Introduction au Machine Learning

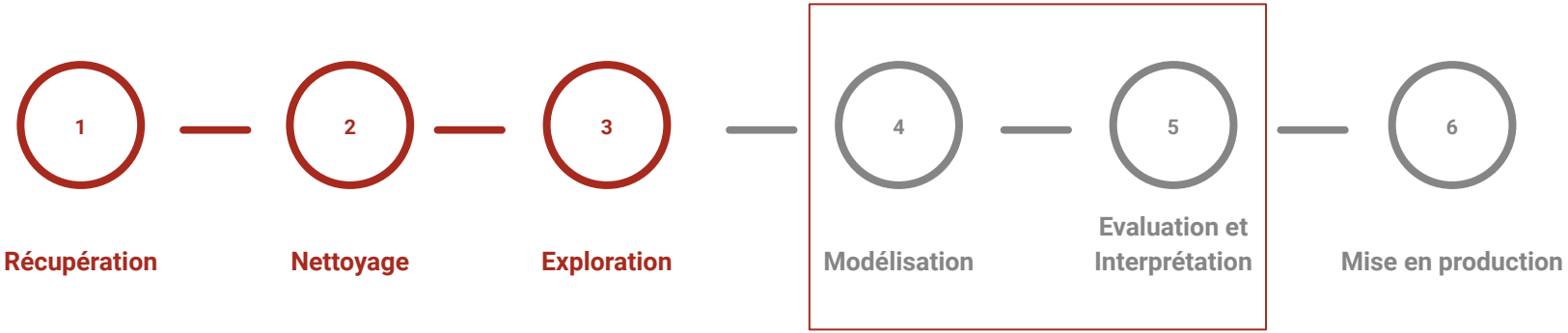
Partie 1 - La régression linéaire

Octobre 2020

Les étapes du Machine Learning

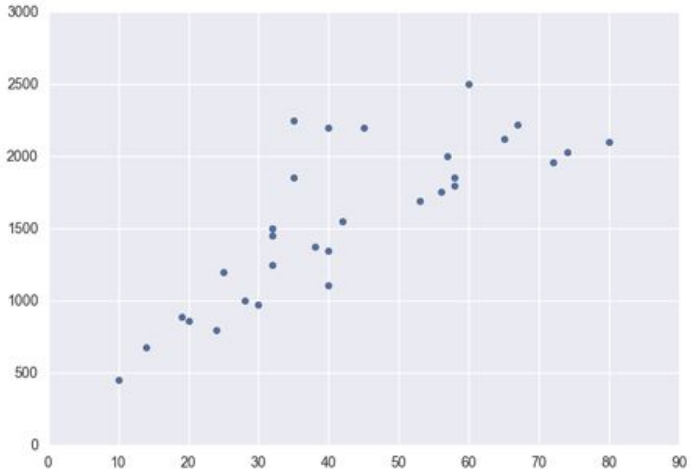
SIMPLON.CO

Cycle de travail du développeur IA

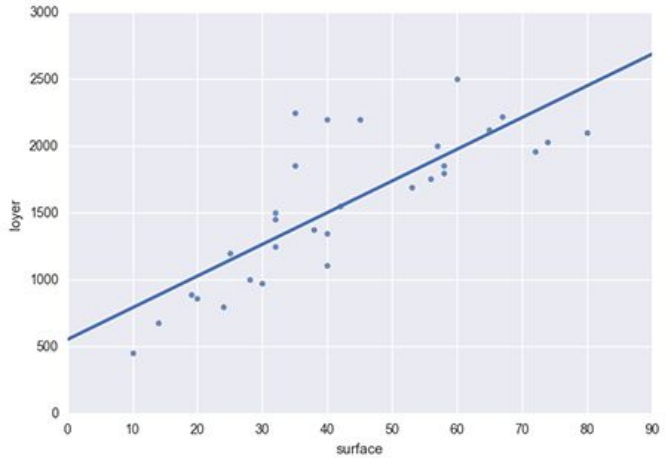


Le principe de la modélisation

En machine learning, l'objectif est de trouver un modèle du phénomène à l'origine des données. C'est-à-dire qu'on considère que chaque donnée observée est l'expression d'une variable aléatoire générée par une distribution de probabilité.



On a représenté ici le loyer d'un appartement en fonction de sa superficie



Une première modélisation simple du phénomène (le prix du loyer) serait donc de considérer la droite la plus "proche" de l'ensemble des points.

Les étapes d'un projet de machine learning (détaillé)

1. Importation des données
2. Nettoyage des données
3. Exploration des données (statistiques univariées et bivariées)
4. Choix du modèle de machine learning
5. Preprocessing (Préparation du jeu de données en vue du machine learning)
    a. échantillonnage (si besoin) en vue de l'apprentissage
    b. sélection de la variable cible et des variables à explicatives
    c. Encoder (créer des dummies)
    d. standardisation et normalisation (si besoin) en fonction du modèle ML choisi
6. Division du jeu de données en training/validation/testing sets
7. Choix de l'hyper-paramètres (grid search cross validation)
8. Apprentissage sur le training set
9. Evaluation du modèle sur le testing set en fonction d'une métrique
10. Mise en production

Les étapes grisées ne sont pas nécessaires pour les modèles linéaires

Commencer un projet de machine learning

1. Importation des données
2. échantillon (pour aller vite)
3. Drop NA (pour pas faire bugger les algo
4. Identifier Y la variable cible et la mettre au bon format
5. Lancer un premier algorithme simple (régression multiple)
6. Mettre le résultat au bon format
7. Soumettre et obtenir un premier score qu'on améliorera progressivement
8. Améliorer le score de manière itérative.

Zoom sur l’évaluation d’une modélisation

L’entraînement d’un modèle revient à mesurer l’erreur de la sortie de l’algorithme avec les données d’exemple et chercher à la minimiser.

Un premier piège à éviter est donc d'évaluer la qualité de votre modèle final à l'aide des mêmes données qui ont servi pour l'entraînement. En effet, le modèle est complètement optimisé pour les données à l'aide desquelles il a été créé. L'erreur sera précisément minimum sur ces données. Alors que l'erreur sera toujours plus élevée sur des données que le modèle n'aura jamais vues !

Pour minimiser ce problème, la meilleure approche est de séparer dès le départ notre jeu de données en deux parties distinctes :

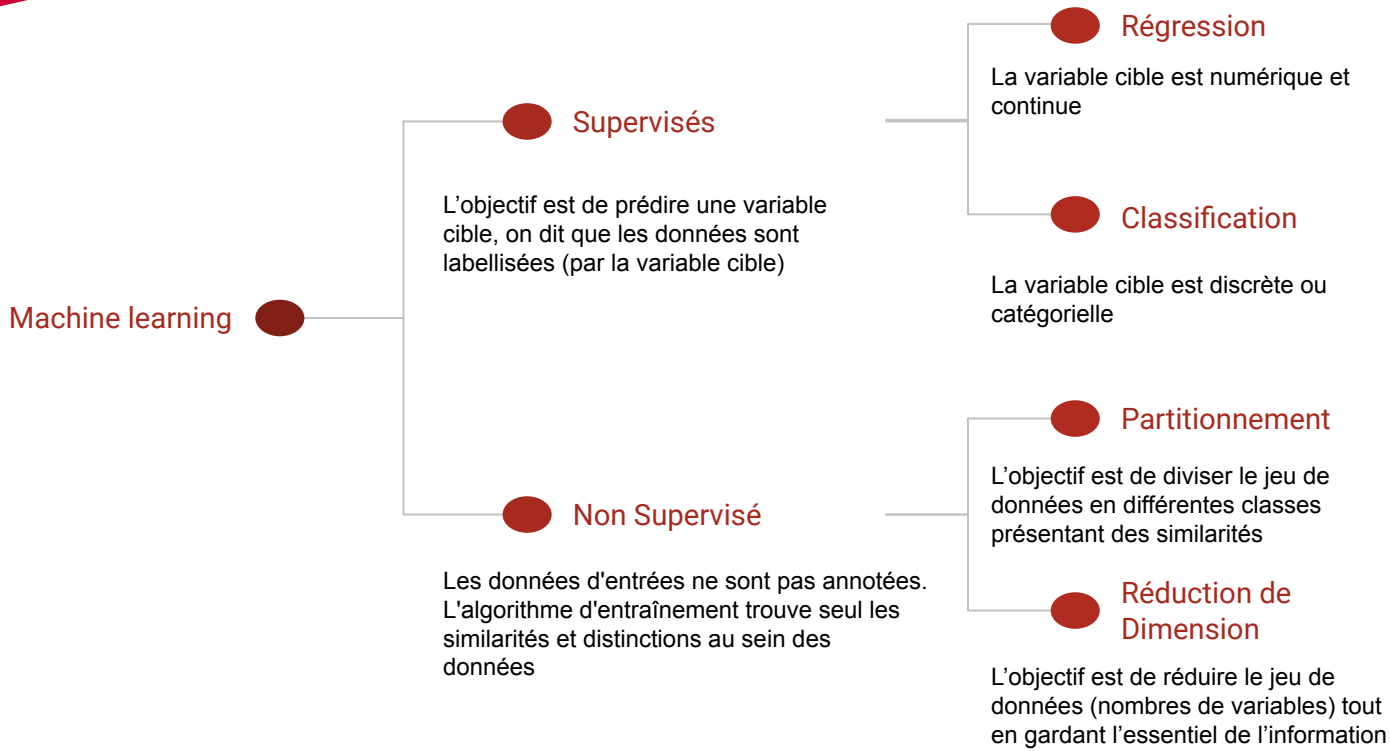
- **Le training set,** qui va nous permettre d’entraîner notre modèle et sera utilisé par l’algorithme d’apprentissage. C'est celui dont on a parlé depuis le début.

