Linear regression

R. Flamary, A. Rakotomamonjy

January 10, 2019

Sommaire

Linear Prediction

Linear function

Function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, can be expressed as

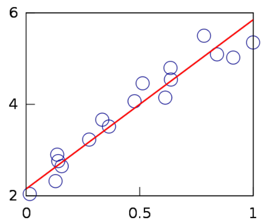
$$f(\mathbf{x}) = \sum_{i=1}^{d} w_i x_i + b = \mathbf{x}^\top \mathbf{w} + b = [\mathbf{x}^\top 1]\boldsymbol{\alpha} \tag{1}$$

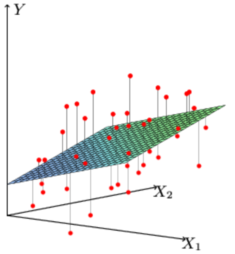
with $\mathbf{w} \in \mathbb{R}^d$ a vector defining an hyperplane in $\mathbb{R}^d$ et $b \in \mathbb{R}$ a bias term dislacing the function along the normal $\mathbf{w}$ of the hyperplane. All parameters can be stored in a unique vector $\boldsymbol{\alpha} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$ of dimensionality $\mathbb{R}^{d+1}$ concatenating $\mathbf{w}$ and b.

Objective of linear prediction

- Regression: $f(\cdot) \in \mathbb{R}$.
- Classification: $\text{sign}(f(\cdot)) \in \{-1, 1\}$.

Linear regression

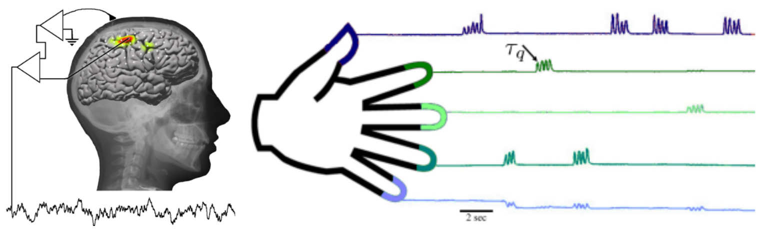




Objective

Train a linear function $f(\cdot)$ that can predict a continuous value $y \in \mathbb{R}$ from an observation $\mathbf{x} \in \mathbb{R}^d$.
In practice we want to find the coefficients $(\mathbf{w}, b)$ of $f(\cdot)$ using a training dataset $\{\mathbf{x}_i, y_i\}_{i=1,\dots,n}$.

Application for a Brain Computer Interface (BCI)



BCI Competition IV, Dataset 4

- Data: Recordings of ECoG brain signals and of simultaneous finger flexion of a subject (using a glove).
- Objective of the competition: predict movement of the 5 fingers of the subject from its recorded ECoG.
- Best performances wree obtained using a linear model.

How do we store training data ?

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top & 1 \\ \mathbf{x}_2^\top & 1 \\ \vdots & \vdots \\ \mathbf{x}_i^\top & 1 \\ \vdots & \vdots \\ \mathbf{x}_n^\top & 1 \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1d} & 1 \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2d} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{id} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{nd} & 1 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}$$

Training data

- $\mathbf{x}_i \in \mathbb{R}^d$ observations for $i = 1, \dots, n$.
- $y_i \in \mathbb{R}$ values to predict for $i = 1, \dots, n$.

Matrix form:

- $\mathbf{X} \in \mathbb{R}^{n \times (d+1)}$ such that $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \mathbf{e}]^\top$ with $\mathbf{e} \in \mathbb{R}^d$ and $e_i = 1, \forall i$
- $\mathbf{y} \in \mathbb{R}^n$ such that $\mathbf{y} = [y_1, y_2, \dots, y_n]^\top$.
- $\boldsymbol{\alpha} \in \mathbb{R}^{d+1}$ is a vector such that $\boldsymbol{\alpha} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$

Performance measure

How to measure performance of a prediction?

Let $\mathbf{y}$ be the values to predict and $\hat{\mathbf{y}}$ the predictions.

Mean square error

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- 0 for a perfect prediction.
- Not normalized (depends on the variance of $\mathbf{y}$)

Correlation coefficient

$$r = \frac{\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sigma_y \sigma_{\hat{y}}}$$

- $\bar{y} = \frac{1}{n} \sum_i y_i$ mean of $\mathbf{y}$ and σ_y its.
- 1 for perfect prediction.
- Normalized : $r \in (-1, 1)$.

