Data Science

Régression Linéaire

Hachem Kadri

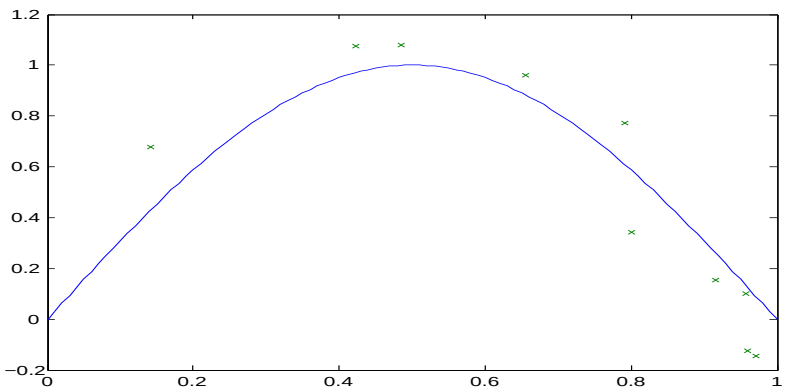
2019-2020

Problème de régression

On observe des données

$$(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$$

On cherche à exprimer la dépendance entre x et y par une fonction.



Un exemple : USCrime (L. Wasserman)

- These data are crime-related and demographic statistics for 47 US states in 1960.
- The data were collected from the FBI's Uniform Crime Report and other government agencies to determine how the variable crime rate depends on the other variables measured in the study.

Un exemple : USCrime (suite)

1. R: Crime rate: # of offenses reported to police per million population
2. Age: The number of males of age 14-24 per 1000 population
3. S: Indicator variable for Southern states (0 = No, 1 = Yes) Ed: Mean # of years of schooling x 10 for persons of age 25 or older
4. Ex0: 1960 per capita expenditure on police by state and local government
5. Ex1: 1959 per capita expenditure on police by state and local government
6. LF: Labor force participation rate per 1000 civilian urban males age 14-24
7. M: The number of males per 1000 females
8. N: State population size in hundred thousands
9. NW: The number of non-whites per 1000 population
10. U1: Unemployment rate of urban males per 1000 of age 14-24
11. U2: Unemployment rate of urban males per 1000 of age 35-39
12. W: Median value of transferable goods and assets or family income in tens of $
13. X: The number of families per 1000 earning below 1/2 the median income

Un exemple : USCrime (suite)

R	Age	S	Ed	Ex0	Ex1	LF	M	N	NW	U1	U2	W	X
79.1	151	1	91	58	56	510	950	33	301	108	41	394	261
163.5	143	0	113	103	95	583	1012	13	102	96	36	557	194
57.8	142	1	89	45	44	533	969	18	219	94	33	318	250
196.9	136	0	121	149	141	577	994	157	80	102	39	673	167
123.4	141	0	121	109	101	591	985	18	30	91	20	578	174
68.2	121	0	110	118	115	547	964	25	44	84	29	689	126
96.3	127	1	111	82	79	519	982	4	139	97	38	620	168
155.5	131	1	109	115	109	542	969	50	179	79	35	472	206
85.6	157	1	90	65	62	553	955	39	286	81	28	421	239
70.5	140	0	118	71	68	632	1029	7	15	100	24	526	174
167.4	124	0	105	121	116	580	966	101	106	77	35	657	170
84.9	134	0	108	75	71	595	972	47	59	83	31	580	172
51.1	128	0	113	67	60	624	972	28	10	77	25	507	206
66.4	135	0	117	62	61	595	986	22	46	77	27	529	190
79.8	152	1	87	57	53	530	986	30	72	92	43	405	264
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Expliquer la variable R utilisant des variables explicatives (les autres attributs)

Modélisation de la régression

- Une variable aléatoire $Z = (X, Y)$ à valeurs dans $\mathbb{R}^d \times \mathbb{R}$
- Les exemples sont des couples $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ tirés selon la distribution jointe
  $P(Z = (x, y)) = P(X = x)P(Y = y|X = x)$.
- Un échantillon S est un ensemble fini d'exemples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ i.i.d. selon P.

Modélisation de la régression (suite)

On cherche une fonction : $f : \mathbb{R}^d \to \mathbb{R}$.
Fonction de perte (loss function)

$$L(y, f(x)) = (y - f(x))^2.$$

La fonction risque (ou erreur) : espérance mathématique de la fonction de perte.

$$R(f) = \int L(y, f(x)) dP(x, y) = \int_{\mathbb{R}^d \times \mathbb{R}} (y - f(x))^2 dP(x, y).$$

Le problème général de la régression :

étant donné un échantillon $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, trouver un classifieur f qui minimise le risque $R(f)$.

