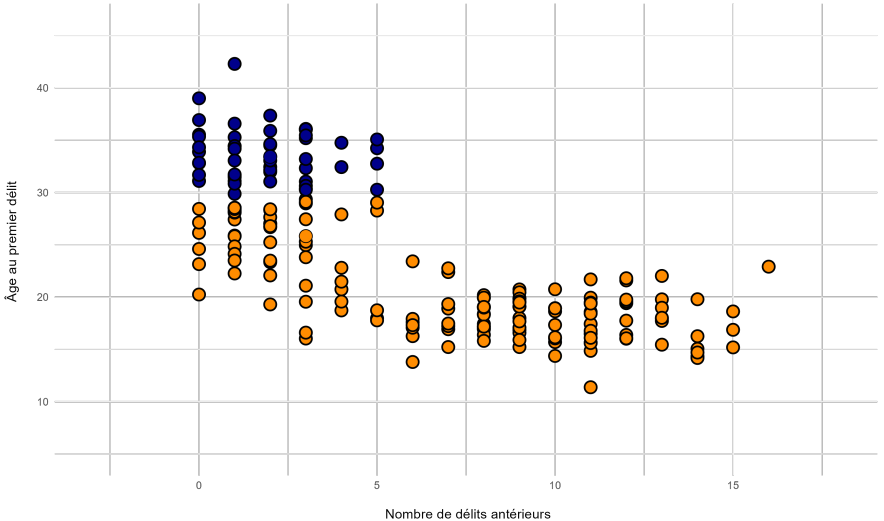
Algorithme


Âge au premier délit
40
30
20
10
0
5
10
15
Nombre de délits antérieurs

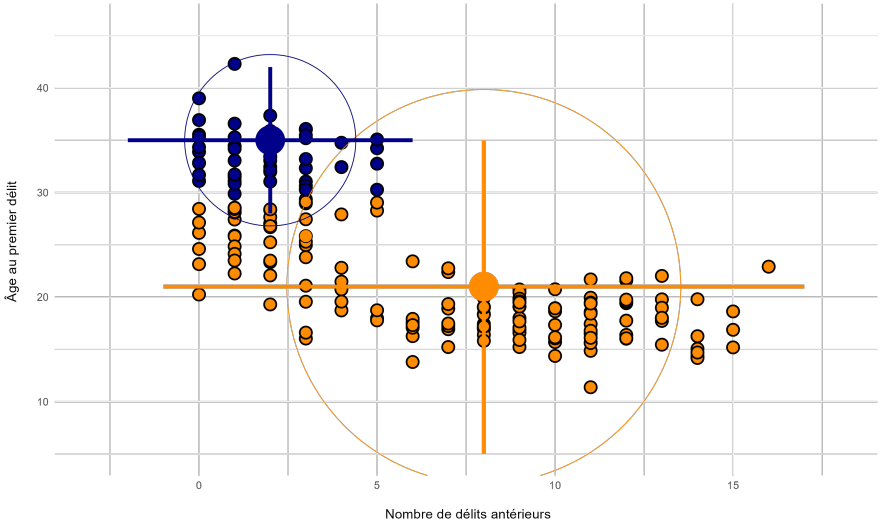
Algorithme


Âge au premier délit
40
30
20
10
0
5
10
15
Nombre de délits antérieurs

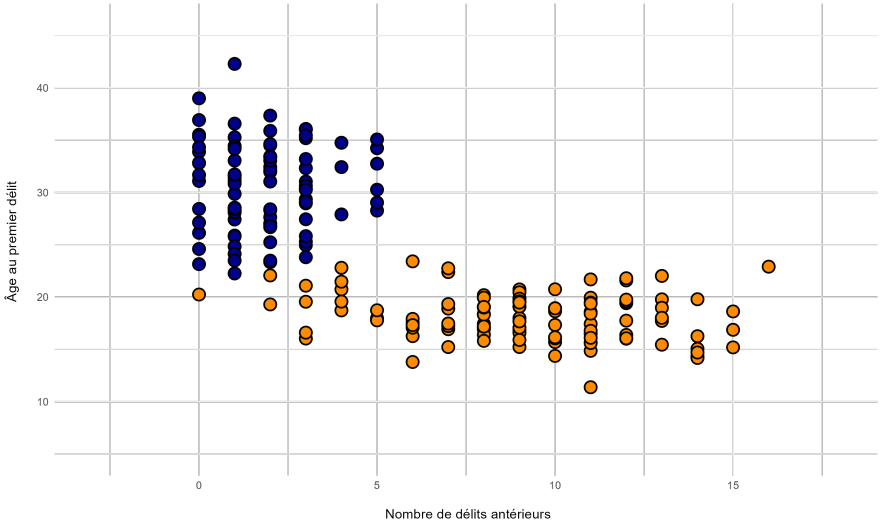
Algorithme


40
30
20
10
Âge au premier délit
0
5
10
15
Nombre de délits antérieurs

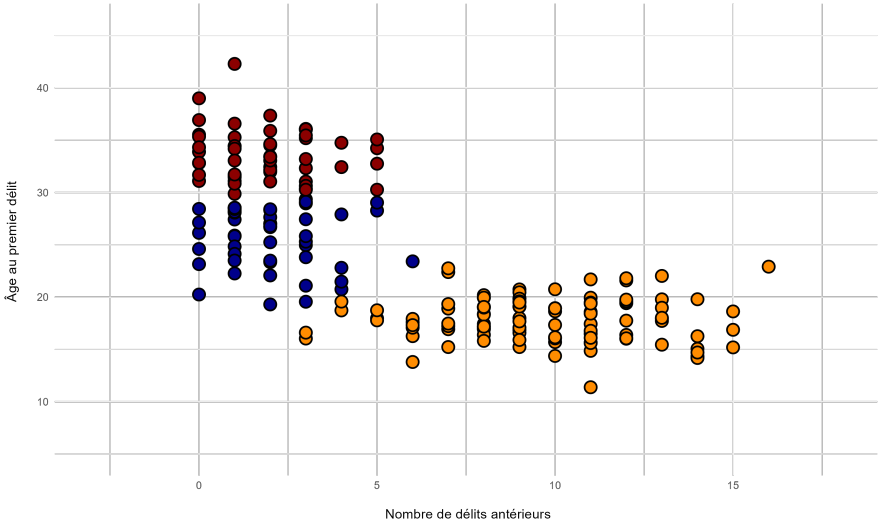
Algorithme


40
30
20
10
Âge au premier délit
0
5
10
15
Nombre de délits antérieurs

Algorithme


40
30
20
10
Âge au premier délit
0
5
10
15
Nombre de délits antérieurs

Algorithme


40
30
20
10
Âge au premier délit
0
5
10
15
Nombre de délits antérieurs

Algorithme

L’exemple que j’ai montré contenait deux prédicteurs...

Comment généraliser l’approche à trois prédicteurs? 4? p?