Warning

Always measure performance of a model on data that has NOT been used for training or else there is a risk of **over-fitting**

Error and squared error

Principle

We have the following model:

$$y = \mathbf{x}^\top \mathbf{w} + b = \tilde{\mathbf{x}}^\top \boldsymbol{\alpha} \tag{2}$$

where $\tilde{\mathbf{x}} = [x_1, ..., x_d, 1]^\top$ is $\mathbf{x}$ concatenated with 1.
We seek for parameters $(\mathbf{w}, b) \equiv \boldsymbol{\alpha}$ of function $f(\cdot)$ that works well on training data.

Residuals

The residual of sample i is the prediction error :

$$\epsilon_i = y_i - \mathbf{x}_i^\top \mathbf{w} - b = y_i - \tilde{\mathbf{x}}_i^\top \boldsymbol{\alpha} \tag{3}$$

We want the residuals to be the smallest possible in average. To this end we can measure the error as the square error:

$$\epsilon_i^2 = (y_i - \mathbf{x}_i^\top \mathbf{w} - b)^2 = (y_i - \tilde{\mathbf{x}}_i^\top \boldsymbol{\alpha})^2 \tag{4}$$

Least square optimization problem

We see the function $f(\cdot)$ minimizing the squared error on the training samples :

$$\min_f \quad \frac{1}{2}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2 = \frac{1}{2}\sum_{i=1}^{n}\epsilon_i^2 \tag{5}$$

using the linear form of $f(\cdot)$, we obtain the following optimization problem :

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \mathbf{w} - b)^2 \tag{6}$$

that is equivalent to

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{i=1}^{n}(y_i - \tilde{\mathbf{x}}_i^\top \boldsymbol{\alpha})^2 \tag{7}$$

Interpretation of least squares

$$J(\boldsymbol{\alpha}) = \frac{1}{2}\sum_{i=1}^{n}\big(\underbrace{y_i - \tilde{\mathbf{x}}_i^\top \boldsymbol{\alpha}}_{\varepsilon_i}\big)^2$$

The problem can be seen as finding an hyperplane $\mathbf{x}^T\mathbf{w} + b = y$ in a $\mathbb{R}^{d+1}$ space that best fits a point cloud $(\mathbf{x}_i, y_i)$ $i = 1, n$ with respect to the y dimension.

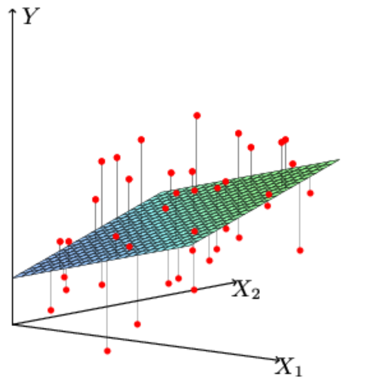


Figure: Residuals for the regression

Matrix form of least squares (1)

Resudial as a vector

The residuals (error) on the samples can be expressed as:

$$\epsilon_i = y_i - \mathbf{x}_i^\top \mathbf{w} - b = y_i - \tilde{\mathbf{x}}_i^\top \boldsymbol{\alpha} \tag{8}$$

Similarly to the training data, they can be stored in a vector $\boldsymbol{\epsilon} \in \mathbb{R}^n$ such that:

$$\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\alpha} \tag{9}$$

Matrix form for lerast square

The leats square optimization proble cen be expressed as:

$$\min_{\alpha} \quad \|\boldsymbol{\epsilon}\|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|^2 \tag{10}$$

where $\|\cdot\|$ is the euclidean norm of a vector such that $\|\boldsymbol{\epsilon}\|^2 = \sum_{i=1}^{n}\epsilon_i^2$

Matrix form of least squares (2)

The optimization problem can be expressed as:

$$\min_{\alpha} \quad J(\boldsymbol{\alpha}) \qquad \text{avec} \quad J(\boldsymbol{\alpha}) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}||^2$$

Using the properties of scalar product we get :

$$\begin{aligned}
\min_{\alpha} \quad J(\boldsymbol{\alpha}) &= \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}||^2 \\
&= \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}) \\
&= \frac{1}{2}\mathbf{y}^\top\mathbf{y} - \frac{1}{2}\boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{y} - \frac{1}{2}\mathbf{y}^\top\mathbf{X}\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha} \\
&= \frac{1}{2}\mathbf{y}^\top\mathbf{y} - \boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{y} + \frac{1}{2}\boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha}
\end{aligned}$$

$\mathbf{y}^\top\mathbf{y}$ is a scalar, $\mathbf{X}^\top\mathbf{y}$ is a vector $\mathbb{R}^d$ and $\mathbf{X}^\top\mathbf{X}$ is a $(d+1)\times(d+1)$ matrix.