La fonction de régression

- Il existe une fonction qui minimise l'écart quadratique moyen : la *fonction de régression*

$$r(x) = \int_Y y dP(y|x)$$

- Pour chaque x, $r(x)$ est égal à la moyenne des observations

- Comme la fonction de Bayes en classification, la fonction de régression est le plus souvent inaccessible.

Minimisation du risque empirique

- Le risque empirique $R_{emp}(f)$ de f est la moyenne des carrés des écarts à la moyenne de f calculée sur S :

$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2.$$

- Principe de minimisation du risque empirique : calculer

$$\underset{f}{ArgMin} R_{emp}(f)$$

Méthode des moindres carrés

Régression linéaire

On suppose que

$$Y = \langle \alpha, X \rangle + \beta + \epsilon$$

où

- X prend ses valeurs dans $\mathbb{R}^d$,
- $\alpha \in \mathbb{R}^d$ et $\beta \in \mathbb{R}$,
- ϵ est une variable aléatoire telle que $E(\epsilon) = 0$ et $V(\epsilon) = \sigma^2$ (variance indépendante de X).

La fonction de régression est

$$r(x) = \langle \alpha, x \rangle + \beta = \alpha_1 x_1 + \ldots \alpha_d x_d + \beta.$$

Régression linéaire - cas $d = 1$

On suppose que

- X prend des valeurs dans $\mathbb{R}$,
- $Y = \alpha X + \beta + \epsilon$ où $E(\epsilon) = 0$ et $V(\epsilon) = \sigma^2$ (variance indépendante de X).

La fonction de régression est

$$r(x) = \alpha x + \beta.$$

Régression linéaire : estimateurs des moindres carrés

Soit $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ un échantillon. Les valeurs de $\hat{\alpha}$ et $\hat{\beta}$ qui minimisent

$$\sum_{i=1}^{n} (y_i - (\hat{\alpha} x_i + \hat{\beta}))^2$$

sont

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \text{ et } \hat{\beta} = \overline{y} - \hat{\alpha}\overline{x}$$

où

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \text{ et } \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Régression linéaire : estimateurs des moindres carrés (suite)

La fonction de régression estimée est alors

$$\hat{r}(x) = \hat{\alpha}x + \hat{\beta}.$$

Les erreurs estimées sont

$$\widehat{\epsilon_i} = y_i - \widehat{y_i} = y_i - (\hat{\alpha}x_i + \hat{\beta}).$$

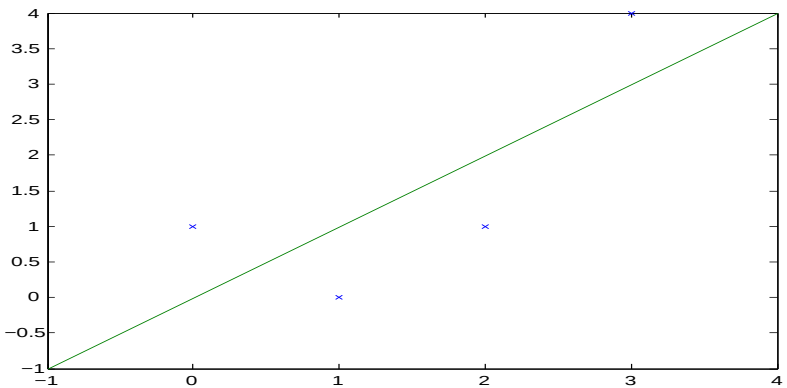
La variance estimée est

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \widehat{\epsilon_i}^2.$$

Estimateurs des moindres carrés : exemple

Soit $S = \{(0,1),(1,0),(2,1),(3,4))\}$ un échantillon.
On trouve

$$\overline{x} = 3/2, \overline{y} = 3/2, \hat{\alpha} = 1 \text{ et } \hat{\beta} = 0.$$



On a $\widehat{\epsilon_1} = 1, \widehat{\epsilon_2} = -1, \widehat{\epsilon_3} = -1, \widehat{\epsilon_4} = 1$ et $\hat{\sigma}^2 = 2$.