Algorithme

L’approche se généralise très bien avec plusieurs dimensions:

$$\min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

Essentiellement: on minimise la variante intra-cluster

Quiz

Rapport de police #00123

Date : 15 novembre 2024
Heure : 22h30
Lieu : Rue Saint-Denis, Montréal, QC
Type d'incident : Vol à main armée

Description des faits :
Un individu masqué, armé d'un couteau, a menacé le propriétaire d'un dépanneur pour obtenir la caisse. L'agresseur portait un manteau noir et des gants rouges. Aucun blessé n'a été signalé, mais le suspect a pris la fuite avec environ 500 $ en espèces.

Contexte :
Un chercheur obtient une centaine de rapports policiers. Il cherche à les regrouper selon la ressemblance des dossiers.

Question : Comment est-il possible de transformer le texte en données exploitables?

Question : Comment mettre en place un algorithme de clustering dans ce contexte?

Aller un peu plus loin

Clustering hiérarchique: le nombre de classes est décidé par l'algorithme

Analyse en composantes principales: condense plusieurs variables X en un nombre réduit (les composantes principales)

$$z_1 = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

Questions?

Méthodes supervisées paramétriques

Régression linéaire

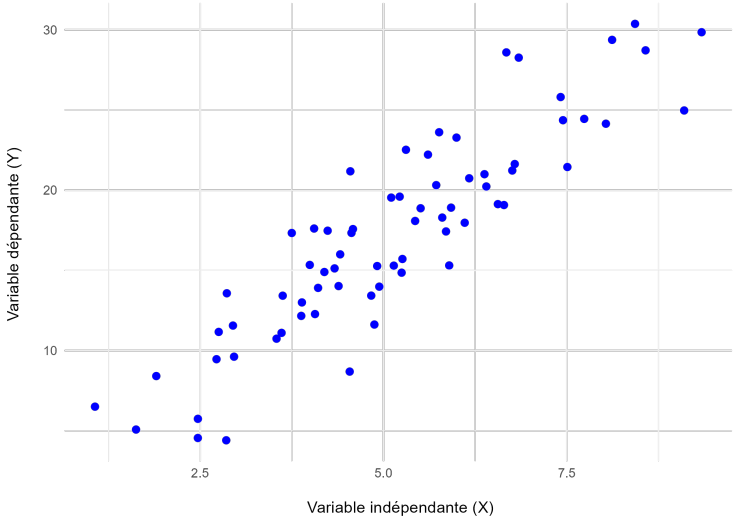
Régression linéaire simple: une seule variable explicative