Convex optimization basics

Optimization problem

We want to solve

$$\min_{\alpha} \quad J(\boldsymbol{\alpha}) \qquad \text{avec} \quad J(\boldsymbol{\alpha}) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}||^2$$

where $J(\boldsymbol{\alpha})$ is a convex function.

Minimum of a convex function

Let $J(\boldsymbol{\alpha})$ be a convex function $\mathbb{R}^d \to \mathbb{R}$ $\mathbb{R}$. $\alpha^\star$ is a minimum $J(\boldsymbol{\alpha})$ if and only if

$$\nabla J(\boldsymbol{\alpha}^\star) = \mathbf{0} \tag{11}$$

where $\nabla J(\boldsymbol{\alpha}) \in \mathbb{R}^d$ is the gradient of the function in $\boldsymbol{\alpha}$ such that

$$\nabla J(\boldsymbol{\alpha})_i = \frac{\partial J(\boldsymbol{\alpha})}{\partial \alpha_i} \quad \forall i$$

In order to find the minimum we need to find $\boldsymbol{\alpha}^\star$ such that the gradient is $\mathbf{0}$.

Gradient computation

$$\begin{array}{rcccccl} J(\boldsymbol{\alpha}) & = & \frac{1}{2}\mathbf{y}^\top\mathbf{y} & - & \boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{y} & + & \frac{1}{2}\boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha} \\ \dfrac{\partial J(\boldsymbol{\alpha})}{\partial \alpha_i} & = & 0 & - & p_i & + & \frac{1}{2}\sum_{j=1}^{d+1}(M_{ij} + M_{ji})\alpha_j \end{array}$$

with $\mathbf{p} = \mathbf{X}^\top\mathbf{y}$ and $\mathbf{M} = \mathbf{X}^\top\mathbf{X}$

- $\dfrac{\partial \boldsymbol{\alpha}^\top\mathbf{p}}{\partial \alpha_i} = \dfrac{\partial \sum_{j=1}^{d+1} p_j\alpha_j}{\partial \alpha_i} = p_i$
- $\dfrac{\partial \boldsymbol{\alpha}^\top\mathbf{M}\boldsymbol{\alpha}}{\partial \alpha_i} = \dfrac{\partial \sum_{j=1}^{d+1}\sum_{k=1}^{d+1} \alpha_j\alpha_k M_{jk}}{\partial \alpha_i} = \sum_{j=1}^{d+1} \alpha_j M_{ji} + \sum_{k=1}^{d+1} \alpha_k M_{ik}$

because $(uv)' = uv' + u'v$ with $u = \alpha_j$ et $v = \sum_{k=1}^{d+1} \alpha_k M_{jk}$

$$\nabla J(\boldsymbol{\alpha}) = -\mathbf{p} + \mathbf{M}\boldsymbol{\alpha} = -\mathbf{X}^\top\mathbf{y} \; + \; \mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha}$$

Least Squares solution

Minimizing the cost $J(\boldsymbol{\alpha})$ corresponds to finding the parameter $\boldsymbol{\alpha}$ that lads to a null gradient:

$$\nabla J(\widehat{\alpha}) = 0 \quad \Leftrightarrow \quad -\mathbf{X}^\top\mathbf{y} \; + \; \mathbf{X}^\top\mathbf{X}\widehat{\alpha} = \mathbf{0}$$

The solution of the minimization problem for Least Square is the vector $\widehat{\boldsymbol{\alpha}}$ defined as

$$\widehat{\alpha} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$$

Hypothesis

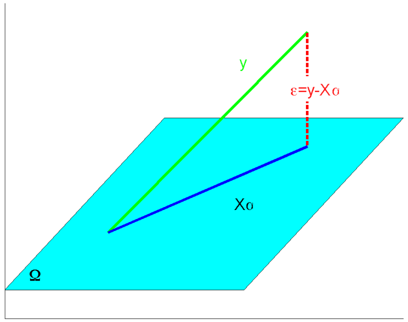
$\mathbf{X}$ is a matrix of rank $d+1$ which means that $\mathbf{X}^\top\mathbf{X}$ is invertible.
In practice it means that $n > d+1$, this method requires that we have more training samples than parameter to estimate.