Propriétés de l'estimateur des moindres carrés

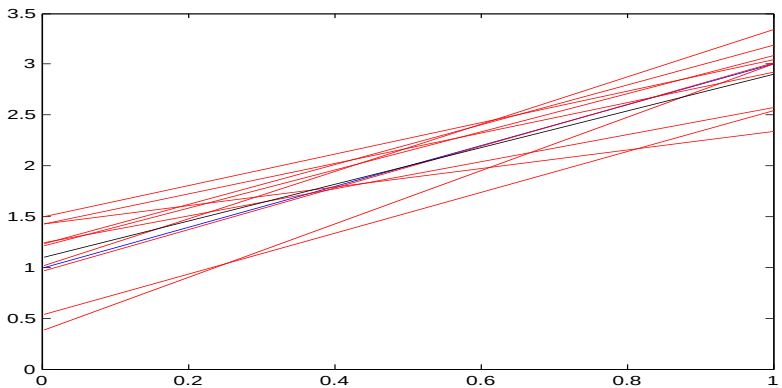
- $\hat{\alpha}, \hat{\beta}$ et $\hat{\sigma}^2$ sont des *estimateurs non biaisés* de α, β et σ^2 : si l'on répète un grand nombre d'expériences avec le même modèle, les moyennes des estimations convergent vers les paramètres du modèle.
- $\hat{\alpha}, \hat{\beta}$ et $\hat{\sigma}^2$ sont des *estimateurs consistants* de α, β et σ^2 : plus on dispose d'observations, plus les estimations se rapprochent des paramètres du modèle.
- si ϵ suit une loi normale, l'estimateur des moindres carrés est aussi l'*estimateur du maximum de vraisemblance* : celui qui maximise la probabilité des observations.

Estimateur non biaisé : illustration

X prend 11 valeurs équidistantes dans $[0, 1]$;
$Y = 2 * X + 1 + Norm(0, 1)$.

On réalise 10 expériences.

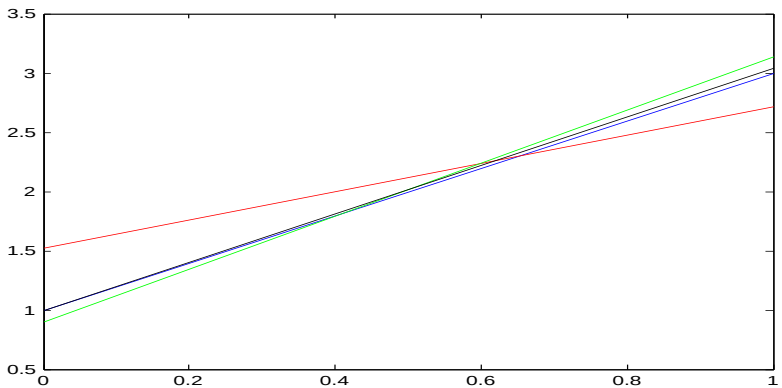
- en bleu : la droite de régression
- en rouge : chaque estimation
- en noir : la moyenne des estimations.



Estimateur consistant : illustration

X prend N valeurs équidistantes dans $[0, 1]$;
$Y = 2 * X + 1 + Norm(0, 1)$.

- en bleu : la droite de régression
- en rouge : $N = 11$
- en vert : $N = 101$
- en noir : $N = 1001$.



Régression linéaire multiple

Soit $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset$ l'échantillon d'apprentissage.

Soit X la matrice $n \times d$ dont la i-ème ligne est x_i.
Soit Y le vecteur colonne de taille n composé des étiquettes y_i.

n est le nombre d'exemples et d est le nombre d'attributs.

- avec des notations matricielles :

$$\sum_{i=1}^{n}(y_i - \alpha^\top x_i)^2 = \|Y - X\alpha\|^2$$

- Gradient par rapport à α :

$$\nabla_\alpha(\|Y - X\alpha\|^2) = -2X^\top(Y - X\alpha)$$

Régression linéaire multiple

La solution du problème de régression par moindres carrés est obtenue en calculant le vecteur de parametre α qui annule la dérivée.

L'estimateur des moindres carrés est

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

où X^T désigne la matrice transposée de X.

Si $X^T X$ n'est pas inversible, ou si $det(X^T X) \simeq 0$, ... il est nécessaire de transformer le problème.

Régression linéaire multiple

Pour considérer le terme biais, c-a-d le modèle $y = \alpha^\top x + \beta + \epsilon$, on augmente la dimension de l'espace d'entrée et on ajoute une composante égale à 1 ($\tilde{x} = (x, 1)$).