$$Y = \beta_0 + \beta_1 X + \epsilon$$

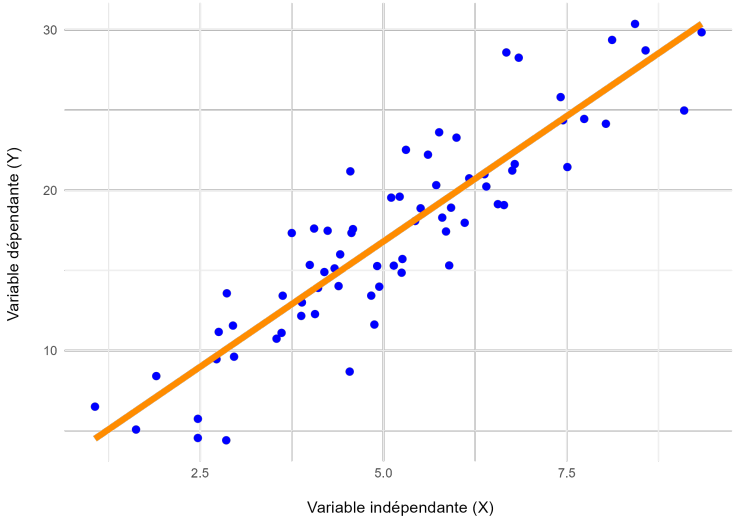
But: prédire Y le mieux possible

Moindres carrés ordinaires: méthode pour estimer β_0 et β_1

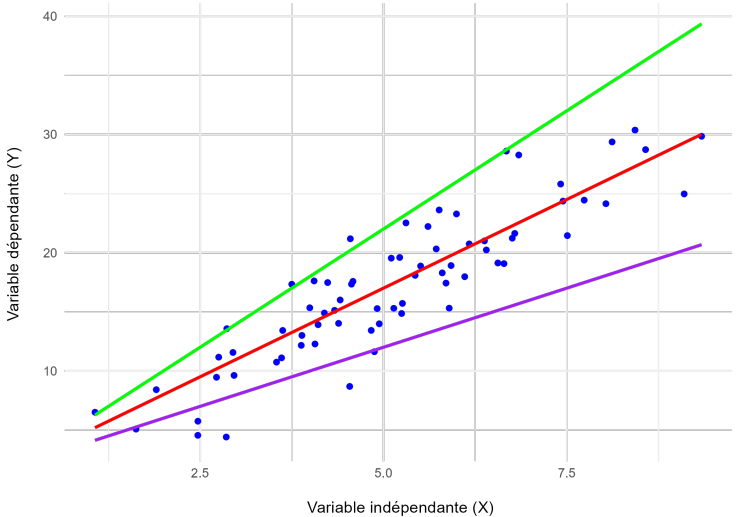
Régression linéaire



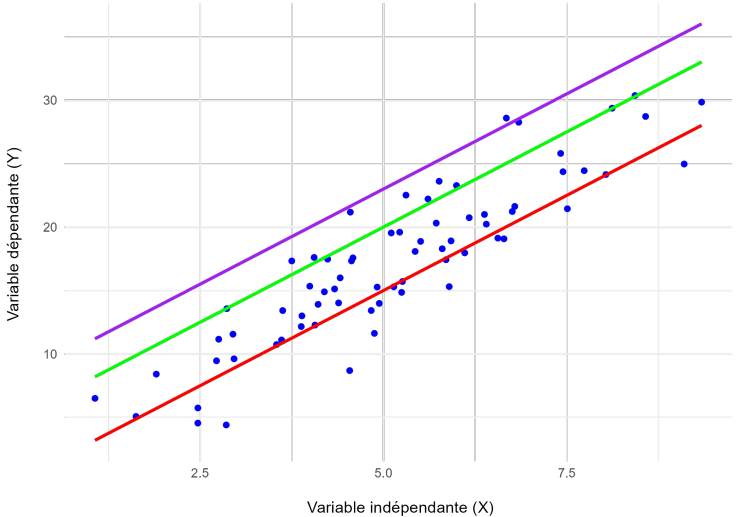
Régression linéaire



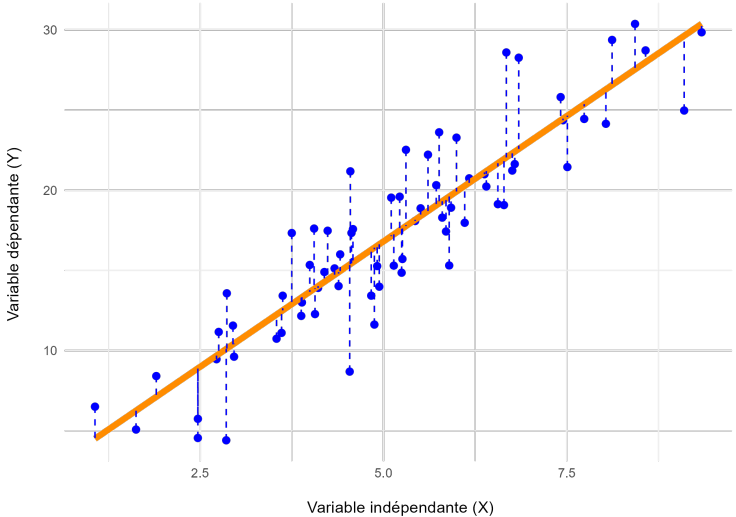
Régression linéaire



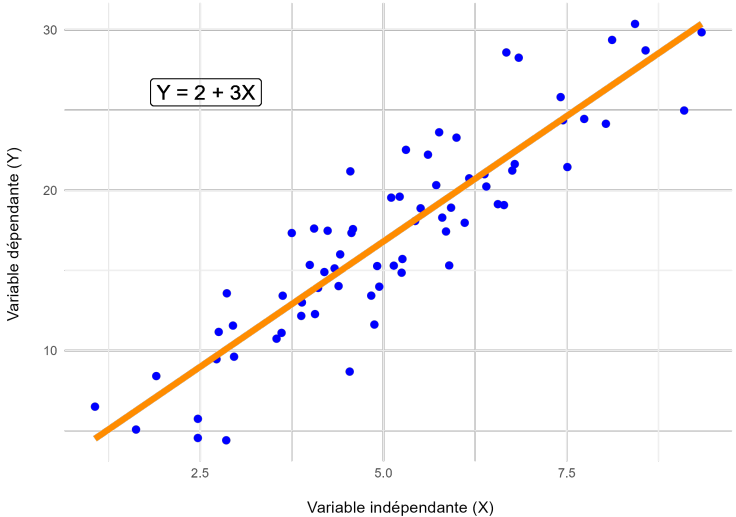
Régression linéaire



Régression linéaire



Régression linéaire



Régression linéaire

Comment les estimés ont-ils été trouvés? En minimisant la somme des erreurs au carré:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$\Rightarrow \epsilon = Y - \beta_0 - \beta_1 X$$

$$\hat{\beta}_0, \hat{\beta}_1 = \arg\min \sum_i \epsilon_i^2 = \arg\min \sum_i (Y_i - \beta_0 - \beta_1 X_i)^2$$

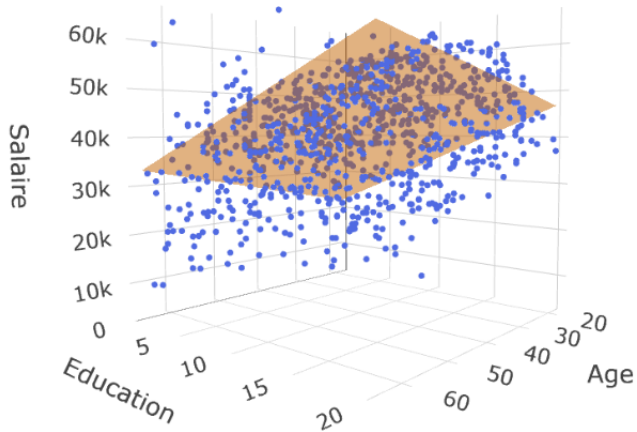
Régression linéaire

Régression linéaire multiple: ajout de variables explicatives

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

$$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 = \arg\min \sum_i \epsilon_i^2 = \arg\min \sum_i (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2})^2$$

Régression linéaire



Régression linéaire: multicolinéarité

Les coefficients d'une régression peuvent être instables s'ils sont fortement corrélés

Exemple: je veux prédire la récidive à partir de la durée de sentence et la sévérité du crime

$$Recidive = \beta_0 + \beta_1 Sentence + \beta_2 Severite + \epsilon$$

Quel est le problème? Comment interpréter les coefficients?

Autre exemple

On veut prédire la récidive à partir de 3 variables et le chercheur veut prendre en compte les interactions (l'effet d'une variable pourrait varier selon une autre variable)

Il veut estimer

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \epsilon$$

Il y a 7 paramètres à estimer. Avec $p = 20$, on a 211 paramètres; avec $p = 40$, on a 821 paramètres

Sans inclure les termes d'interaction d'ordre 3!

Problèmes

Il y a donc deux potentiels problèmes avec la régression linéaire:

Multicolinéarité forte: des variables sont très corrélées et les coefficients sont instables

Trop de prédicteurs: avec MCO, il faut nécessairement que $N > P$

Que faire?

Ridge et Lasso

Les régressions Ridge et Lasso permettent de réduire la dimensionnalité

$\Rightarrow$ se débarrasser de variables non-pertinentes ou redondantes

Comment?

Ridge et Lasso

Soit le modèle linéaire suivant:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i.$$

La méthode MCO consiste à minimiser la somme des erreurs au carré (RSS, pour *residual sum of squares)*:

$$RSS = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})^2$$

Ridge et Lasso

$$\hat{\beta}_{MCO} = \arg\min RSS$$

Dans la régression Ridge et Lasso, on ajoute une pénalité aux paramètres:

$$\hat{\beta}_{Ridge} = \arg\min RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

$$\hat{\beta}_{Lasso} = \arg\min RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

λ est un hyperparamètre (un paramètre choisi par l'utilisateur)

Ridge et Lasso

Comme avec le MCO, on cherche à s'approcher des données le plus possible (en cherchant à diminuer RSS) mais...

- La pénalité Ridge ($\lambda \sum_{j=1}^{p} \beta_j^2$) force certains coefficients à se rapprocher de zéro
- La pénalité Lasso ($\lambda \sum_{j=1}^{p} |\beta_j|$) force certains coefficients directement à zéro

Ridge et Lasso

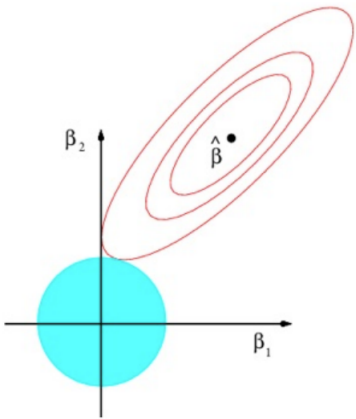


Figure 5: Pénalité Ridge

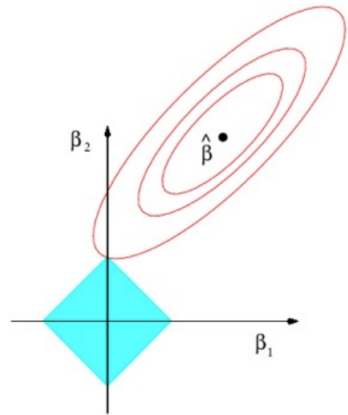


Figure 6: Pénalité Lasso

Ridge et Lasso

Si $\lambda = 0$, alors les estimateurs reviennent à $\hat{\beta}_{MCO}$

Si $\lambda \uparrow$, alors la pénalité augmente

Comment trouver λ?