- **Le testing set**, qui permet de mesurer l’erreur du modèle final sur des données qu’il n’a jamais vues. On va simplement passer ces données comme s'il s'agissait de données que l’on n'a encore jamais rencontrées (comme cela va se passer ensuite en pratique pour prédire de nouvelles données) et mesurer la performance de notre modèle sur ces données.

C'est à vous de définir la proportion du dataset que vous souhaitez allouer à chaque partie. En général, les données sont séparées avec les proportions suivantes : **80 % pour le training set et 20 % pour le testing set.**

(**Le validation set** – qui permet de mesurer l’erreur de prédiction pour choisir entre plusieurs modèles – est aussi souvent utilisé. Nous l'étudierons dans les prochains cours.)

Les grands types de problèmes du ML



Les principaux algorithmes

Algorithmes	Problèmes correspondant
Régression Linéaire: Cherche à établir une relation linéaire entre deux variables ou davantage	Régression (Supervisé)
Régression logistique (LOGIT): permet d'étendre la régression linéaire aux données discrètes et catégorielles.	Classification (Supervisé)
Classification and Regression Trees (CART): Labellise une instance en suivant le processus d'un arbre de décision. Ces arbres peuvent être combiné de manière séquentielle (**XGboost**) ou non (**Random Forest**)	Régression et Classification (Supervisé)
K nearest neighbors (KNN): Prédire une valeur cible à partir des valeurs des instances qui lui ressemble	Régression et Classification (Supervisé)
Support Vector Machine (SVM): Généralisation des **classificateurs linéaires** qui cherchent à diviser un jeu de données en fonction de leur ressemblance (les multiples divisions ne sont pas séquentielles)	Régression et Classification (Supervisé)
Réseaux de Neurones: Analyse de la données à la fois de manière parallélisée et séquentielles	Régression et Classification (Supervisé) Réduction dimension et partitionnement
K mean: divise un jeu de données en k groupes, souvent appelés clusters, de façon à minimiser une certaine fonction	Partitionnement (Non Supervisé)
Analyse en composantes principales (ACP): transforme des variables corrélées en nouvelles variables décorrélées les unes des autres	Réduction de dimension (Non Supervisé)

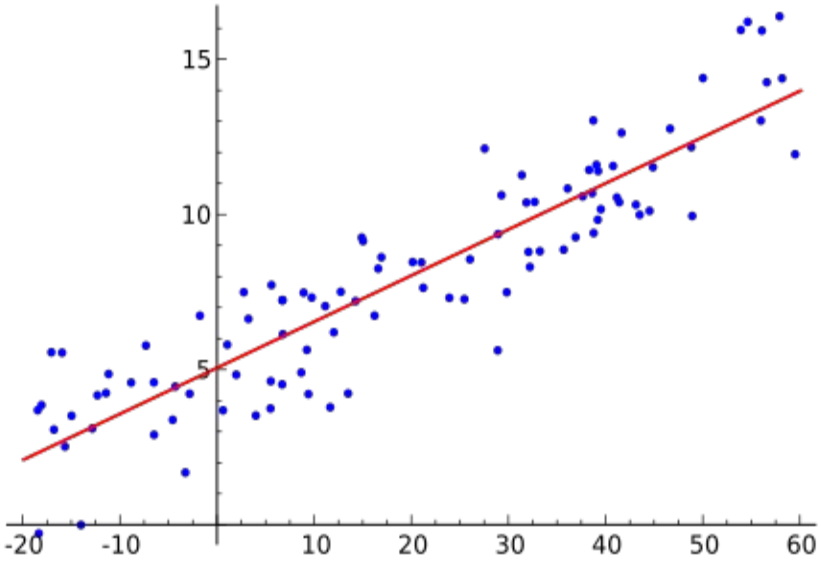
La régression linéaire simple

SIMPLON.CO

La régression linéaire: le principe

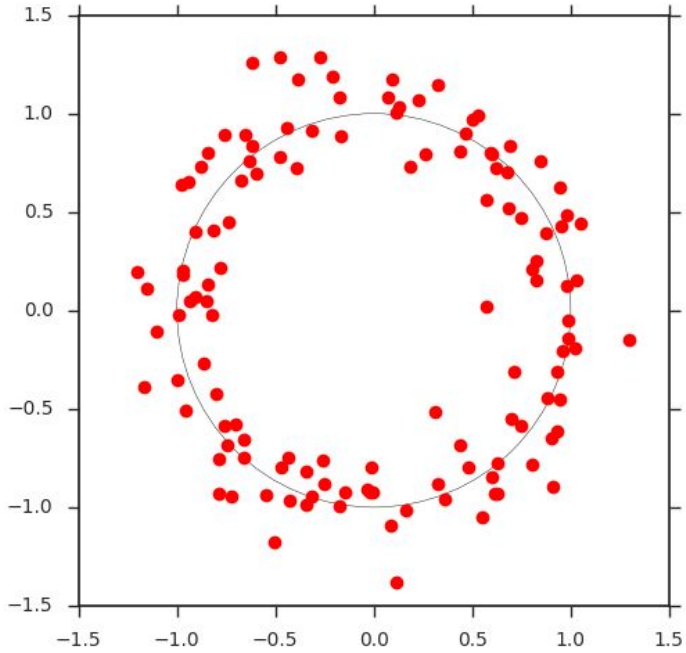
On considère que toute donnée (point bleu) est la combinaison d'un modèle sous jacent (droite rouge) et d'un certain bruit ou aléa indépendant. (écart entre le point et la droite)

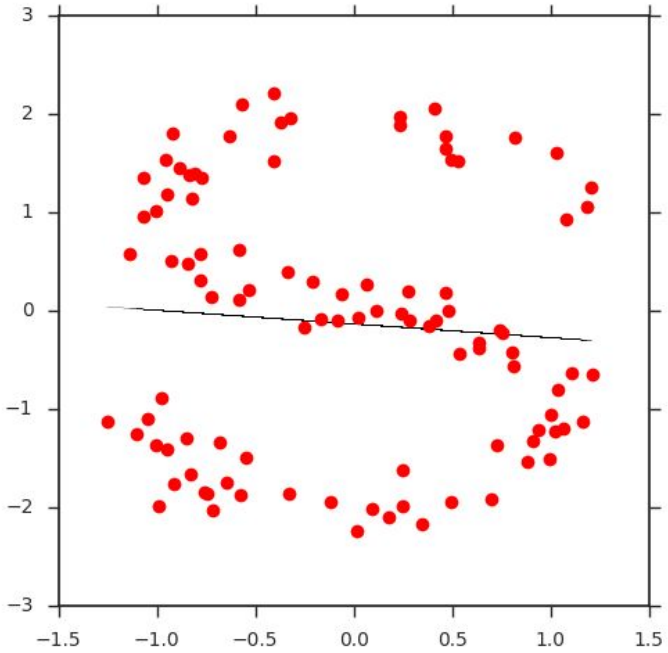
Dans le cas de la régression linéaire, le modèle sous jacent est une droite.



La régression linéaire

Il est important de choisir le modèle sous jacent adapté aux données. Dans les cas suivants il serait impossible de résumer convenablement les nuages de points à l'aide d'une droite:





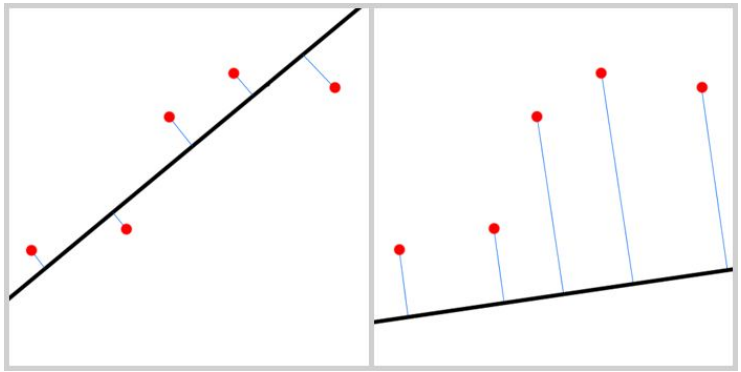
Fonction loss

En apprentissage supervisé, la notion principale est celle de **perte d'information** (loss) due à l'approximation. Elle détermine à quel point notre modélisation du phénomène perd de l'information par rapport à la réalité observée à travers les données d'exemple.

L'apprentissage se résume souvent à une méthode itérative qui converge vers un minimum de cette fonction. Plus la perte d'information diminue, plus on se rapproche de la réalité et meilleur est notre modèle.