Soit $\tilde{X}$ la matrice $n \times (d + 1)$ dont la i-ème ligne est $x_i, 1$.

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (X^T X)^{-1} X^T Y$$

Exemple :

$S = \{((0,0),-1),((0,1),1),((1,0),1),((1,1),1)\}$.
On a

$$X = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}, X^T = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \text{ et } Y = \begin{pmatrix} -1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

On vérifie que

$$X^T X = \begin{pmatrix} 2 & 1 & 2 \\ 1 & 2 & 2 \\ 2 & 2 & 4 \end{pmatrix}, (X^T X)^{-1} = \begin{pmatrix} 1 & 0 & -1/2 \\ 0 & 1 & -1/2 \\ -1/2 & -1/2 & 3/4 \end{pmatrix}$$

$$(X^T X)^{-1} X^T = \begin{pmatrix} -1/2 & -1/2 & 1/2 & 1/2 \\ -1/2 & 1/2 & -1/2 & 1/2 \\ 3/4 & 1/4 & 1/4 & -1/4 \end{pmatrix} \text{et } (X^T X)^{-1} X^T Y = \begin{pmatrix} 1 \\ 1 \\ -1/2 \end{pmatrix}$$

soit

$$\hat{\alpha} = (1,1) \text{ et } \hat{\beta} = -1/2.$$

Modèles linéaires généralisés

- Peu vraisemblable, en général, que les données d'observations $(x_i, y_i) \in \mathbb{R}^{d+1}$ puissent être précisément décrites par un modèle linéaire.
- Sous des conditions de régularités très générales, toute fonction f peut être approchée, au moins localement, par une combinaison **linéaire**
  - de *monômes* (développements limités)

$$\sin x = x - x^3/3 + x^5/5 - \dots$$

  - ou de *fonctions trigonométriques* (développements de Fourier)

$$x = 2\left(\sin x - \frac{\sin 2x}{2} + \frac{\sin 3x}{3} - \dots\right)$$

Modèles linéaires généralisés (suite)

- Soient $h_1, \ldots, h_M : \mathbb{R}^d \mapsto \mathbb{R}$.
- On transforme chaque x_i en un vecteur $(h_1(x_i), \ldots, h_M(x_i))$.
- On considère le problème de régression linéaire multivarié suivant : trouver les coefficients $\alpha_1, \ldots, \alpha_M, \beta \in \mathbb{R}$ qui minimisent

$$\sum_{i=1}^{n} \left(y_i - \left(\sum_{j=1}^{M} \alpha_j h_j(x_i) + \beta \right) \right)^2 .$$

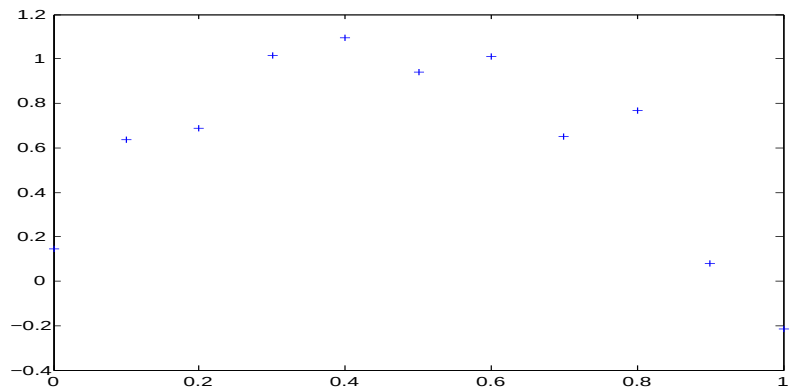
- On obtient ainsi une fonction de régression non linéaire

$$\hat{f}(x) = \sum_{j=1}^{M} \hat{\alpha}_j h_j(x) + \hat{\beta}.$$

Un exemple

On observe les données suivantes :

0	0.1000	0.2000	0.3000	0.4000	0.5000	0.6000	0.7000	0.8000	0.9000	1.0000
0.1434	0.6351	0.6856	1.0160	1.0964	0.9393	1.0098	0.6516	0.7655	0.0796	-0.2138