Ridge et Lasso

On peut essayer plein de valeurs différentes pour λ et calculer l'erreur quadratique moyenne (*grid search*) pour chacune, puis on sélectionne le λ qui minimise l'erreur

$$\lambda = 0, 1, 2, 3, 4, \ldots, 100$$

Risque: surapprentissage (aucune garantie que les prédictions soient bonnes dans l'échantillon de test)

Idée: avoir une idée de la performance du modèle avant de passer à l'échantillon de test

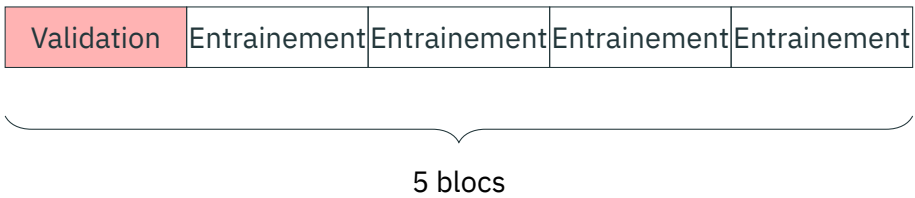
Ridge et Lasso

Validation-croisée à k blocs: produire un échantillon de validation

Itération 1: le premier bloc est utilisé pour la validation

...

Itération k: le dernier bloc est utilité pour la validation



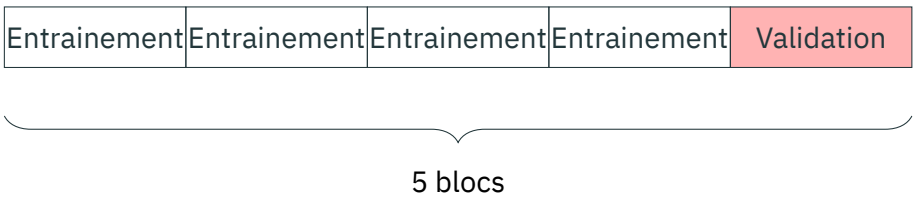
Ridge et Lasso

Validation-croisée à k blocs: produire un échantillon de validation

Itération 1: le premier bloc est utilisé pour la validation

...

Itération k: le dernier bloc est utilité pour la validation



Ridge et Lasso

Table 2. Purpose of datasets, ML models and their evaluation.

Dataset	Type of Recurrence	Purpose	ML Model	Evaluation Metrics	Evaluation Value
Thailand	Other	Recidivism in drug distribution	Logistic Regression	ACC	0.90
MnSTARR+	General	General recidivism	LogitBoost	ACC AUC	0.82 0.78
LS/CMI	General	General recidivism	Random Forest	ACC AUC	0.74 0.75
NCRA+	General	General recidivism	Glmnet	AUC	0.70
RisCanvi	Violent	Violent Recidivism	MLP	AUC	0.78
FDJJ	Sexual	Sexual recidivism in Youth	Random Forest	AUC	0.71
RITA+	Other	General and violent recidivism in male	Random Forest	AUC	0.78
HCR-20+	General	General recidivism	Ensemble model with NBC, kNN, MLP, PNN, SVM	ACC	0.87
YLS/CMI	Other	General recidivism in Youth	Random Forest	ACC AUC	0.65 0.69
SAVRY+	Other	Violent recidivism in youth	Logistic Regression	AUC	0.71
StatRec	General	General Recidivism	Logistic Regression	ACC AUC	0.73 0.78
	Sexual	Sexual recidivism	LDA	ACC AUC	0.96 0.73
	Violent	Violent recidivism	Logistic regression	ACC AUC	0.78 0.74
DOI	General	General recidivism	L1–Logistic Regression	ACC AUC	0.78 0.73
	Sexual	Sexual recidivism	L1–Logistic Regression	ACC AUC	0.96 0.77
	Violent	Violent recidivism	Penalized LDA	ACC AUC	0.78 0.74

ACC: accuracy; AUC: area under the curve; Thailand: data by central correctional institution for drug addicts and central women correctional institution in Thailand; MnSTARR+: Minnesota Screening Tool Assessing Recidivism Risk + Minnesota Sex Offender Screening Tool-3; LS/CMI: Level of Service/Case Management Inventory; NCRA+: NeuroCognitive Risk.

Ridge et Lasso: Exemple

Rapport de police #00123

Date : 15 novembre 2024
Heure : 22h30
Lieu : Rue Saint-Denis, Montréal, QC
Type d'incident : Vol à main armée
Individu : Homme blanc

Description des faits :
Un individu **masqué**, **armé** d'un couteau, a menacé le propriétaire d'un dépanneur pour obtenir la caisse. L'agresseur portait un manteau **noir** et des gants **rouges**. Aucun blessé n'a été signalé, mais le suspect a pris la fuite avec environ 500 $ en espèces.

J'obtiens une centaine de rapports policiers comme celui de gauche

Je veux voir s'il y a une association entre les **adjectifs** utilisés et l'ethnicité du contrevenant

Ridge et Lasso: exemple

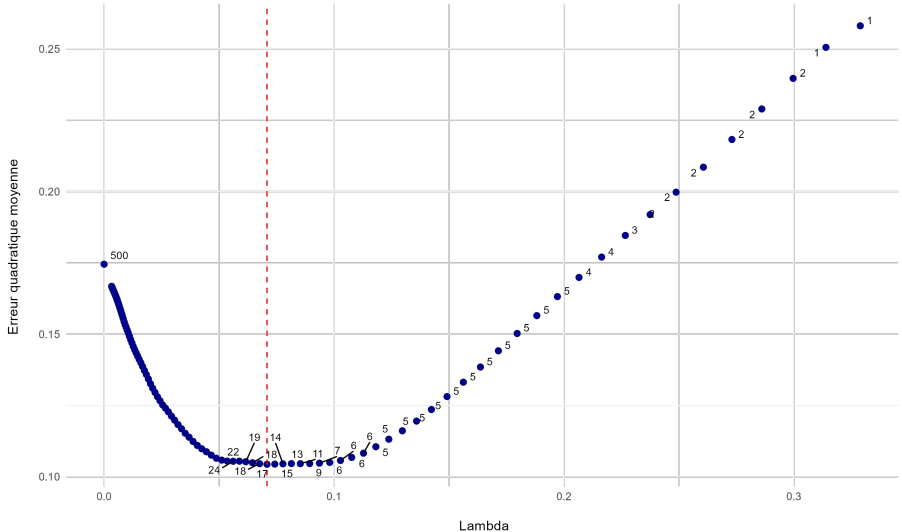
Les données ressemblent à ceci:

500

100

Rapport	Ethnicité	Masqué	Armé	Noir	Rouge	Agressif	Intoxiqué	...
123	Blanc	1	1	1	1	0	0	
124	Blanc	1	0	0	0	0	0	
125	Noir	0	1	1	0	1	1	
...								