Geometrical interpretation

$\Omega = \textbf{span}\{\mathbf{X}\}$

Ω is the linear subspace of $\mathbb{R}^n$ generated by the columns of matrix $\mathbf{X}$.

$\mathbf{z} \in \Omega \quad \Leftrightarrow \quad \exists\, \boldsymbol{\alpha} \in \mathbb{R}^{d+1} \quad \mathbf{z} = \mathbf{X}\boldsymbol{\alpha}$



The least square is the projection of $\mathbf{y}$ onto Ω. We have :

$$\mathbf{X}\widehat{\alpha} = \mathbf{H}\mathbf{y}$$

with the orthogonal projection operator $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$

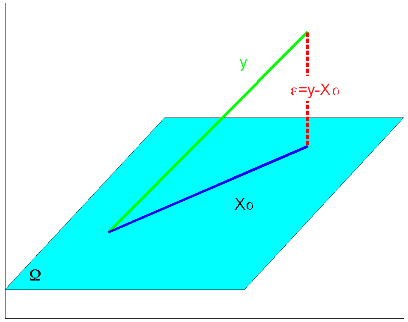
Estimation and orthogonality

Objective value $\|\boldsymbol{\epsilon}\|^2$ is minimal for $\boldsymbol{\alpha}$ such that $\mathbf{z} = \mathbf{X}\boldsymbol{\alpha}$ is the orthogonal projection of $\mathbf{y}$ on Ω. This means that

$$\forall \mathbf{z} \in \Omega \quad \mathbf{z}^\top \boldsymbol{\epsilon} = 0$$

Which means that the residual is orthogonal to all columns of $\mathbf{X}$

$$\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\alpha}}) = 0$$



$$\begin{aligned}
\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\alpha}}) = 0 \quad &\Leftrightarrow \quad \mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\alpha}} = 0\\
&\Leftrightarrow \quad \mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\alpha}} = \mathbf{X}^\top\mathbf{y}\\
&\Leftrightarrow \quad \widehat{\boldsymbol{\alpha}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}
\end{aligned}$$

Probablistic interpretation of least squares (1)

Observation model

We suppose that the observation model is the following:

$$y = \mathbf{x}^\top\mathbf{w} + b + \epsilon = \hat{\mathbf{x}}^\top\boldsymbol{\alpha} + \epsilon$$

- ϵ is a a centered random variable such that $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$.
- The probability of a given observation $(\mathbf{x}, y)$ when the parameters $\boldsymbol{\alpha}$ are know is then

$$p(\mathbf{x}, y|\boldsymbol{\alpha}) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(y - \hat{\mathbf{x}}^\top\boldsymbol{\alpha})^2}{2\sigma_n^2}\right)$$

Likelihood on the dataset

The likelihood for the whole dataset can be expressed as

$$\mathcal{L}(\boldsymbol{\alpha}) = \prod_{i=1}^n p(\mathbf{x}_i, y_i|\boldsymbol{\alpha}) = \left(\frac{1}{\sqrt{2\pi\sigma_n^2}}\right)^n \prod_{i=1}^n \exp\left(-\frac{(y_i - \hat{\mathbf{x}}_i^\top\boldsymbol{\alpha})^2}{2\sigma_n^2}\right)$$

Probablistic interpretation of least squares (2)

Maximum likelihood estimator

The MLE estimator is the solution of

$$\max_{\boldsymbol{\alpha}} \quad \mathcal{L}(\boldsymbol{\alpha})$$

In practice people often maximize the log-likelihood in order to have a simpler problem with the same solution.

Maximizing the Log-likelihood

$$\begin{aligned}
\log\left(\mathcal{L}(\boldsymbol{\alpha})\right) &= n \log\left(\frac{1}{\sqrt{2\pi\sigma_n^2}}\right) + \sum_{i=1}^n \log\left(\exp\left(-\frac{(y_i - \hat{\mathbf{x}}_i^\top\boldsymbol{\alpha})^2}{2\sigma_n^2}\right)\right)\\
&= n \log\left(\frac{1}{\sqrt{2\pi\sigma_n^2}}\right) - \frac{1}{2\sigma_n^2}\sum_{i=1}^n (y_i - \hat{\mathbf{x}}_i^\top\boldsymbol{\alpha})^2 = cst - J(\boldsymbol{\alpha})
\end{aligned}$$

Maximizing the log-likelihood wrt $\boldsymbol{\alpha}$ is equivalent to minimizing $J(\boldsymbol{\alpha})$.