Les données ne semblent pas alignées : on les transforme par les fonctions

$$h_1(x) = x, h_2(x) = x^2 \text{ et } h_3(x) = x^3.$$

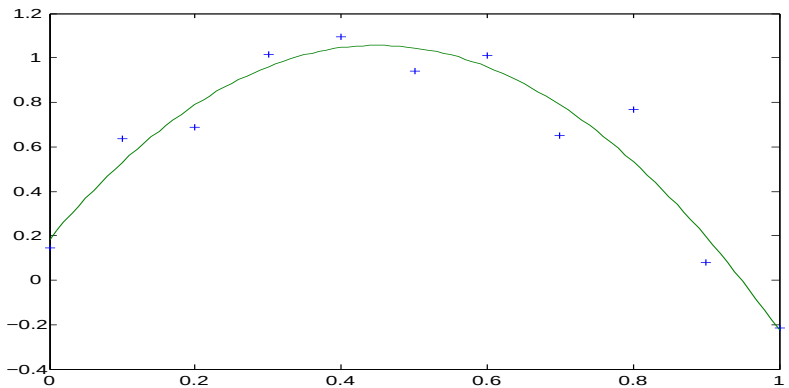
Un exemple (suite)

Une régression linéaire multivariée permet de trouver

$$\beta = 0.1848, \alpha_1 = 3.8960, \alpha_2 = -4.3942 \text{ et } \alpha_3 = 0.0878$$

soit le polynôme

$$p(x) = 0.1848 + 3.8960x - 4.3942x^2 + 0.0878x^3.$$



Sélection de variables

Les observations x peuvent dépendre d'un très grand nombre de variables.

- Constat : trop de variables nuisent à la qualité de la prédiction.
- Le modèle induit peut être difficile à interpréter s'il fait intervenir toutes les variables au même niveau.
- Question : comment trouver un bon compromis entre biais et variance ? *underfitting* et *overfitting* ?

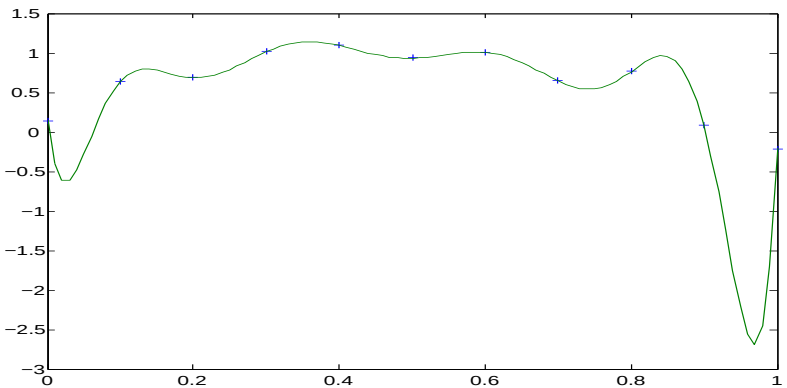
En trouvant un compromis entre l'adéquation aux données et la complexité du modèle.

Un exemple (suite)

Si l'on prend comme base les monômes $h_i(x) = x^i$ pour $1 \leq i \leq 10$, on trouve le polynôme

$$p(x) \simeq 0.1463 - 76.0471x + 2394.5864x^2 - 28139.8810x^3 + 173816.7283x^4 - 634281.2333$$

$$1437424.4162x^6 - 2044405.0970x^7 + 1774215.9551x^8 - 858102.0064x^9 + 177152.2278x^{10}.$$



Quels sont les monômes les plus significatifs ?

Sélection de variables

Soit $\mathcal{V}$ l'ensemble de variables et soit S un sous-ensemble de $\mathcal{V}$.

Soit $\hat{r}_S$ la fonction de régression calculée sur S.

On note l_S la log-vraisemblance du modèle $\hat{r}_S$.

Deux critères pour sélectionner S :

- Maximiser $l_S - |S|$ (méthode AIC, pour Akaike Information Criterion)
- Maximiser $l_S - \frac{|S|}{2} \log d$ (méthode BIC, pour Bayesian Information Criterion)

Dans les deux cas : compromis entre l'adéquation aux données et la complexité du modèle.

Deux stratégies pour sélectionner S

Pour sélectionner S, on peut envisager tous les ensembles S possibles et évaluer S selon l'un des deux critères précédents.

Problème : si le nombre de variables est trop important, c'est impossible !

Deux stratégies alternatives gloutonnes : forward and backward stepwise regression.

- Partir du modèle vide et ajouter des variables, l'une après l'autre, tant que le score associé au critère choisi croît.
- Partir de l'ensemble des variables $\mathcal{V}$ et les supprimer l'une après l'autre, tant que le score associé au critère choisi croît.

