Chap. 2
Linear regression

I) Model specification

$x = (x_1, x_2, \dots, x_D)$ → one input

output $y = \underbrace{\sum_{i=1}^{D} w_i x_i}_{= w^T x} + \varepsilon$ ↳ Gaussian noise, variance σ^2

If we want to model a non-linear relationship, we can replace x by $\phi(x)$ (this is called basis function expansion)

(x is 1-dim and $\phi(x)$ can be, ex., $\phi(x) = [1, x, x^2, \dots, x^{D-1}]$)

the model has a parameter $\theta = \begin{pmatrix} \sigma^2 \\ w \end{pmatrix}$ (in general, the parameter in a model is denoted by θ).

II) Maximum likelihood estimation (least squares)

A common way to estimate the parameter is the MLE, defined as:

$$\hat{\theta} := \arg\max_{\theta} \underbrace{\log p(\mathcal{D}|\theta)}_{= \ell(\theta)} \qquad (\mathcal{D} = \text{data})$$

More precisely: if the $(x^{(i)}, y^{(i)})$ we have access to are iid, we write:

$$\ell(\theta) = \sum_{i=1}^{N} \log p(y^{(i)} | x^{(i)}, \theta)$$

↙

where: $y \mapsto p(y|x,\theta)$ is the density of y knowing x [supposing we know the vector w] and [supposing ε is $\mathcal{N}(0,\sigma^2)$], that is

$$p(y|x,\theta) = \frac{\exp\left(-\frac{(w^T x - y)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}}$$

Instead of maximizing $l(\theta)$ (= log-likelihood), we could minimize the negative-log-likelihood (NLL):

$$NLL(\theta) := -\sum_{i=1}^{N} \log p(y^{(i)} | x^{(i)}, \theta)$$

(because some software packages are designed to find minima and not maxima)

We have: $l(\theta) = \sum_{i=1}^{N} \log \left[\left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp\left(-\frac{1}{\sigma^2} (y^{(i)} - w^T x^{(i)})^2 \right) \right]$

$$= -\frac{1}{2\sigma^2} RSS(w) - \frac{N}{2} \log(2\pi\sigma^2)$$

where RSS is the residual sum of squares

$$RSS(w) := \sum_{i=1}^{N} (y_i - w^T x)^2 .$$

RSS is also called the sum of squared errors (or SSE) and $\frac{SSE}{N}$ is called the mean square error (or MSE)

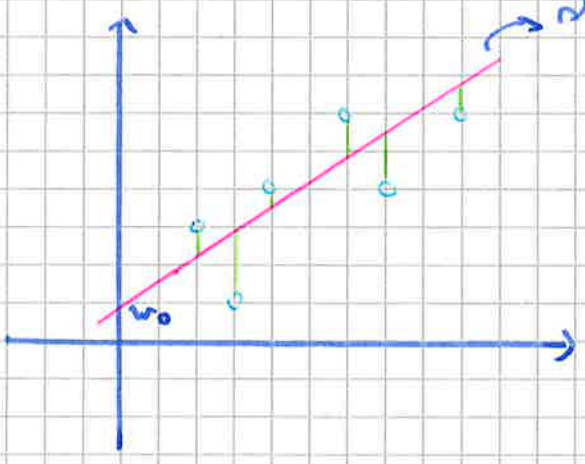
It can also be written as the square of the l^2 norm of the vector of residual errors:

$$RSS(w) = \sum_{i=1}^{N} (\varepsilon^{(i)})^2 = \|\varepsilon\|_2^2$$

↳ where $\varepsilon^{(i)} := y^{(i)} - w^T x^{(i)}$

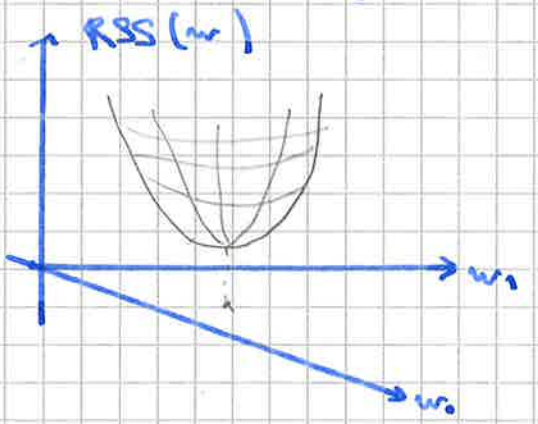
Suppose σ is known and we only look at the vector w minimizing the RSS → this method is known as least squares.