Why regularize ?

Least Squares

We minimize the prediction error on the training data:

$$\min_{\boldsymbol{\alpha}} \quad J(\boldsymbol{\alpha}) \qquad \text{avec} \quad J(\boldsymbol{\alpha}) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}||^2$$

Problem solution is

$$\widehat{\boldsymbol{\alpha}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$$

Numerical problems

- When $n < d + 1$, matrix $\mathbf{X}^\top\mathbf{X}$ is non-invertible.
- There exists an infinity of solutions, problem is ill-posed.

$\Rightarrow$ regularization (among all possible solutions, pick the simplest).

Ridge Regression

Optimization problem

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \mathbf{w} - b)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 \qquad (12)$$

- We add a regularization term $\|\mathbf{w}\|^2$ weighted by the regularization coefficient $\lambda \geq 0$.
- Parameter λ can be chosen to limit over-fitting on the data.
- This regularization promotes parameters $\mathbf{w}$ of minimal norm.
- It make the optimization problem strictly convex (a unique solution).
- When $\lambda = 0$ problem boils down to the least squares (special case).

Matrix form of ridge regression

$$\min_{\alpha} \quad \left\{ J'(\alpha) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}||^2 + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \mathbf{S}\boldsymbol{\alpha} \right\} \qquad (13)$$

with $\mathbf{S} \in \mathbb{R}^{(d+1)\times(d+1)}$ a matrix defined as

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \qquad S_{i,j} = \begin{cases} 1 & \text{si } i = j \text{ et } i \leq d \\ 0 & \text{sinon} \end{cases} \qquad (14)$$

$\mathbf{S}$ is a diagonal matrix containing 1 on the diagonal except for the last term equal to 0 . Problem (12) and (13) are equivalent because

$$\boldsymbol{\alpha}^\top \mathbf{S}\boldsymbol{\alpha} = \sum_{i,j=1}^{d+1} \alpha_i \alpha_j S_{i,j} = \sum_{i=1}^{d} \alpha_i^2 = \sum_{i=1}^{d} \mathbf{w}_i^2 = \|\mathbf{w}\|^2 \qquad (15)$$

Gradient computation

$$\begin{array}{rcccccl} J'(\boldsymbol{\alpha}) & = & \frac{1}{2}\mathbf{y}^\top\mathbf{y} & - & \boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{y} & + & \overbrace{\frac{1}{2}\boldsymbol{\alpha}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\alpha} + \frac{\lambda}{2}\boldsymbol{\alpha}^\top\mathbf{S}\boldsymbol{\alpha}}^{\frac{1}{2}\boldsymbol{\alpha}^\top(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{S})\boldsymbol{\alpha}} \\ \dfrac{\partial J'(\boldsymbol{\alpha})}{\partial \alpha_i} & = & 0 & - & p_i & + & \frac{1}{2}\sum_{j=1}^{d+1}(M_{ij} + M_{ji})\alpha_j \end{array}$$

with $\mathbf{p} = \mathbf{X}^\top\mathbf{y}$ and $\mathbf{M} = \mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}$

- $\dfrac{\partial \boldsymbol{\alpha}^\top \mathbf{p}}{\partial \alpha_i} = \dfrac{\partial \sum_{j=1}^{d+1} p_j \alpha_j}{\partial \alpha_i} = p_i$
- $\dfrac{\partial \boldsymbol{\alpha}^\top \mathbf{M}\boldsymbol{\alpha}}{\partial \alpha_i} = \dfrac{\partial \sum_{j=1}^{d+1}\sum_{k=1}^{d+1} \alpha_j \alpha_k M_{jk}}{\partial \alpha_i} = \sum_{j=1}^{d+1} \alpha_j M_{ji} + \sum_{k=1}^{d+1} \alpha_k M_{ik}$

because $(uv)' = uv' + u'v$ with $u = \alpha_j$ and $v = \sum_{k=1}^{d+1} \alpha_k M_{jk}$

$$\nabla J'(\widehat{\alpha}) = -\mathbf{p} + \mathbf{M}\boldsymbol{\alpha} \qquad = -\mathbf{X}^\top\mathbf{y} \ + \ (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S})\widehat{\boldsymbol{\alpha}}$$