Régression Ridge

Autre idée : pénaliser la taille du modèle. On cherche à minimiser

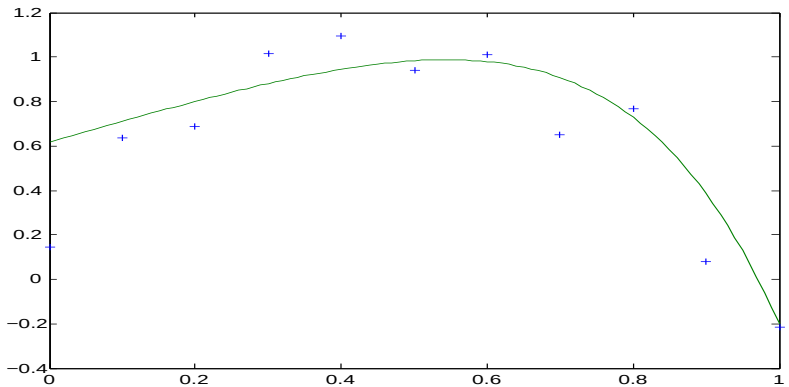
$$\sum_{i=1}^{n}(y_i - (\alpha x_i + \beta))^2 + C(\beta^2 + ||\alpha||^2).$$

- se résoud aussi simplement que la régression multivariée sans pénalisation.
- mais comment trouver la valeur de C ? Par exemple, par validation croisée.

Un exemple (suite)

Avec les monômes $h_i(x) = x^i$ pour $1 \leq i \leq 10$ et $C = 0.1$, on trouve les coefficients

$$0.6189\ ;\ 0.9368\ ;\ -0.0655\ ;\ -0.3620\ ;\ -0.4026$$
$$-0.3529\ ;\ -0.2745\ ;\ -0.1906\ ;\ -0.1095\ ;\ -0.0341\ ;\ 0.0350.$$



Régression Lasso

Autre idée : on cherche à minimiser

$$\sum_{i=1}^{I}(y_i - (\hat{\alpha}x_i + \hat{\beta}))^2 + C(|\beta| + \sum |\alpha_i|).$$

- Un grand nombre de coefficients s'annule.
- Il faut toujours déterminer une bonne valeur pour C.

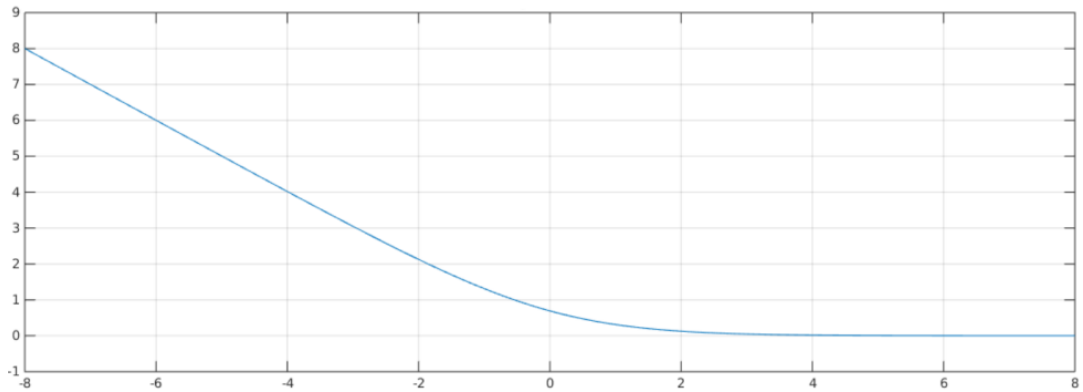
Régression Logistique

- Régression régularisée

$$\sum_{i=1}^{I} \ell(y_i, f_\alpha(x_i)) + C||\alpha||^2$$

- Fonction coût logistique

$$\ell(y_i, f_\alpha(x_i)) = \log(1 + e^{-yf_\alpha(x)})$$



Régression Logistique

- La fonction logistique est différentiable
- Pour calculer la solution, on peut utiliser l'algorithme de descente de gradient
- $\nabla_\alpha(\sum_{i=1}^{l} \ell(y_i, f_\alpha(x_i))) = \sum_{i=1}^{l} x_i \frac{-y_i}{1 + e^{y_i x_i^\top \alpha}}$
- $\alpha_t = \alpha_{t-1} - \gamma(\sum_{i=1}^{l} x_i \frac{-y_i}{1 + e^{y_i x_i^\top \alpha_{t-1}}} + 2C\alpha_{t-1})$
- La solution de la régression logistique a une interprétation probabilistique