Ridge et Lasso: exemple



Aller un peu plus loin

Régression elastic-net: un mélange entre Ridge et Lasso

$$\hat{\beta}_{EN} = \arg\min RSS + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

La régression Lasso peut prendre différentes formes: linéaire, logit, probit, Poisson

Questions?

Méthodes supervisées non-paramétriques

Arbres de régression

Nous allons maintenant couvrir quelques méthodes non-paramétriques: les abres de décision

Deux types d'arbres:

- Arbres de **classification**: prédire un outcome binaire (*cette image comporte-t-elle un chat?*)
- Arbres de **régression**: prédire un outcome continu (*combien y a-t-il de chats sur cette image?*)

Nous allons nous concentrer sur les arbres de régression

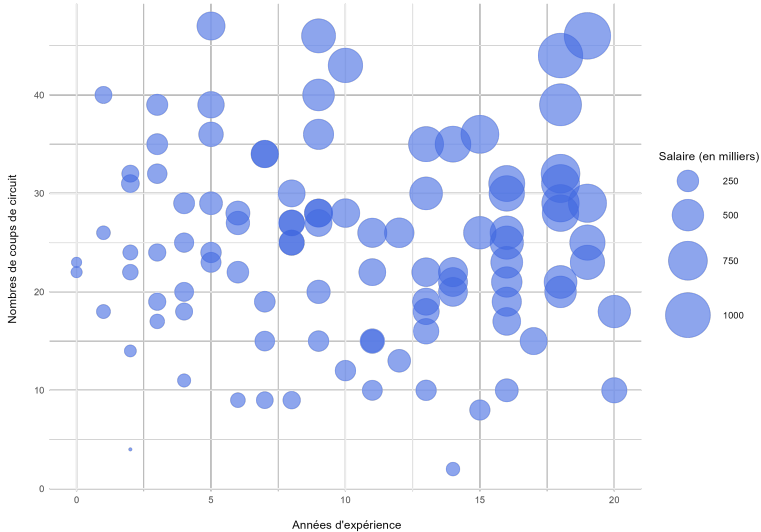
Arbres de régression

Commençons avec un exemple!

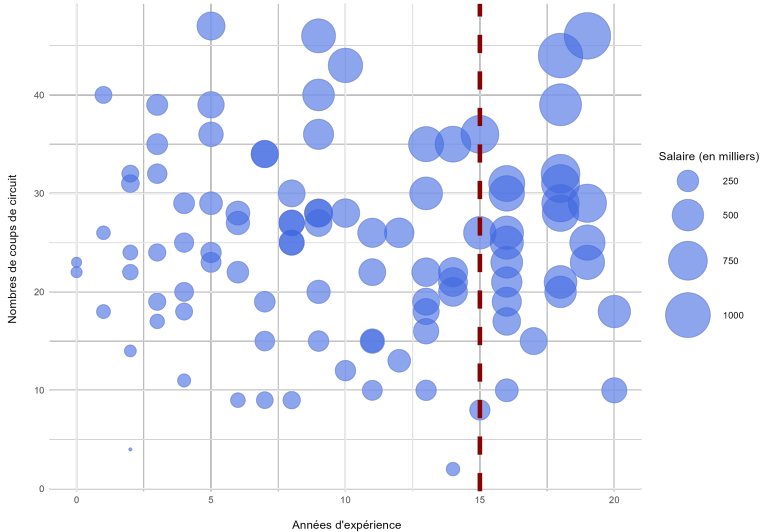
Nous avons des données sur le **salaire** de joueurs de baseball, ainsi que des informations sur leur **nombre d'années d'expérience** et le **nombre de coups de circuit** (lors de la dernière saison)

On cherche à prédire le salaire

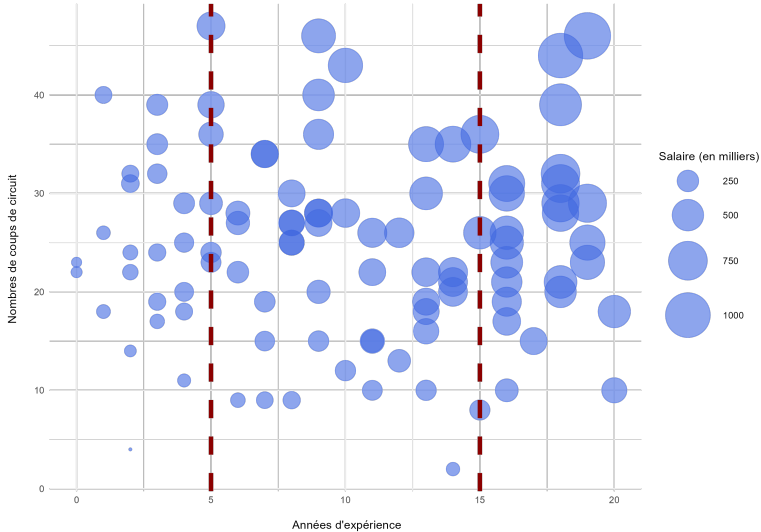
Arbres de régression



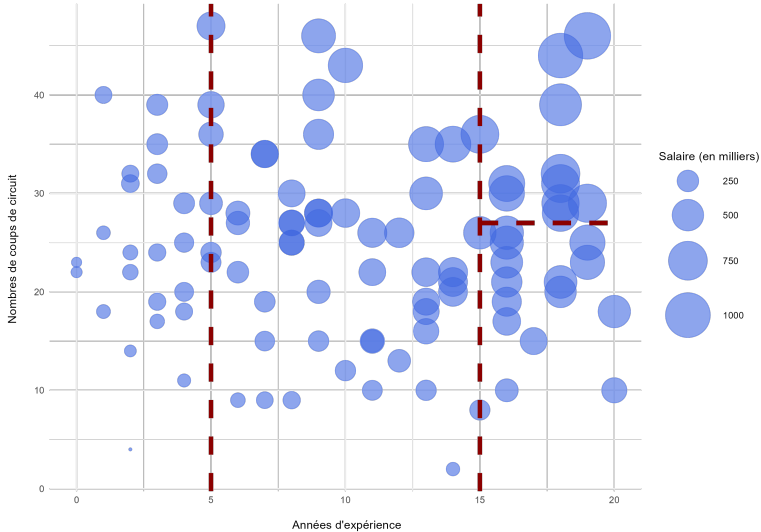
Arbres de régression



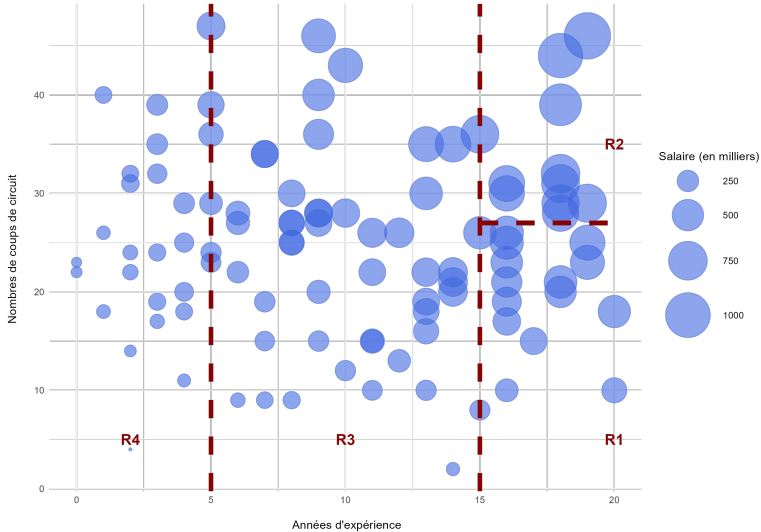
Arbres de régression



Arbres de régression



Arbres de régression



Arbres de régression

Notre algorithme *visuel* a déterminé 4 régions ($R1$ à $R4$)