Ridge regression solution

Minimizing the cost $J'(\boldsymbol{\alpha})$ corresponds to finding the parameter $\boldsymbol{\alpha}$ that lads to a null gradient:

$$\nabla J'(\widehat{\alpha}) = 0 \Leftrightarrow -\mathbf{X}^\top\mathbf{y} \ + \ (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S})\widehat{\boldsymbol{\alpha}} = 0$$

The solution of the minimization problem for ridge regression is the vector $\widehat{\boldsymbol{\alpha}}$ defined as :

$$\widehat{\boldsymbol{\alpha}} = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}\right)^{-1}\mathbf{X}^\top\mathbf{y}$$

Regularization

Matrix $\mathbf{S}$ adds λ on the diagonal of $\mathbf{X}^\top\mathbf{X}$, making the matrix $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}$ invertible. The roblem is now well posed and has a unique solution.

Probabilistic interpretation of Ridge Regression

Prior distribution for the parameters

- In LS we had no prior information about $\boldsymbol{\alpha}$.
- We suppose that the $\mathbf{w}$ parameter has been drawn from $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma_p^2 \mathbf{I})$.
- Probability of a given $\mathbf{w}$ is : $p(\mathbf{w}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{(w_k)^2}{2\sigma_p^2}\right)$

Maximum likelihood estimator

$$
\begin{aligned}
\log\left(\mathcal{L}(\boldsymbol{\alpha})\right) &= cst - \frac{1}{2\sigma_n^2}\sum_{i=1}^{n}(y_i - \hat{\mathbf{x}}_i^\top \boldsymbol{\alpha})^2 - \frac{1}{2\sigma_p^2}\sum_{k=1}^{d}(w_k)^2 \\
&= cst - \frac{1}{2\sigma_n^2}\sum_{i=1}^{n}(y_i - \hat{\mathbf{x}}_i^\top \boldsymbol{\alpha})^2 - \frac{1}{2\sigma_p^2}\|\mathbf{w}\|^2
\end{aligned}
$$

- Maximizing the log-likelihood wrt $\boldsymbol{\alpha}$ is equivalent to minimizing $J(\boldsymbol{\alpha})$.
- Problems are equivalent when $\lambda = \frac{\sigma_n^2}{\sigma_p^2}$.

Lasso estimator

$$
\min_{\mathbf{w},b} \quad \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \mathbf{w} - b)^2 + \lambda \sum_{k=1}^{d} |w_k| \tag{16}
$$

Optimization problem

- $\|\mathbf{w}\|_1 = \sum_{k=1}^{d} |w_k|$ is the L1 norm of vector $\mathbf{w}$.
- Objective function is non differentiable in $w_k = 0, \forall k$.
- For a large enough λ the solution of the problem is sparse (some components of $\mathbf{w}$ are exactly 0).
- The problem is equivalent to

$$
\min_{\mathbf{w},b,\|\mathbf{w}\|_1 \leq \mu} \quad \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \mathbf{w} - b)^2 \tag{17}
$$

I.e. there exists a μ that leads to the same solution of the problem for a given λ.

Lasso with no bias

Data and model

- If the model has no bias b it means that the prediction can be expressed as

$$
f(\mathbf{x}) = \sum_i x_i w_i = \mathbf{x}^\top \mathbf{w}
$$

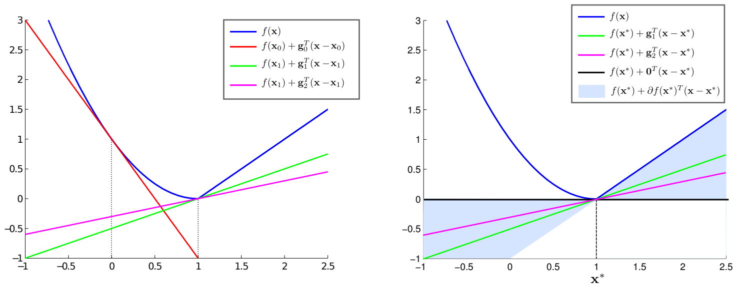
- $\mathbf{X}' \in \mathbb{R}^{n \times d}$ is the $\mathbf{X}$ matrix without the last columns containing ones.
- Prediction can be done with done with $\mathbf{X}'\mathbf{w}$ and $\mathbf{w}$ is teh only parameters

Matrix form of the optimization problem

$$
\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{X}'\mathbf{w}\|^2 + \lambda\|\mathbf{w}\|_1 \tag{18}
$$

- In the remaining we will focus on the Lasso with no bias for readability.
- Extension of the result when adding the bias is tedious but straightforward.