Exemple de fonction de perte : L'erreur quadratique

C'est la distance euclidienne (trait bleu) entre la donnée (point rouge) et le modèle (**ligne noire**)



Régression linéaire: aspect théorique

On a un jeu de N données, portant sur le loyer “y” et la surface “x” d’appartements. On cherche un modèle permettant d’expliquer le loyer “y” par la surface ‘x”

On cherche donc une droite qui ferait le lien entre x et y. L’équation d’une droite s’écrit: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Minimiser l’écart entre nos données réelles et cette droite, revient à écrire: $\text{Argmin}_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2$

On cherche alors le point où la dérivée (selon β0 et β1) et on obtient après simplification:

$$\begin{cases} \hat{\beta}_1 = \dfrac{\text{cov}(X, Y)}{\text{var}(X)} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

Empiriquement β0 et β1 se calculent à partir d’une estimation de la matrice de variance covariance.

Régression linéaire: évaluation de la modélisation

Trois indicateurs clefs (vrais au delà de la régression linéaire)

La variation expliquée par la régression (Sum of Squares Explained [par la régression]). C'est la somme des différences entre la valeur prédite et la valeur moyenne observée.

$$\mathrm{SSE} = \mathrm{SCE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

La variation expliquée par les résidus (Sum of Squared Residuals). C'est la différence entre la valeur prédite et la valeur théorique.

$$\mathrm{SSR} = \mathrm{SCR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

La variation totale (*Sum of Squares Total*) correspond à la variance de y*n. D'après le calcul:

SST= $\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i).$

Or dans le cas de la régression le linéaire simple, le dernier élément est souvent nul. On a donc:

SST= SSE + SSR

Une partie de la variance est expliqué par le modèle, une partie reste résiduelle.

Régression linéaire: évaluation de la modélisation

On comprend que plus un modèle sera performant plus la proportion de la variance sera forte et plus celle expliquée par le résidu sera faible.

C'est tout le principe du **Coefficient de détermination R2**

$$R^2 = \frac{\text{SSE}}{\text{SST}} = \frac{\text{SST} - \text{SSR}}{\text{SST}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Le coefficient de détermination varie entre 0 et 1. Lorsqu'il est proche de 0, le pouvoir prédictif du modèle est faible et lorsqu'il est proche de 1, le pouvoir prédictif du modèle est fort.

Puisque le calcul du R2 passe par celui du SST pourquoi est ce qu'on ne se sert pas directement du SST (distance entre la valeur prédite et la valeur réelle) directement comme étiquette? Le problème c'est que cette valeur augmente avec le nombre de données, il faut donc la retravailler un peu. C'est le principe de **RMSE (racine de l'erreur quadratique moyenne)**:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2}$$

Plus le RMSE est faible, meilleure est la prédiction. Cependant le RMSE dépend de l'ordre de grandeur du jeu de données, il peut donc être plus difficile à interpréter et il ne rend pas possible les comparaisons entre deux jeux de données qui n'ont pas le même ordre de grandeur

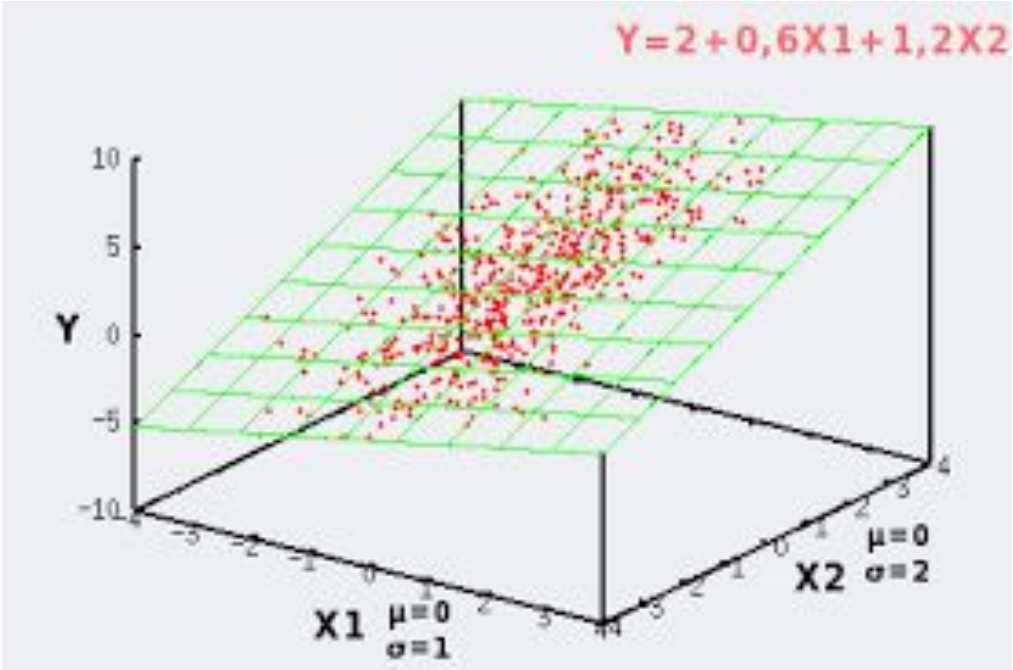
La régression multilinéaire

SIMPLON.CO

Régression multilinéaire: le principe

La régression linéaire multiple est une généralisation de la régression linéaire simple pour décrire les variations d'une variable endogène (variable cible) associée aux variations de plusieurs variables exogènes (variable explicative)

Graphiquement, elle consiste à résumer un nuage de point à N dimension par un hyperplan linéaire (à N-1 dimension):



Régression multilinéaire: aspect théorique

De nouveau on cherche à modéliser un loyer, mais plus seulement à l'aide de la surface x1 mais également à l'aide d'autres variables telles que x2 (la présence d'un balcon), x3 le nombre de pièces, x4 (le nombre de commerces dans un rayon de 1km) ….

L'équation de l'hyperplan linéaire se note ainsi: $Y_i = a_0 + a_1 X_{i1} + a_2 X_{i2} + \ldots + a_p X_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n$

n correspond aux nombres de données dans l'échantillon et p au nombre de variables explicatives.

Pour décrire ce problème, on adopte souvent une écriture matricielle:

$$\begin{cases} y_1 = a_0 + a_1 x_{1,1} + \ldots + a_p x_{1,p} + \varepsilon_1 \\ y_2 = a_0 + a_1 x_{2,1} + \ldots + a_p x_{2,p} + \varepsilon_2 \\ \ldots \\ y_n = a_0 + a_1 x_{n,1} + \ldots + a_p x_{n,p} + \varepsilon_n \end{cases}$$

peut s'écrire:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Ce qui donne: $y = Xa + \varepsilon$ avec

- y est de dimension $(n, 1)$
- X est de dimension $(n, p+1)$
- a est de dimension $(p+1, 1)$
- ε est de dimension $(n, 1)$

Régression multilinéaire: les variables explicatives

Il est possible d'intégrer comme variables explicatives à une régression linéaire multiple des variables numériques continues et des variables discrètes ou catégorielles. Cependant les variables catégorielles doivent être transformées en variables **"dummies".** C'est à dire qu'une variable "couleur des cheveux" prenant comme valeur "brun", "blond" "roux" sera transformé en trois variables "brun", "blond", "roux" prenant comm valeur 0 ou 1. (get_dummies (Pandas) ou OneHotEncoder (Scikit-learn))

Les modèles linéaires simples et multilinéaires possèdent de nombreuses hypothèses théoriques souvent ignorées dans la pratique. Cependant dans le cadre multilinéaire une des hypothèses a des conséquences pratiques, il s'agit de l'**absence de colinéarité entre les variables explicatives**.

Au moment où je vais évaluer mon modèle et les différentes variables que j'y ai intégré, si variables explicatives sont corrélées alors l'une d'entre elle (ou même les deux, voire le modèle lui même) pourrait apparaître comme non significative alors qu'elle l'est dans la réalité.

Il convient donc avant de construire son modèle linéaire d'étudier la corrélation entre les variables explicatives. Dans mon exemple le nombre de pièces semble pouvoir être fortement corrélé à la surface de l'appartement, je dois donc explorer la corrélation entre ces deux variables.

Concrètement si deux variables explicatives ont un **R2>0,7** il faut à minima être vigilant mais sans doute ne pas les garder dans le même modèle

Il est possible d'évaluer la corrélation linéaire entre une variable continue et une variable catégorielle à l'aide du rapport de corrélation de l'ANOVA

Régression multilinéaire: aspects théorique et évaluation

Pour trouver l'hyperplan linéaire optimal on va de nouveau utiliser la méthode des moindres carrés, à savoir trouver les paramètres a0, a1, … ap qui minimisent la distance euclidienne entre mes données et l'hyperplan.

$\hat{a} = (X^T X)^{-1} X^T Y$ est l'estimateur qui minimise cette valeur

XT étant la transposé de X et â de dimension p+1 est l'estimateur empirique de a

L'évaluation du modèle se fait de nouveau à l'aide du coefficient de détermination **R2** et l'erreur quadratique moyenne **MSE** (ou RMSE). Cependant comme le R2 augmente avec le nombre de variable explicative, on peut lui préférer le **R2 ajusté**. Cependant, dans la pratique on privilégie des modèles régularisés.

Contrairement à la régression linéaire simple, dans le cadre de la régression multilinéaire, on se pose la question: Est ce que toutes mes variables sont vraiment utiles dans mon modèle? Cela s'appelle le test d'hypothèse, on va s'intéresser à la significativité du modèle et de ses variables.