Ex: $D = 2$, $w = (w_0, w_1)$



$x^{(i)}$ = blue points

We look for the pink line minimizing the sum of the squares of the green distances.



1) Derivation of the MLE

$$NLL(w) = \frac{1}{2}(y - Xw)^T(y - Xw) = \frac{1}{2} w^T (X^T X) w - w^T (X^T y) + \frac{1}{2} y^T y$$

$$\begin{cases} y = (y^{(1)}, \dots, y^{(N)}) \quad \text{(column vector)} \\ w = (w_1, w_2, \dots, w_D) \quad \text{(column vector)} \\ X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_D^{(1)} \\ \vdots & & & \vdots \\ x_1^{(N)} & x_2^{(N)} & \dots & x_D^{(N)} \end{bmatrix} \end{cases}$$

where

$$\begin{cases} X^T X = \sum_{i=1}^{N} \begin{bmatrix} (x_1^{(i)})^2 & \dots & x_1^{(i)} x_D^{(i)} \\ \vdots & & \vdots \\ x_1^{(i)} x_D^{(i)} & \dots & (x_D^{(i)})^2 \end{bmatrix} \\ X^T y = \sum_{i=1}^{N} x^{(i)} \times y_i \end{cases}$$

We compute the gradient :

$$\nabla(NLL(w)) = (X^T X) w - X^T y$$

Indeed if A is a symmetric square matrix

$$A = \begin{bmatrix} a_{11} & \dots & a_{1D} \\ \vdots & & \vdots \\ a_{D1} & \dots & a_{DD} \end{bmatrix}$$

$$w^T A w = \sum_{i=1}^{D} \sum_{j=1}^{D} w_i a_{ij} w_j$$

whose gradient is a vector with line k :

$$\frac{\partial}{\partial w_k} \left(\sum_{i=1}^{D} \sum_{j=1}^{D} w_i a_{ij} w_j \right) = \sum_{j=1}^{D} a_{kj} w_j + \sum_{i=1}^{D} w_i a_{ik}$$

$$= 2 \sum_{j=1}^{D} a_{kj} w_j$$

i.e. the vector $2Aw$

and the gradient of $w^T \eta$ (for η a vector) is a vector with line k:

$$\frac{\partial}{\partial w_k}\left(\sum_{i=1}^{D} w_i \eta_i\right) = \eta_k$$

i.e. the vector η.

the critical point (the w for which $\nabla(NLL(w)) = 0$) is:

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

→ under the assumption that this matrix is invertible ($X^T X$) (explanation later)

2) Geometric interpretation

We assume $N > D$. We set $E = \text{Vect}\left(\begin{pmatrix} x_i^{(1)} \\ \vdots \\ x_i^{(N)} \end{pmatrix}, 1 \le i \le D\right)$.

E is a sub-space of dimension $\le D$ of $\mathbb{R}^N$. Let us suppose $\dim(E) = D$. We set (for $1 \le i \le D$):