On peut maintenant faire des prédictions en prenant la moyenne de salaire dans chaque région

Dans $R1$, la moyenne des salaires est de 161 000$

$\Rightarrow$ ce sera ma prédiction pour un nouvel individu qui tombe dans cette région

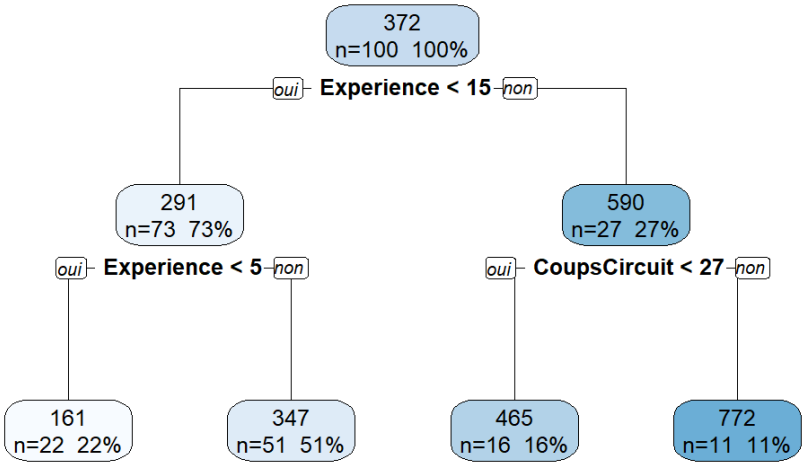
Arbres de régression

Chaque région est définie par un ensemble de caractéristiques, par exemple:

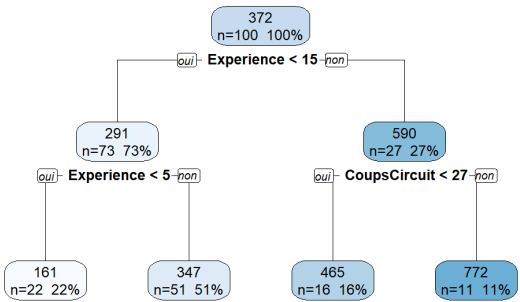
$$R1 : \{Experience \geq 15 \quad \& \quad CoupsCircuits < 27\}$$

C'est tout! J'ai un arbre de régression!

Arbres de régression



Arbres de régression



Terminologie

Noeuds: prédictions et nombre d'observations

Feuilles (noeuds finaux): prédictions finales

Branches: critères

Arbres de régression

Question importante: comment trouver les régions? Comment trouver les critères?

Comme avant, on cherche à minimiser le RSS:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

- J noeuds
- $\hat{y}_{R_j}$: prédiction

Arbres de régression

Idée: essayer tous les arbres possibles

Ce serait un algorithme très long et coûteux

Plutôt, on va adopter l'approche *top-down*: fractionnement binaire récursif

Arbres de régression

1. On commence avec toutes les observations dans le premier noeud (la *racine*)
2. Parmi tous les prédicteurs, on cherche le prédicteur X_j et le seuil s tel que les deux régions descendantes

$$R_{gauche} = \{X | X_j < s\}$$
$$R_{droite} = \{X | X_j \geq s\}$$

qui mènent aux prédictions qui réduisent le plus le RSS

Arbres de régression

L'algorithme essaie tous les prédicteurs et tous les seuils

Pour tout j et pour tout s, on définit la paire:

$$R_1(j,s) = \{X | X_j < s\}$$
$$R_2(j,s) = \{X | X_j \geq s\}$$

Et on choisit les j et s:

$$j, s = \arg\min \sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2$$

Arbres de régression

3. On répète le processus sur les sous-régions déjà trouvées
4. On continue jusqu'à ce qu'un critère d'arrêt soit atteint (par exemple: profondeur limite, nombre d'observations par feuille, etc.)

Comme toute méthode de machine learning, on fait généralement ces étapes sur l'échantillon d'entraînement

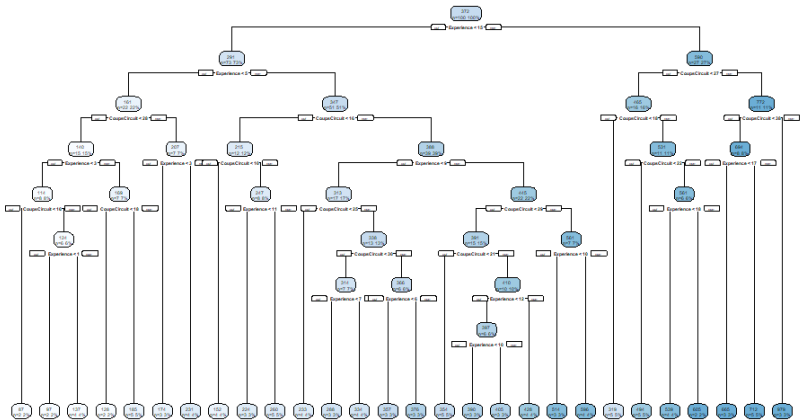
Arbres de régression

L'algorithme tel que défini pose un risque de surapprentissage: l'arbre peut devenir très profond

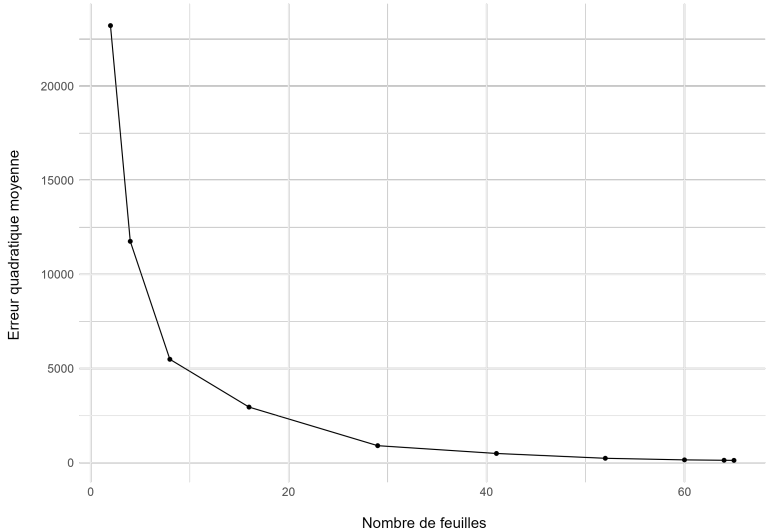
Étape importante: émondage (*pruning*)