Test d'hypothèses:
Q1: Est ce qu'au moins une des variables est utiles dans mon modèle

On test l'hypothèse: $\mathbf{H0 : a1 = a2 = ... = ap = 0}$

L'idée derrière H0 c'est : "Certe quand on calcule l'hyperplan optimal, on trouve que ces valeurs ne sont pas nulles, mais en vérité si ces coefficients étaient tous égaux à 0 ça serait statistiquement à peu près la même chose." énoncé autrement: 'est ce qu'on a eu de la chance de trouver une relation linéaire ou est ce bien le modèle sous jacent"

On va tester cette hypothèse, si après le test, on est contrait de la rejeter alors il faudra accepter son alternative:
H1 : Au moins un des a est différent de 0 (il y a au moins dans mon modèle une variable qui est réellement utile pour prédire le loyer)

Pour évaluer l'hypothèse nulle (H0)
on calcule la statistique F qui suit une loi de Fisher.

$$F_{calc} = \frac{\frac{R^2}{p}}{\frac{1-R^2}{n-p-1}}$$

On prend notre décision en fonction de la **p-value** associée à F pour notre modèle. La p-value est la probabilité dans le cas où l'hypothèse nulle seraient vraie d'obtenir une telle valeur pour F voire une valeur plus extrême.

Concrètement si la **P-value<0,05** on considère que l'hypothèse nulle est rejetée. C'est à dire qu'on accepte l'hypothèse alternative et donc qu'on considère qu'au moins une des variables à un intérêt.

(Si la p-value est entre 0,001 et 0,05 on ne s'enflamme pas, le modèle est pas incroyable.

Test d'hypothèses:
Q2: Est ce que toutes mes variables sont utiles dans mon modèle?

Pour chaque coefficient a1, a2,...ap, on teste: **$H0 : aj = 0$** (pour j entre 1 et p)

L'idée derrière H0 c'est : "Est ce que l'influence de cette variable dans mon modèle est justifiée statistiquement ou est ce qu'il aurait très bien pu être nul"

On va tester ces hypothèses l'une après l'autre, si après un test, on est contraint de rejeter l'une de ces hypothèse, alors il faudra accepter son alternative:
H1 : aj est différent de 0 (L'influence de ma variable dans mon modèle est justifiée)

Pour évaluer l'hypothèse nulle (H0)
on calcule la statistique t qui suit une loi de Student. $$t = \frac{\hat{a}_j - a_j}{\hat{\sigma}_{\hat{a}_j}} \sim \mathrm{T}(n - p - 1)$$

On prend notre décision en fonction de la p-value associée à t pour notre coefficient. La p-value est la probabilité dans le cas où l'hypothèse nulle seraient vraie d'obtenir une telle valeur pour t voire une valeur plus extrême.

Concrètement si la P-value<0,05 on considère que l'hypothèse nulle est rejetée. C'est à dire qu'on accepte l'hypothèse alternative et donc qu'on considère que notre coefficient est significativement différent de 0 donc que l'influence de la variable explicative est statistiquement justifiée.

(Si la p-value est entre 0,001 et 0,05 on ne s'enflamme pas, le modèle est pas incroyable.

Résumé des tests hypothèses:

R2: Vous apprend à quelle point la variation de vos données est expliquée par votre modèle: est ce que le modèle est bon ou non.

P-value: (du modèle ou du coefficient) Vous dit à quel point votre modèle est fiable ou s'il est juste dû au hasard.

Pour un modèle	R2 > 0,7	R2 < 0,7
p-value de F < 0,05	Modèle bon et fiable	Modèle pas terrible mais fiable
p-value de F > 0,05	Modèle bon mais pas fiable, donc pas bon	Modèle ni bon ni fiable

Pour une variable	Amélioration du R2 ajusté	Régression du R2 ajusté
p-value de t < 0,05	On conserve la variable	On rejette la variable
p-value de t > 0,05	On rejette la variable	On rejette la variable

Remarque: Une variable explicative est souvent peu significative, soit parce qu'elle n'a pas une influence linéaire sur la variable cible, soit parce qu'elle est corrélée linéairement avec une autre variable explicative

Interprétation des coefficients:
L'Influence des coefficients

La régression linéaire donne l'avantage de fournir une interprétation des coefficients du modèle. La lecture des coefficients dans la régression multilinéaire fonctionne de la même manière que celle dans la régression simple. Cependant cette lecture ne fournit pas directement l'influence de chaque variable dans le modèle car les différents coefficient dépendent de l'ordre de grandeur des variables.

- Exemple: même si la surface à plus d'influence que la présence d'un balcon, le coefficient rattaché au balcon sera sans doute plus important car la variable varie entre 0 et 1 quand la surface varie entre 0 et 1000.

Pour connaître l'influence des coefficients dans la régression, il faut que l'ensemble des variables partagent le même ordre de grandeur. Pour parvenir à cela il faut les transformer en **variables centrées réduites:**

- **centrée:** la moyenne est nulle (on soustrait à chaque valeur la moyenne)
- **réduite:** l'écart type est égal à 1 (on divise toutes les valeurs par l'écart type)

Une fois la variables centrée réduite, il faut à nouveau appliquer la régression linéaire. Pour cette nouvelle régression linéaire, plus un coefficient aura une valeur absolue important, plus son role dans le modèle sera important.

Interprétation des coefficients:
La notion de “contrôle” et de “toutes choses égales par ailleurs”

Dans les études scientifiques mais également dans toute analyse de données, se pose souvent la question de la corrélation réelle entre deux variables.

Exemple: J’étudie les revenus en fonction du genre. Je constate qu’il y a un écart entre les hommes et les femmes. Je peux alors me poser cette question:

- est ce que les femmes sont en effet moins payées que les hommes à niveau d’étude identique
- ou est ce que les femmes ont en général un niveau d’étude plus faible et c’est ce qui explique qu’au global, elles sont moins bien payées

Ce que j’aimerais, c’est pouvoir étudier les revenus des hommes et des femmes à niveau d’étude identique. On appelle ça:

- étudier la corrélation entre les variables “revenu” et “genre” en **contrôlant** par la variable “niveau d’étude”
- étudier la corrélation entre les variables “revenu” et “genre” **toutes choses égales par ailleurs** (on suppose ici que le niveau d’étude et le seul paramètre pouvant biaiser notre étude)

Techniquement, il suffit pour ça de réaliser une régression linéaire multiple. Dans une régression multiple, les variables explicatives se contrôle les unes les autres. Leurs coefficients nous informent donc de l’influence des variables indépendamment des autres. C’est pour cela qu’il est nécessairement que les variables explicatives ne soit pas corrélées entre elles car sinon leurs coefficients s’annulent entre eux.

Régularisation: l’enjeu

Régularisation: l'enjeu

Rappel: l'erreur quadratique (MSE) est la différence entre la valeur réelle et la valeur prédite se note ainsi:

$$MSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \mathbb{E}\left((y - \hat{y})^2 \right)$$

E correspond à l'espérance. (la moyenne d'une variable aléatoire)

Ce qui donne: (après simplification du 3ème terme de l'identité remarquable

$$MSE = \mathbb{E}\left((y - \hat{y})^2 \right) = \mathbb{E}\left((y + \bar{y} - \bar{y} - \hat{y})^2 \right) = \mathbb{E}((y - \bar{y}))^2 + \mathbb{E}\left((\hat{y} - \bar{y})^2 \right)$$

On retrouve ainsi le fameux arbitrage entre le biais et la variance:

$$MSE = Biais^2 + Variance$$

- **Le Biais:** C'est l'écart entre l'espérance de la prédiction et la vraie valeur. Indique les insuffisances intrinsèques du modèle qui n'est pas assez précis
  - Il est du à des variables explicatives manquantes, ou forme de la relation non captée.
- **La Variance**: C'est la dispersion de la prédiction autour de sa propre espérance. Témoigne de l'instabilité du modèle, sa dépendance aux fluctuations de l'échantillon d'apprentissage. (surapprentissage)
  - Elle est due à des problèmes de colinéarité et surdimensionnalité (trop de variables explicatives)

Pour limiter le surapprentissage (erreur due à la variance), il est possible d'adapter le modèle mais cette adaptation à souvent tendance à augmenter le biais.