$$\vec{x}_i = \begin{pmatrix} x_i^{(1)} \\ \vdots \\ x_i^{(N)} \end{pmatrix}$$

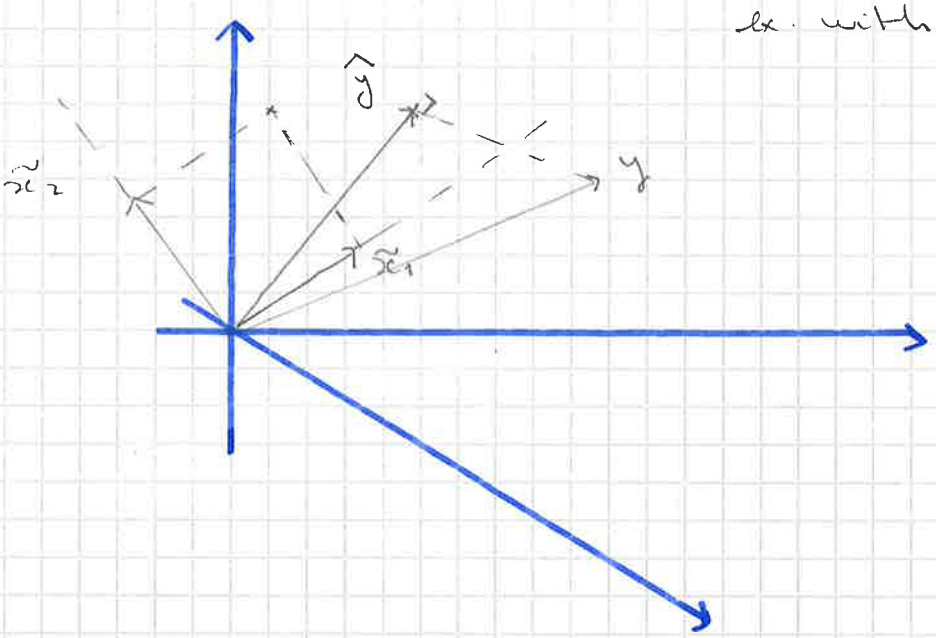
What we do in the linear regression is that we seek a vector $\hat{y} \in \mathbb{R}^N$ such that $\hat{y} \in E$ and is as close as possible to y: $\hat{y} = \underset{\eta \in \text{span}(\vec{x}_1, \dots, \vec{x}_D)}{\text{argmin}} \|y - \eta\|_2$

Or to put it differently, we write $\eta = w_1 \vec{x}_1 + \dots + w_D \vec{x}_D$ ($\in \text{span}(\vec{x}_1, \dots, \vec{x}_D)$)

i.e. $\eta = Xw$.

To minimize the norm of $y - \eta$, we want $y - \eta$ to be orthogonal to $\text{span}(\vec{x}_1, \dots, \vec{x}_D) = E$

ex. with $\begin{cases} N=3 \\ D=2 \end{cases}$



So $\tilde{x}_j^T (y - \hat{y})$ for $j = 1, 2, \dots, D$. Hence:

$$\tilde{x}_j^T (y - Xw) \text{ for } j = 1, 2, \dots, D$$

$$X^T (y - Xw) = 0$$

And so $w = (X^T X)^{-1} X^T y$

$$\hat{y} = Xw = X (X^T X)^{-1} X^T y .$$

this is the <u>orthogonal projection</u> of y onto $\text{span}(\tilde{x}_1, \dots, \tilde{x}_D)$.

the projection matrix $P := X(X^T X)^{-1} X^T$ is called the <u>hat matrix</u> since it "puts a hat on y".

<u>Why is $X^T X$ invertible in the case $N > D$?</u>

For any $y \notin \text{Span}(\tilde{x}_1, \dots, \tilde{x}_D)$, $\exists!$ w such that:

$X^T (y - Xw) = 0$ (also true for $y \in \text{Span}(\tilde{x}_1, \dots, \tilde{x}_D)$)

i.e. $X^T X w = X^T y$

$X^T y$: a generic element of $(x_1, \dots, x_N)$ (remember $x_i = \begin{pmatrix} x_1^{(i)} \\ \vdots \\ x_D^{(i)} \end{pmatrix}$)

$w \in \mathbb{R}^D$

this means that $X^T X$ is invertible.