Cette étape consiste à couper des branches à l'arbre

Arbres de régression



Arbres de régression



Arbres de régression

Dans l'échantillon d'entraînement, plus l'arbre est profond, plus l'erreur quadratique moyenne est faible

Cas extrême: une observation par feuille (EQM = 0)

Comme avec le Lasso, nous allons ajouter une pénalité au calcul

Arbres de régression

L'idée est de construire un arbre profond T_0

On considère un sous arbre $T \subset T_0$ (par exemple: retirer la dernière couche)

Notation: cet arbre contient $|T|$ feuilles

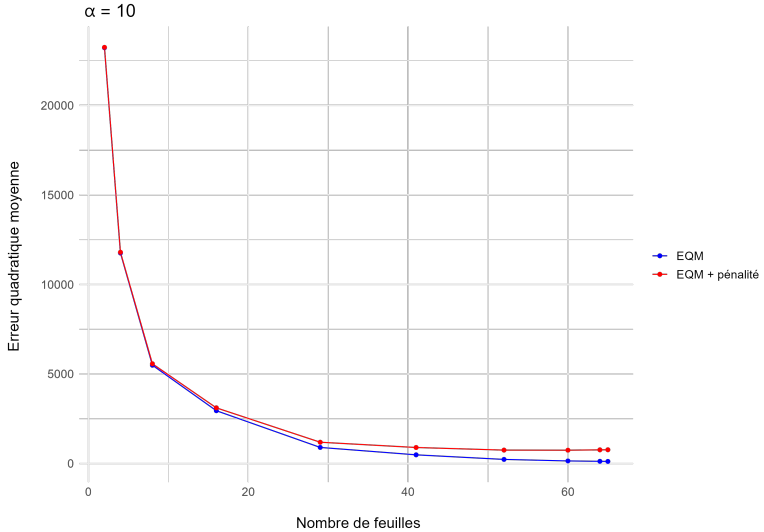
Par définition, on a que $|T| < |T_0|$

Arbres de régression

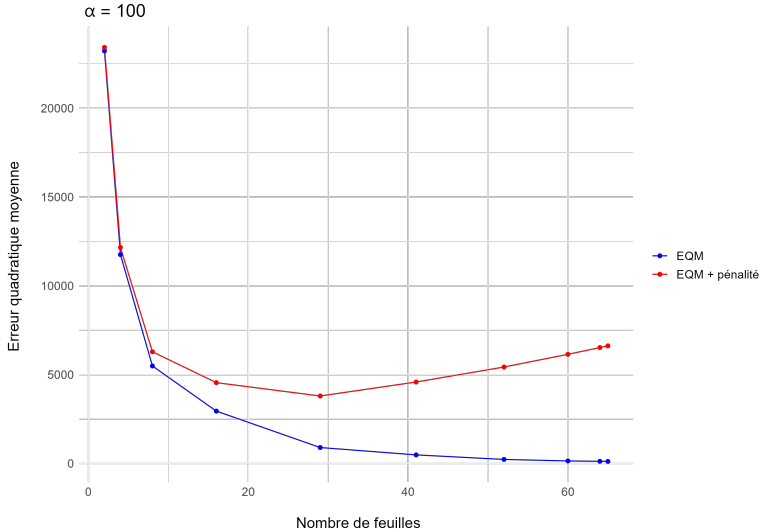
$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

$$RSS_\alpha = \sum_{m=1}^{|T|} \sum_{i: x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

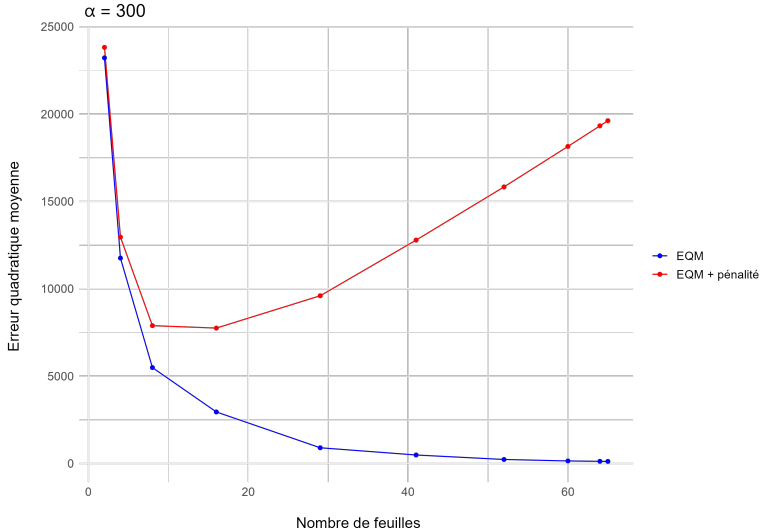
Arbres de régression



Arbres de régression



Arbres de régression



Arbres de régression

La pénalité force l'erreur quadratique moyenne à augmenter avec la profondeur de l'arbre *même* dans l'échantillon d'entrainement

Comment trouver α? Validation croisée!

On sélectionne le α qui minimise le plus l'erreur

Arbres de régression versus régression linéaire?

Est-ce qu'un arbre de régression est *meilleur* qu'une régression linéaire? Pas nécessairement...

Régression linéaire:

$$f(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

Arbre de régression:

$$f(X) = \sum_{m=1}^{M} c_m 1\{X \in R_m\}$$

Arbres de régression versus régression linéaire?

Quelle forme est la meilleure? Ça dépend du problème.

Si la relation entre les prédicteurs et l'outcome est linéaire, la régression peut être tout à fait appropriée

Avantages des arbres

Les arbres présentent plusieurs avantages:

- Faciles à expliquer (j'espère!)
- Reproduisent une façon de réfléchir commune chez les humains
- Se présentent bien graphiquement
- Peuvent gérer les prédicteurs numériques et non-numériques
- Prennent en compte les interactions mécaniquement

Désavantages des arbres

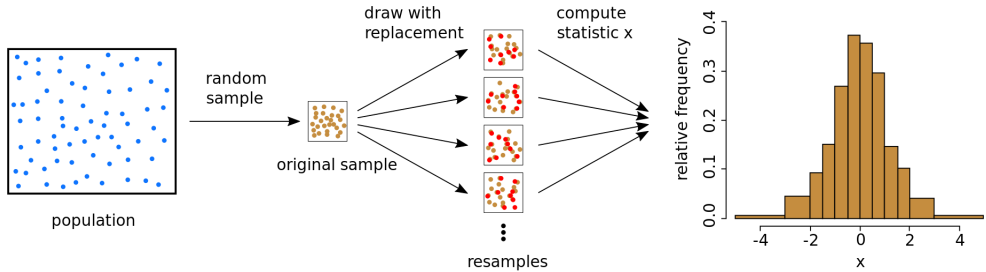
Mais ils ont aussi quelques désavantages:

- Ne produisent pas les meilleures prédictions (nous verrons comment les améliorer)
- Ils sont généralement peu robustes (changer un peu les données change de beaucoup l'arbre résultant)

Améliorer les prédictions

Première méthode: *bagging* (bootstrap aggregating)

D'abord, qu'est-ce que le bootstrap? $\Rightarrow$ rééchantillonnage avec remise



Améliorer les prédictions

Le problème de l'arbre est son manque de robustesse (grande variance)

Idéalement, on ferait la procédure sur plein d'échantillons d'entrainement, mais on n'en a qu'un seul $\rightarrow$ bootstrap

$$\hat{f}_{bagging}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x)$$

Améliorer les prédictions

Désavantage du bagging: perte d'interprétation

Difficile de visualier 100 arbres (ou 1000) $\rightarrow$ impossible de savoir les variables qui ont mené aux prédictions

Importance d'une variable: nombre de fois qu'une variable est choisie comme critère (ou à quel point le RSS diminue quand on choisit cette variable)

Améliorer les prédictions

Autre désavantage du bagging: les arbres risquent de se ressembler

Exemple: l'algorithme choisit peut-être la variable X_1 pour le premier noeud dans tous les arbres

La forêt aléatoire est un ensemble d'arbres *décorrélés*

Améliorer les prédictions

En forêt aléatoire, on ne permet pas à tous les prédicteurs d'être considérés à chaque étape

À *chaque* itération de *chaque* arbre, on permet à m prédicteurs d'être considérés

Souvent: $m \approx \sqrt{p}$

Si $m = p \rightarrow$ bagging

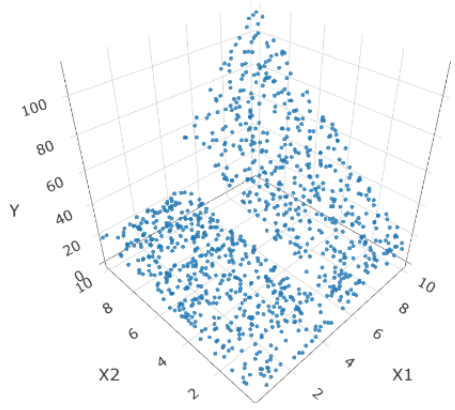
Exemple

J'ai simulé le processus suivant:

$$X_1 \sim N(0, 10)$$
$$X_2 \sim N(0, 10)$$

$$Y = \begin{cases} 1.5 \times X_1 + 1.5 \times X_1 X_2 & \text{si } X_1 > 5 \\ 0.5 \times X_1 + 2 \times X_2 & \text{si } X_1 \leq 5 \end{cases}$$

Exemple



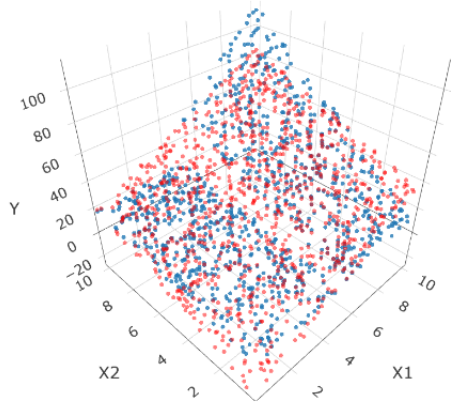
Exemple

Approche 1: régression linéaire

Je ne connais pas le processus générateur de données... j'estime donc:

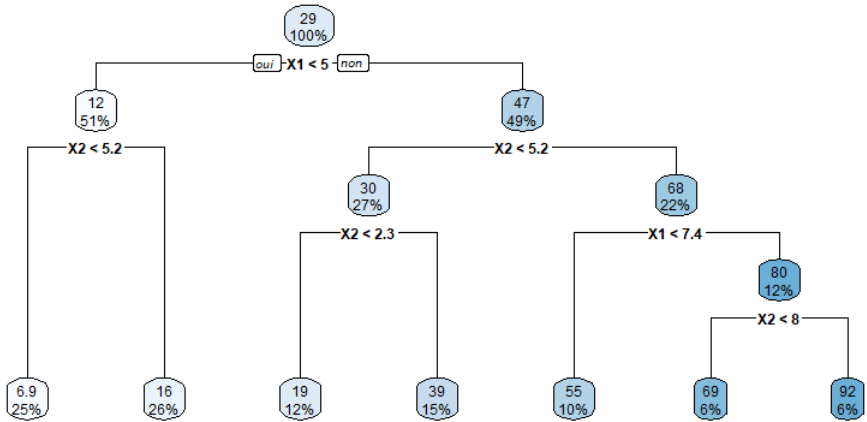
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Exemple

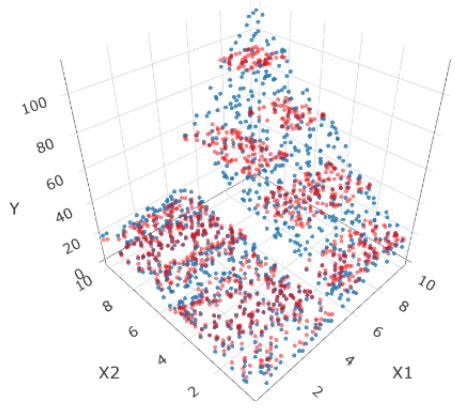


Exemple

Approche 2: arbre de régression



Exemple

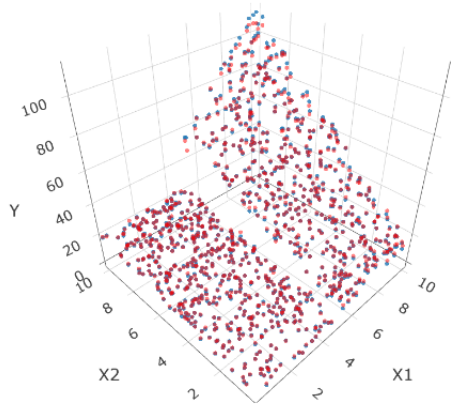


Exemple

Approche 3: forêt aléatoire

Je reprends la procédure de l'arbre en faisant du bagging (500 arbres) et en *décorrélant* les arbres

Exemple



Aller un peu plus loin

Forêts aléatoires causales: identifier des effets de traitement hétérogènes

Support Vector Machines: hyperplan optimal qui sépare les observations

Boosting: arbres et forêts qui corrigent les erreurs des itérations passées

Questions?

Conclusion

Conclusion

Nous avons vu plusieurs méthodes tirées de l'apprentissage-machine

Méthodes non-supervisées:

1. Clustering

Méthodes paramétriques:

1. Régression Ridge
2. Régression Lasso

Méthodes non-paramétriques:

1. Arbres de régression
2. Forêt aléatoire

Conclusion

Les méthodes *fancy* n'ont pas toujours les réponses à tout: un modèle bien simple peut nous mener loin (tout en assurant une meilleure interprétabilité)

Conseil: commencez simple! On se sert de méthodes pour résoudre des problèmes, et non pour impressionner

Merci!

william.arbour@umontreal.ca