Régularisation: les différentes techniques

Adapter le modèle consiste concrètement à diriger (réguler) un peu plus fermement la modélisation en imposant des contraintes sur les paramètres estimés de la régression (contraintes sur les valeurs que pourront prendre les coefficients dans leur ensemble pour éviter qu'elles soient totalement erratiques)

Les différentes techniques:

1. **La régression RIDGE**
   - Il faut dans ce cadre centrer réduire les variables. On limite ensuite la plage de valeur que prennent les coefficients (skrinkage: la somme des coefficients carré (norme L2) doit être inférieure à une certaine valeur)). Il faut pour cela déterminer un coefficient de pénalité λ empiriquement (grid search) ou par le calcul. (Plus λ est grand, plus le modèle est régularisé).
   - Intuitivement: plus un paramètre a de l'importance, moins les autres paramètres peuvent en avoir, on peut ainsi avoir beaucoup de paramètre ayant un rôle déterminant (réduit la dimension)
2. **La régression LASSO**
   - Tout comme RIDGE, LASSO utilise un procédé de shrinkage (sur la somme des valeurs absolues cette fois-ci (norme L1))
   - LASSO permet de sélectionner les variables avants de calculer le modèle. (modèle parcimonieux)
3. **Elastic net : combinaison de RIDGE et de LASSO et donc double avantage;**
   - Capacité de sélection de variables du LASSO conservée (coefficients nuls) : exclusion des variables non pertinentes
   - Groupe de variables prédictives corrélées, partage des poids (comme Ridge) et non plus sélection arbitraire
   - Inconvénient: il y à présent deux hyperparamètres: λ et a

Fiche d’identité de la régression linéaire

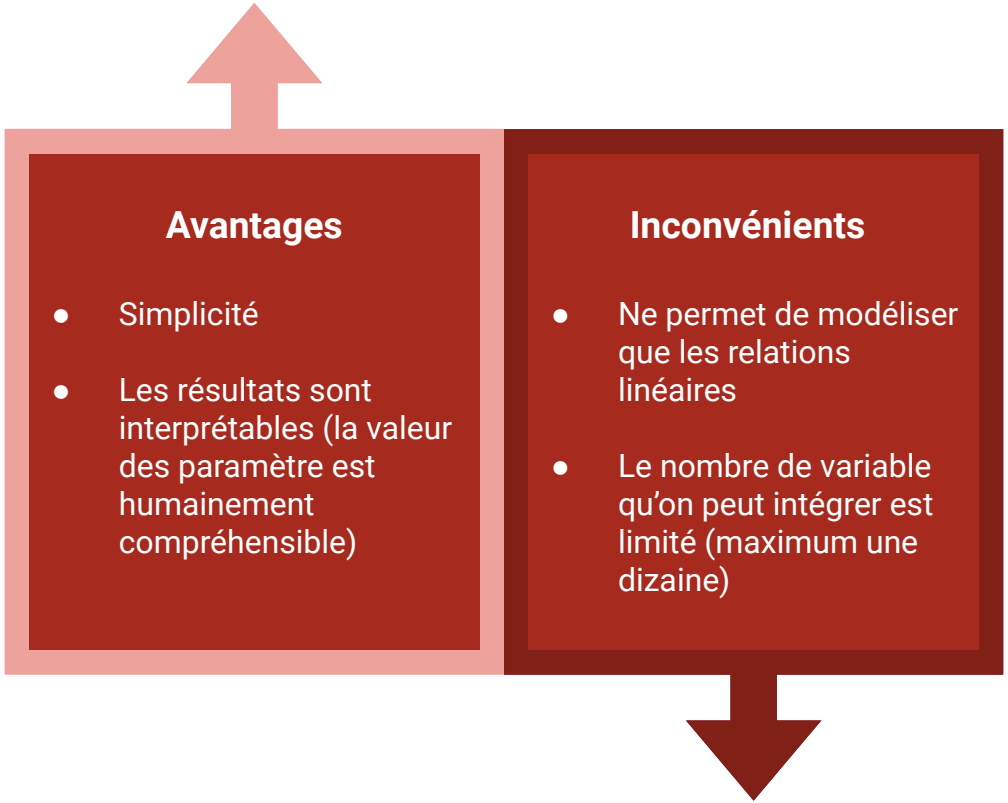
Résumer en une fiche, quand utiliser ce modèle et comment s’en servir.

Nom	Régression linéaire.
Problème résolu	Régression (supervisée)
Type de données	Continues et catégorielles (mais transformée en dummies, encodée)
Contraintes	• Absence de colinéarité entre les variables explicatives • Nombre limité de variables explicatives (<10) • Sensibilité aux valeurs aberrantes • Relation linéaire entre la variable cible et les variables explicatives.
Avantages	• Fonctionne sur un nombre limité de valeur • Simplicité de modélisation • Interprétation humaine des

Fiche d’identité de la régression linéaire

Nom	Régression linéaire.
Problèmes résolus	Régression
Contraintes	Corrélation linéaire entre variable cible et variable explicative Absence de colinéarité entre les variables explicatives Limites des variables explicatives à une dizaine
Avantages	Interprétabilité des résultats
Mesure d’évaluation	R2 et MSE
Techniques de Régularisation	Ridge, Lasso, ElasticNet

Conclusion: Avantages et inconvénients de la régression linéaire



Usage réelle de la régression linéaire:

- La régression linéaire même régularisée est trop simple pour fournir un bon modèle de machine learning
- Cependant on l'utilise souvent a posteriori pour interpréter les résultats d'un autre modèle. En particulier elle permet de comprendre clairement le rôle de chaque attribut et la significativité de ces derniers.
- Elle peut également servir lors de la phase exploratoire pour se donner une première idée de la relation entre ses données.

La régression Logistique

SIMPLON .CO

La régression logistique: le principe

La régression logistique est un algorithme de classification. Le but est de prédire à quelle “classe” chaque instance appartient en fonction de la valeur qu’elle prend pour chaque variable explicative.

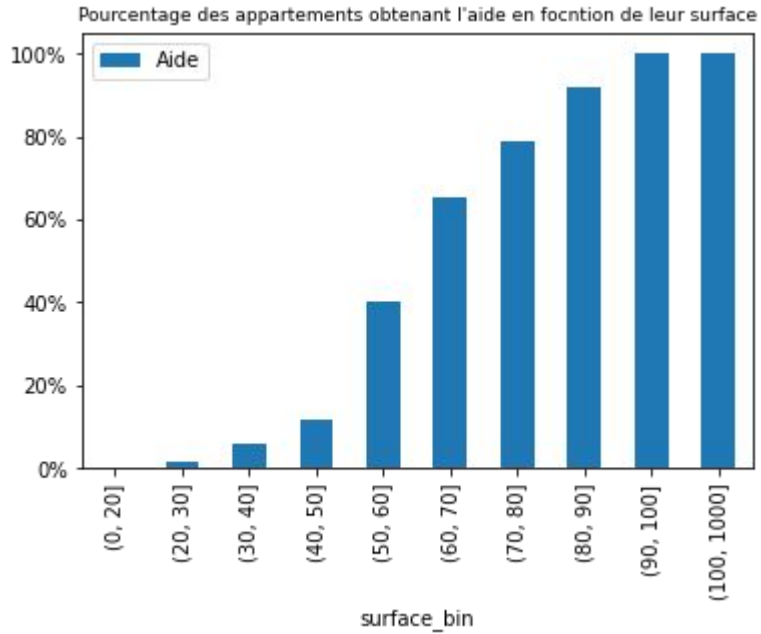
Prenons un exemple proche de ce que nous avons vu précédemment:

“La mairie de Paris a décidé de mettre en place une aide pour tous les loyers supérieurs à 2000 euros. Je souhaite savoir si pour mon appartement je pourrais bénéficier de cette aide. Malheureusement pour moi, (mais heureusement pour les vertus pédagogiques de cet exemple), je n’ai plus accès dans ma base de données aux informations concernant le loyer. Je sais uniquement pour chaque appartement, leur surface et s’ils bénéficient de l’aide (oui ou non).”

Je vais donc devoir modéliser **l’obtention de l’aide de la mairie Y** (Y=1 s’il bénéficie de l’aide, Y=0 s’il n’en bénéficie pas) **à partir de X, la surface de mon appartement**.

La régression logistique: le principe

Pour savoir si je peux obtenir l'aide en fonction de la surface de mon appartement, une première approche intuitive serait discrétiser la variable surface, c'est à dire la diviser en plusieurs tranches et pour chaque calculer le pourcentage d'appartement qui obtienne l'aide dans cette tranche:



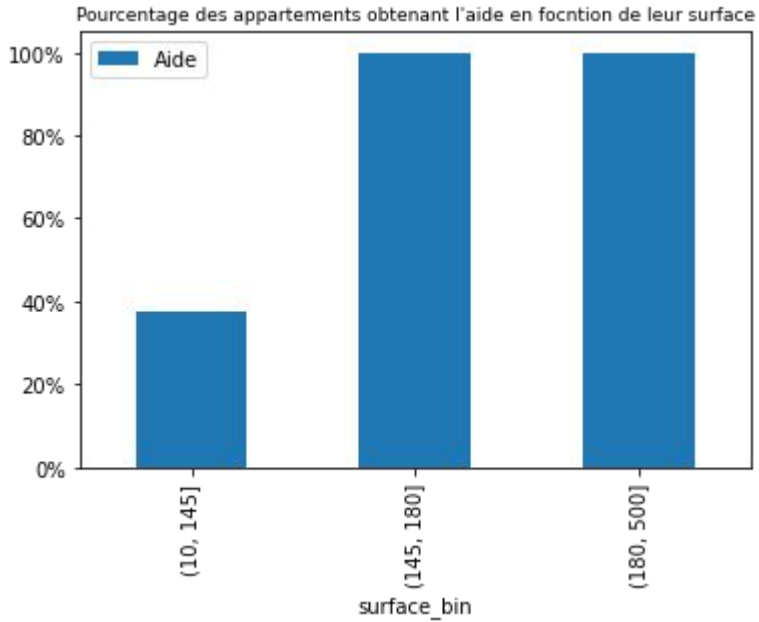
Je peux en déduire que pour chaque tranche, le pourcentage d'appartements obtenant l'aide correspond à la probabilité d'obtenir l'aide si on a appartient à cette tranche.