III) Ridge regression

the MLE can overfit because it picks the parameters that are good for the training data. But the training data can be noisy so there is no reason to pick the best parameters for the training data.

the alternative to linear regression is to use a Bayesian statistics approach. We suppose $y_i = w_0 + w^T x + \varepsilon_i$

We assume an a priori distribution of the $w_1, \dots, w_D$ (corresponding to a prior knowledge / an preconceived idea (...)) Here, we have in mind that they should not be too big so we chose the following a priori:

$$p(w) = \prod_{j=1}^{D} \mathcal{N}(w_j | 0, \tau^2)$$

↳ meaning each w_j is ind. of the others and of law $\mathcal{N}(0, \tau^2)$

the likelihood is the density of $y_1, \dots, y_D$ under the assumption that $w_1, \dots, w_D$ is what it is:

$$p(y | x_i, \underbrace{\theta}_{= w_1, \dots, w_D}) = \prod_{i=1}^{N} \mathcal{N}(y_i | w_0 + w \cdot x_i, \sigma^2)$$

↳ knowing w_i, x_i, y_i has a Gaussian density

the observations $y_1, \dots, y_D$ are fixed. the higher $p(y | x_i, \theta)$ is, the more likely it is that $w_1, \dots, w_D$ has a certain value.

the a posteriori distribution is the product of the a priori distribution and the likelihood, renormalized to be a probability density (in w).

We want to find the MAP (Maximum A Posteriori),

so we look for w maximizing

$$\sum_{i=1}^{N} \log \mathcal{N}(y_i \mid \underbrace{w_0 + w^T x_i}_{w_0 + w^T x_i}, \sigma^2) + \sum_{j=1}^{D} \log \mathcal{N}(w_j \mid 0, \tau^2)$$

(we took the log and threw away the normalization constant)

or, equivalently, minimizing:

$$\frac{1}{N}\left[\sum_{i=1}^{N} \frac{(y_i - \overbrace{w_0}^{\bar{y}} - w^T x_i)^2}{2\sigma^2} + \sum_{j=1}^{D} \frac{w_j^2}{2\tau^2}\right]$$

or (equ.) minimizing $\frac{1}{N}\left[\sum_{i=1}^{N} (y_i - \underbrace{w_0}_{\bar{y}} - w^T x_i)^2 + \lambda w^T w\right]$

$$\left(\lambda = \frac{\sigma^2}{\tau^2}\right)$$

We can interpret the first term as the NLL of before and the second term as a complexity penalty (we do not want the w_j to be too big). The corresponding solution is given by:

$$\hat{w}_{ridge} = (\lambda I_D + X^T X)^{-1} X^T y$$

this technique is known as ridge regression or penalized least squares. In general, adding a Gaussian prior to the parameters of a model to encourage them to be small is called l_2-regularization or weight-decay.

Projeter la figure 7.8 p. 226 du livre.

IV) The LASSO

We define $\hat{w}_{lasso} = \text{argmin}\left\{\sum_{i=1}^{N} (y_i - w_0 - w^T x_i)^2, w \text{ such that } \sum_{j=1}^{D} |w_j| \leq t\right\}$

(t to be chosen)

- Making t sufficiently small will cause some of the coefficients to be exactly 0.

- If t is chosen larger than $t_0 = \sum_{i=1}^{D} |\hat{w}_i^{LS}|$ (the ones we found in the least-square regression)

  Then $\hat{w}_{lasso} = \hat{w}^{LS}$

- t should be adaptively chosen to minimize an estimate of expected prediction error

Def: LASSO and Ridge regressions are part of what is called shrinkage methods

---

Remarques pour le TP : p. 45 à 55 du O'Reilly

* Le α de la Ridge Regression est le λ du cours
* Le λ du LASSO est le t du cours.
* L'algorithme de python trouve la solution du LASSO par une boucle (on doit converger vers la solution). On peut préciser le nbre d'itérations maximales (max_iter) autorisées (en particulier, on peut relever ce plafond pour s'assurer de la convergence).