Si mon appartement fait 77m2, comme 80% des appartement entre [70m2-80m2] obtiennent une aide je peux me dire que j'ai 80% de chance d'obtenir une aide.

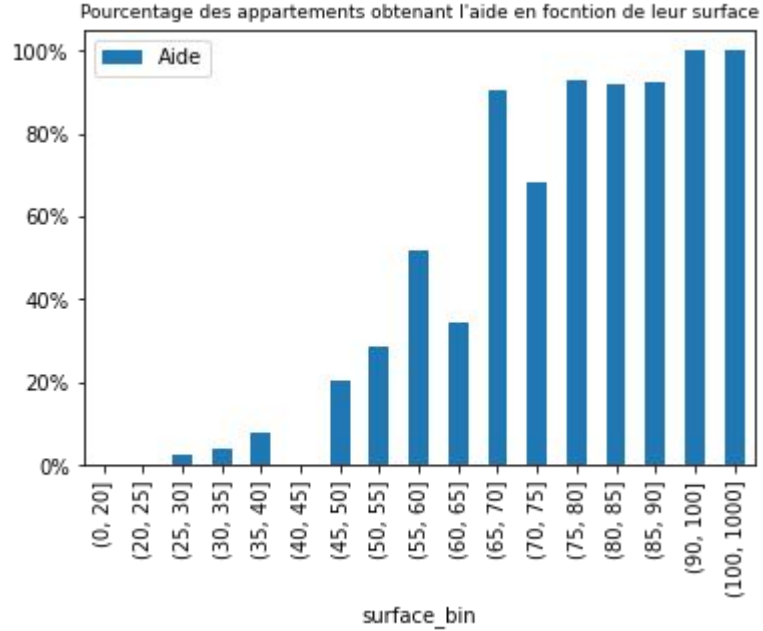
Mathématiquement:
P(Y=1/X=[70-80]=80%

La régression logistique: le principe

L’approche semble pertinente mais il reste une question à résoudre? Comment est ce que je dois découper mes tranches? Plus j’ai de tranche plus mon résultat sera précis, cependant au bout d’un moment, si je fais trop de tranches, il n’y aura plus accès d’appartements par tranche pour avoir une moyenne représentative.



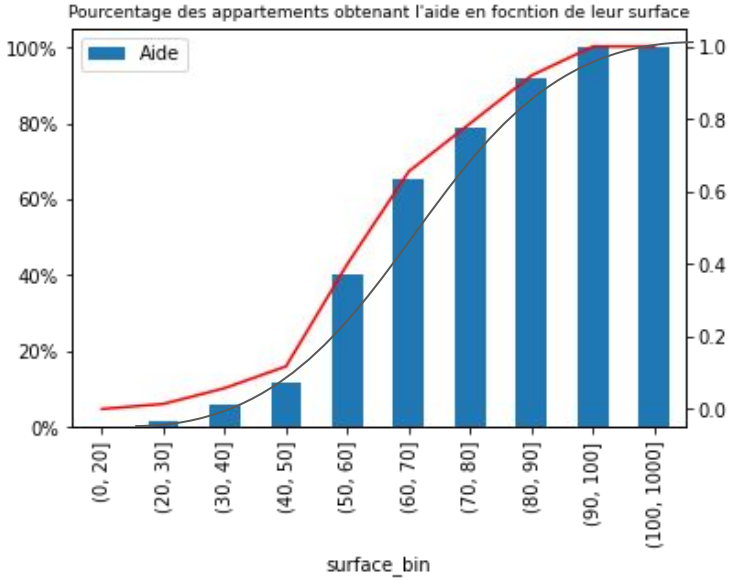
Exemple avec 3 tranches



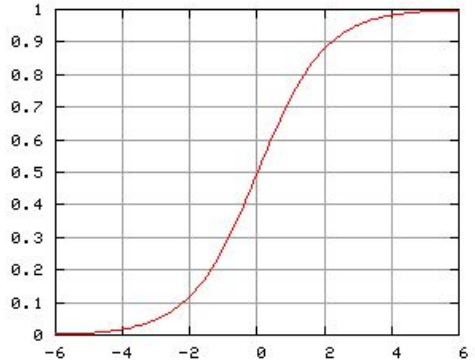
Exemple avec 17 tranches

La régression logistique: le principe

Il faut donc adopter une approche plus statistique.
L'objectif ne sera plus de rechercher P(Y=1 / X appartient à une tranche) mais directement P(Y=1 / X)
Cela revient à rechercher le modèle qui s'approche de la courbe en S rouge



Or un tel modèle existe c'est le **modèle LOGIT**. (Quand on verra les réseaux de neurones on appellera cela une **sigmoïde**).
Sa fonction de répartition ressemble à cela:



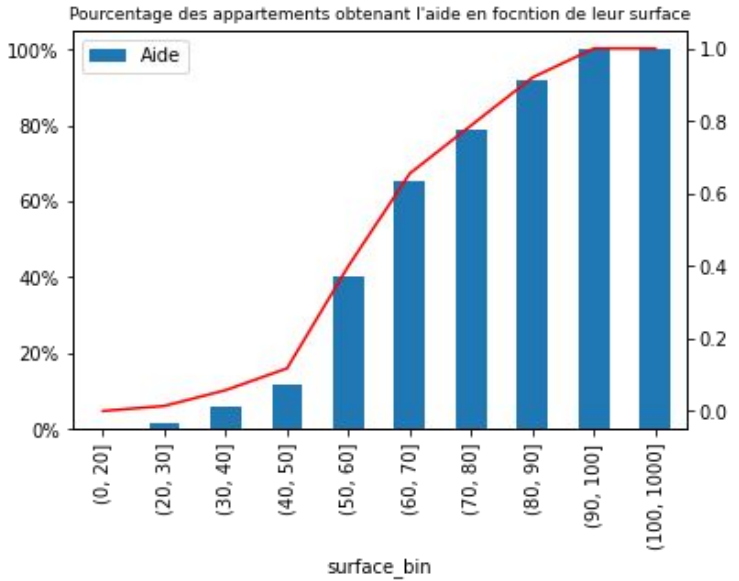
et elle a pour équation: $$P(Y=1 \mid X) = \frac{e^{a_0 + a_1 \cdot X}}{1 + e^{a_0 + a_1 \cdot X}}$$

La régression logistique: le principe

L'objectif est donc de trouver les paramètres a0 et a1 tel que la sigmoïde d'équation: $P(Y=1 \mid X) = \frac{e^{a_0 + a_1 \cdot X}}{1 + e^{a_0 + a_1 \cdot X}}$ colle le mieux à nos données.

Nous l'avons vu, lorsque l'on prend trop de tranches, nous ne disposons pas assez de données pour obtenir une moyenne significative par tranche. En d'autres termes, dans un échantillon empirique, il est impossible d'avoir pour tout X la probabilité P(Y=1 | X) associée. On ne peut donc utiliser comme fonction de perte la distance euclidienne entre ma probabilité empirique et ma probabilité prédite (estimateur des moindres carrées.

Nous sommes contraint d'utiliser une autre fonction de perte. Celle qu'on utilise dans le cadre de la régression logistique est **le maximum de vraisemblance**.



La régression logistique: le maximum de vraisemblance

Soit θ un vecteur de paramètres et x un ensemble d'observations d'un variable aléatoire X.

La **vraisemblance**: L(θ | x) est la probabilité d'obtenir les valeurs x en supposant que notre variable aléatoire X ait pour paramètre θ.

Mathématiquement: $L(\theta|x) = p_\theta(x) = P_\theta(X = x)$

Prenons l'exemple d'un tirage à pile ou face. On veut savoir si notre pièce est truquée ou non. Soit p la probabilité d'obtenir pile. (Donc ici θ=p, c'est le paramètre de notre variable aléatoire X (lancé de pièce):

- Si p=0,5 la pièce est bonne
- Si p=0,4 ou p=0,6, la pièce est truquée.

Je fais 10 lancées x = (P,P,F,F,PF,P,P,P,F) 6 piles et 4 faces. Avec p la probabilité de faire Pile.

- La probabilité de faire (P,P) est p*p soit p^2
- La probabilité de fait (P,F) est p*(1-p)
- La probabilité de faire (P,F,F) est $p \cdot (1-p)^2$

On a donc: $L(\theta \mid x) = L(p \mid x) = P_p(X=x) = p^6 \cdot (1-p)^4 = \prod_{i=1}^{10} p^{1_{x_i=P}} \cdot (1-p)^{1_{x_i=F}}$

Avec $1_{x_i=P} = 1 \; quand \; x_i = Pile \; et \; 0 \; autrement$

La régression logistique: le maximum de vraisemblance

On calcule donc la vraisemblance à l'aide de notre équation: $L(p \mid x) = p^6 \cdot (1-p)^4$

La vraisemblance permet de calculer quel est le paramètre le plus vraisemblable:

- L(0,4 | x) = 5,3 e-4 (0,00053)
- L(0,5 | x) = 9,8 e-4 (0,00098)
- L(0,6 | x) = 1,2 e-3 (0,00120)

Le paramètre le plus vraisemblable est donc 0,6, on en conclut d'après l'estimateur du maximum de vraisemblance que notre pièce est biaisée.

La régression logistique: le maximum de vraisemblance

Rappel

$$L_{\theta=p}(X=x) = P_p(X=x) = \prod_{i=1}^{10} P_p\left(x_i=P\right)^{1_{x_i=P}} \cdot \left(1 - P_p\left(x_i=P\right)\right)^{1_{x_i=F}} = \prod_{i=1}^{10} p^{1_{x_i=P}} \cdot (1-p)^{1_{x_i=F}}$$

Dans le cadre de la régression logistique, on en était resté au fait de rechercher un couple de paramètre (a0,a1) tel que la sigmoïde d'équation $P(Y=1 \mid X) = \frac{e^{a_0 + a_1 \cdot X}}{1 + e^{a_0 + a_1 \cdot X}}$ colle le mieux à nos données.

Cela revient à dire, rechercher les paramètres (a0,a1) telle que la P(Y=1 | X) soit la plus vraisemblable d'après les données observées. En termes mathématique on cherche donc les paramètres (a0,a1) qui maximisent:

$$L_{\theta=(a_0,a_1)}(Y=1 \mid X) = \prod_{i=1}^{n} P_{a_0,a_1}(Y=1 \mid X)^{1_{y_i=1}} \cdot \left(1 - P_{a_0,a_1}(Y=1 \mid X)\right)^{1_{y_i=0}}$$

Les paramètres a0 et a1 qui maximisent cette quantité sont les estimateurs du maximum de vraisemblance de la régression logistique.
(dans la pratique on maximise le logarithme de la fonction de vraisemblance car les calculs sont plus faciles, il n'y a pas de solution explicite à ce problème, on l'obtient donc de manière itérative (on appelle cela la descente de gradient)

La régression logistique: évaluation de la classification

La Matrice de confusion:
À partir de la matrice de confusion on peut calculer différents critères de performance

		Classe réelle	
		-	+
Classe prédite	-	True Negatives *(vrais négatifs)*	False Negatives *(faux négatifs)*
	+	False Positives *(faux positifs)*	True Positives *(vrais positifs)*

Cela permet d'évaluer la **qualité** du test:

- La **précision** **(**en anglais "precision"**)**, c'est-à-dire la proportion de prédictions correctes parmi les points que l'on a prédits positifs.
  - Précision = TP / (TP + FP)
- L'**accuracy** (en anglais exactitude): le taux des bonnes réponses = TN +TP / (TN+TP +FN + FP) (Moyenne des précisions pour chaque classe)

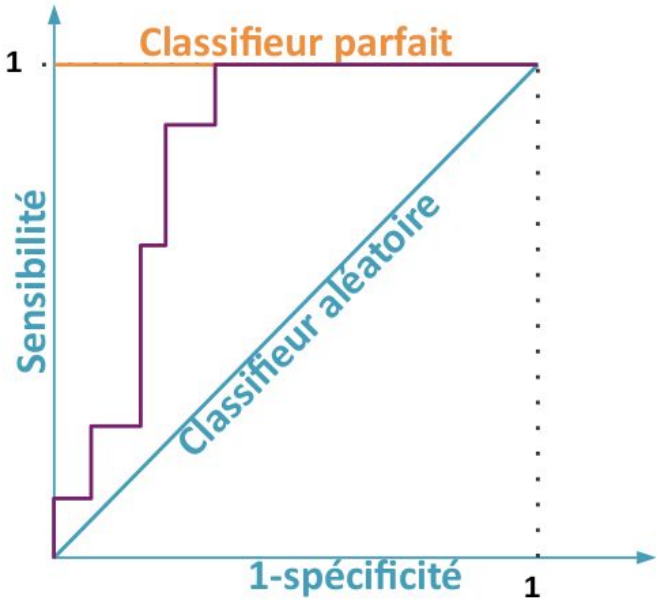
Ou **l'exhaustivité** de ses découvertes: le **rappel** ("recall" en anglais),

- la **sensibilité** ("sensitivity" en anglais), est le taux de vrais positifs, c'est à dire la proportion de positifs que l'on a correctement identifiés.
  - Rappel = TP / (TP +FN)
- la **spécificité** ("specificity" en anglais), qui est le taux de vrais négatifs (n'a de sens que dans le cas binaire)
  - Spécificité = TN / (FP + TN)

Ou un mixte des deux:

- La **F mesure**. Si on prédit que tous les cas sont positifs alors on aura un très bon rappel, a l'inverse si on prédit que peu de cas positif (seulement les meilleurs) on aura une très bonne précision. La F-mesure est un arbitrage entre les deux
  - f-mesure= 2* (Précision * Rappel / Précision + Rappel) = 2TP / (2TP + FP + FN)

La régression logistique: évaluation de la classification



Ordonné: la sensibilité: le taux de vrai positif (ou rappel des positifs):

- Proportion des 1 qu'on a réussi à trouver

Abscisse: L'antispécificité: le taux de faux positif

- Proportion des 0 qu'on n'a pas réussi à trouver.

En bas à gauche: seuil à 0%: je prédis que personne n'est 1, logiquement:

- Je ne trouve aucun de mes 1
- Il n'y a aucun 0 que je n'arrive pas à trouver

En Haut à droite: seuil de 100%: je prédis que tout le monde est 1:

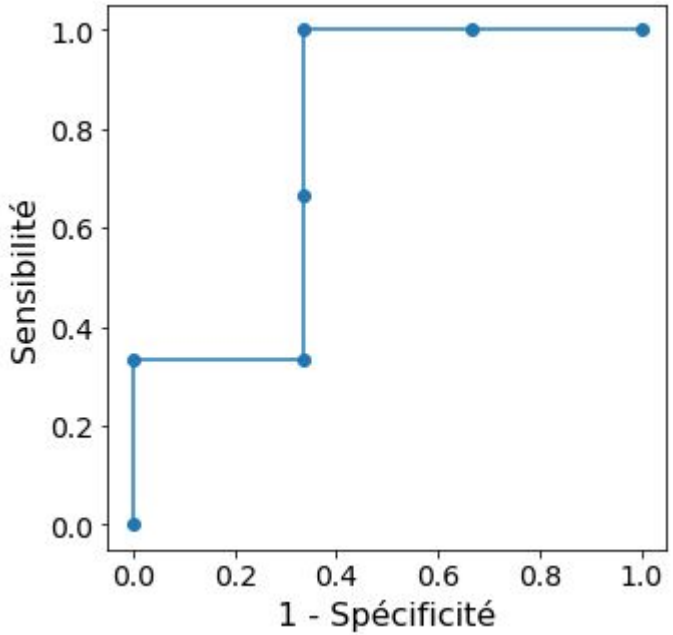
- Je trouve tous mes vrais 1
- Je ne trouve aucun de mes 0 (100% de non trouvés)

Quand mon seuil évolue entre 0% et 100%

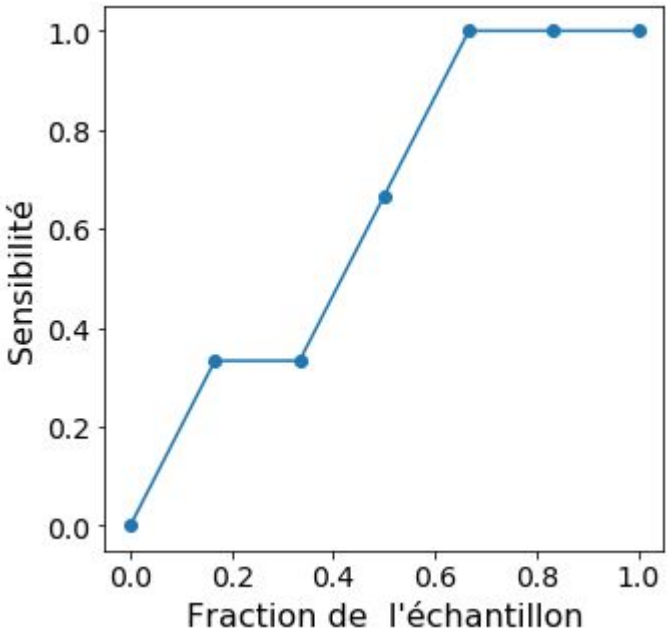
- Classificateur aléatoire: toute personnes que j'accepte à autant de chance d'être 0 que 1: donc les deux taux évolue en même temps
- Classificateur parfait: toute les premières personnes que j'accepte sont des 1, quand j'accepte pour mon premier 0, j'ai déjà trouvé tous mes 1
- Mon classificateur: quand j'ai accepté x% des mauvais par erreur, quel pourcentage y des bons j'ai déjà trouvé.

La régression logistique: évaluation de la classification

La courbe ROC pourrait être un outil pour choisir le seuil, le problème c'est qu'on ne peut pas le lire directement. Si on l'a vu le taux de faux positif est proportionnel au seuil. Ce n'est pas directement le seuil.
Pour déterminer un seuil, on peut représenter le taux de vrai positif en fonction du seuil. C'est la courbe lift.



Courbe ROC



Courbe LIFT

La régression logistique: évaluation de la classification

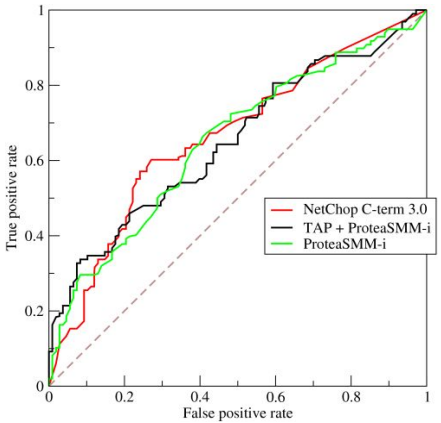
Mesure	Ce qui est évalué	Rapport au seuil. (proportion de personne à qui on dit oui).	Quand est ce qu'on s'en sert
Accuracy	La qualité	évolue avec le seuil Moins on accepte de personne plus il augmente.	Quand on a un seuil fixé et que les deux erreurs sont équivalents
Recall	L'exhaustivité	évolue avec le seuil Plus on accepte de personne, plus il augmente.	Quand la priorité est de ne louper aucun 1 et que le seuil est flexible
F mesure	La qualité et l'exhaustivité	Calculé pour un seuil donné.	Quand le seuil est fixé mais qu'on préfère ne pas louper de 1
AUC	La qualité et l'exhaustivité	Calculé pour tous les seuils.	Quand le seuil n'est pas fixé et que les deux erreurs sont équivalentes.

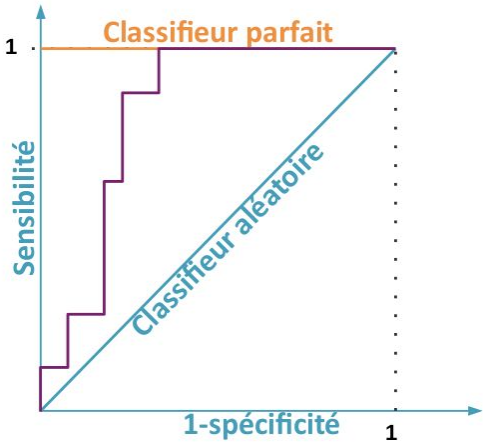
La régression logistique: évaluation d'un score

Dans le cadre d'une classification binaire (0 ou 1). Le score d'une instance correspond à sa probabilité d'être égale à 1
Il est donc possible à partir du score de classer l'ensemble des instances de la plus susceptible d'être égale à 1 à la plus susceptible d'être égale à 0.
La classification finale est souvent réalisée à partir de ce score et d'un **seuil**. Quand le score dépasse un certain seuil alors on considère que la valeur de l'instance est égale à 1.

La courbe ROC

La courbe ROC donne la sensibilité (taux de vrai positifs) en fonction de l'anti spécificité (taux de faux positif) pour toutes les valeurs du seuil (entre 0 et 1)





Lecture

Une courbe ROC se lit en évaluant l'écart entre la courbe générée par score et le classificateur aléatoire (quelle que soit le seuil utilisé, comme le modèle est aléatoire, on aura la même proportion de prédictions positives correctes que de prédictions positives incorrectes. Le taux de faux positifs et le taux de vrais positifs seront donc égaux, et ainsi la sensibilité sera égale au complément à 1 de la sensitivité.

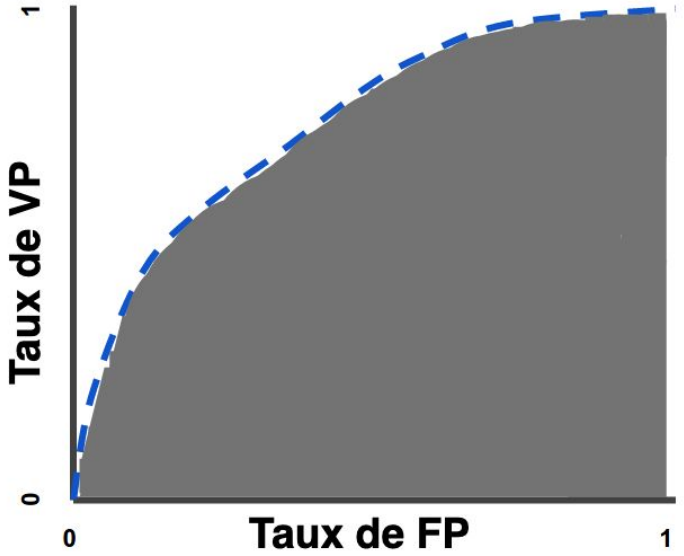
La régression logistique: évaluation d’un score

On peut résumer la courbe ROC par un nombre : "l'aire sous la courbe"(**AUROC**).

- Un classifieur parfait a une AUROC de 1 ;
- un classifieur aléatoire, une AUROC de 0.5.

Cela permet de comparer facilement des modèles.

AUC ou l’AUROC est une mesure de précision. C’est une mesure agrégée des performances pour tous les seuils de classification possibles



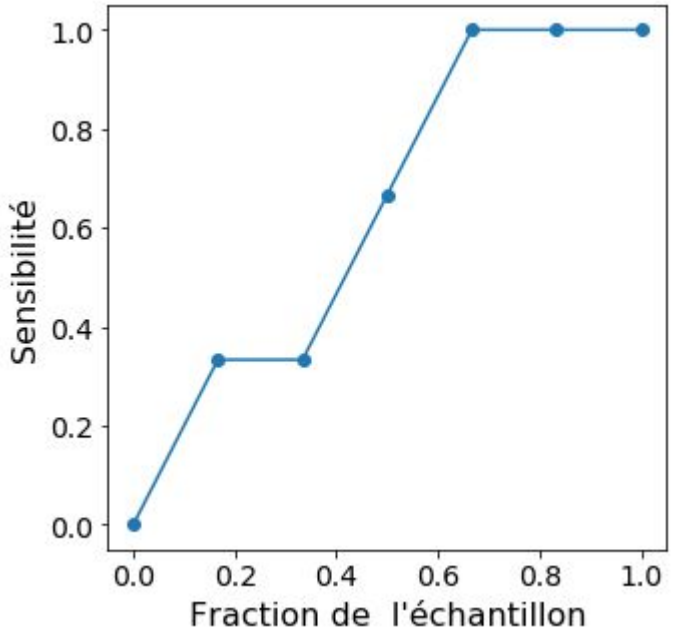
Caractéristiques de l’AUC:

- **L'AUC est invariante d'échelle.** Elle mesure la qualité du classement des prédictions, plutôt que leurs valeurs absolues.
- **L'AUC est indépendante des seuils de classification.** Elle mesure la qualité des précisions du modèle quel que soit le seuil de classification sélectionné.

La régression logistique: évaluation d'un score

La courbe Lift

Surtout utilisée dans le ciblage marketing, se construit aussi en parcourant le jeu de données ordonné par score. On représente en abscisse la fraction du jeu de données parcourue, et en ordonnée le taux de vrais positifs.



Lecture

On peut prendre l'exemple d'une campagne d'emailing. On considère que notre campagne est rentable seulement si 80% des personnes y répondent favorablement. On pourra donc fixer le seuil en fonction de ce critère. Ici on voit qu'au delà de 0,6 les personnes classées 1 ont au moins 80% de chance d'être réellement positive. On choisit donc comme seul 0,6.

La régression logistique: régularisation

Comme pour la régression linéaire multiple, il est possible de régulariser un modèle logistique.

On pourra donc utiliser :

- La régression logistique avec régularisation L2 (comme pour Ridge) pour éviter le sur-apprentissage (dans scikit-learn, c'est même l'implémentation par défaut de la régression logistique) ;
- La régression logistique avec régularisation L1 (comme pour Lasso) pour obtenir un modèle parcimonieux (dans scikit-learn, il suffit d'utiliser l'option'penalty'=l1).

Le machine learning en fiche

Objectif: constituer 7 fiches qui résument les principales étapes du machine learning.

Pour chaque thème décrire:

- Ce que c'est?
- Pour quelles raisons met-on en place cette procédure
- Quels sont les différents moyen de le réaliser à travers la librairie scikit learn
- Exemple à partir de la base de données Titanic

Rendu: un notebook de votre présentation qui sera le support de présentation au groupe et une fiche que tout le monde pourra conserver.

Ressources:
https://scikit-learn.org/stable/user_guide.html

Temps: une demi journée, présentation demain matin.

Le machine learning en fiche

Groupes et sujets:

- Rachid, Marie: Encodage, Onehotencoder et discrétisation
- Kevin, Kevin Tanguy: Feature selection
- Julien Antoine, Phichet : Gestion valeurs manquantes / imputer sciait learn
- Farid, Vivien: normalisation / standardisation/ Polynomial Features
- Camille, Arthur: Cross validation et validation curve.
- Joséphine Giovanny: Courbe ROC / Courbe Lift /Learning curve
- Hatice, Mickael: Transformer / Estimator / Pipeline et Pipeline gridsearch

Sources

Openclassroom: initiez vous au ML:

https://openclassrooms.com/fr/courses/4011851-initiez-vous-au-machine-learning/5869331-decouvrez-le-domaine-de-la-data-science

Wikipédia

https://www.wikiwand.com/fr/R%C3%A9gression_lin%C3%A9aire

Kaggle: Intro to Machine learning

https://www.kaggle.com/learn/intro-to-machine-learning