Subgradients and subdifferential



Non differentiable function

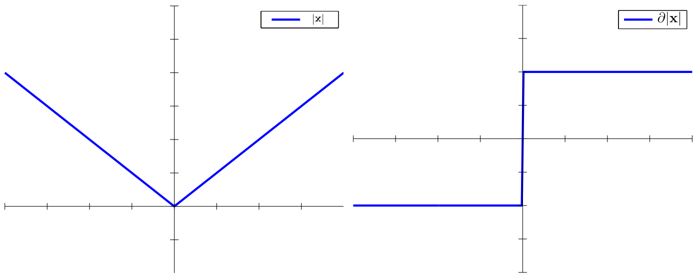
- A non differentiable function might not have a gradient (ex $|\cdot|$ in 0).
- The tool used in place of gradient is the subdifferential and subgradients.
- For a convex function $f(\mathbf{x})$, $\mathbf{g}$ is a subgradient of f in $\mathbf{x}_0$ if

$$
f(\mathbf{x}) \geq f(\mathbf{x}_0) + \mathbf{g}^\top(\mathbf{x} - \mathbf{x}_0) \tag{19}
$$

- The set of all subgradients at $\mathbf{x}_0$ is the subdifferential $\partial f(\mathbf{x}_0)$.
- $\mathbf{x}_0$ is a minimum of the convex function f if $\mathbf{0} \in \partial f(\mathbf{x}_0)$.

Subdifferential for the L1 norm

$$\|\mathbf{w}\|_1 = \sum_{k=1}^{d} |w_k|$$



L1 norm

- The subdifferential is of the form

$$\partial\|\mathbf{w}\|_1 = \begin{cases} \mathbf{g} : \|\mathbf{g}\|_\infty \leq 1 & \text{si } \mathbf{w} = 0 \\ \mathbf{g} : \|\mathbf{g}\|_\infty \leq 1 \text{ et } \mathbf{g}^T\mathbf{w} = \|\mathbf{w}\|_1 & \text{si } \mathbf{w} \neq 0 \end{cases}$$

- Which give

$$\partial\|\mathbf{w}\|_1 = \begin{cases} \mathbf{g} : g_k \in [-1, 1] & \text{si } \mathbf{w} = 0 \\ \mathbf{g} : g_k = sign(w_k) & \text{si } \mathbf{w} \neq 0 \end{cases}$$

Optimality conditions of the Lasso

Optimality conditions

$\mathbf{w}^\star$ is a solution of the optimization problem if

$$\mathbf{0} \in \partial J_{lasso}(\mathbf{w}^\star) \quad \text{with} \quad J_{lasso}(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}'\mathbf{w}\|^2 + \lambda\|\mathbf{w}\|_1$$

This can be reformulated as the following condition

$$-\mathbf{X}'^\top(\mathbf{y} - \mathbf{X}'\mathbf{w}^\star) + \lambda\mathbf{g} = \mathbf{0} \quad \text{with} \quad \mathbf{g} \in \partial\|\mathbf{w}^\star\|_1$$

Conditions on the components of $\mathbf{w}^\star$

$$\begin{array}{lcr} w_k^\star \neq 0 & \Rightarrow & -\mathbf{X}_k'^T(\mathbf{y} - \mathbf{X}'\mathbf{w}^\star) + \lambda \text{sign}(w_k^\star) = 0 \\ w_k^\star = 0 & \Rightarrow & |\mathbf{X}_k'^T(\mathbf{y} - \mathbf{X}'\mathbf{w}^\star)| \leq \lambda \end{array}$$

- $\mathbf{X}_k'$ is the kth column of $\mathbf{X}'$ (feature k).
- The is no closed-form solution except for special cases of $\mathbf{X}'$ (orthogonality).

Interpreting optimality conditions

Correlation with the residue

$$c_k = \mathbf{X}_k'^T(\mathbf{y} - \mathbf{X}'\mathbf{w}^\star) = \|\mathbf{X}_{.,k}\|\|\mathbf{y} - \mathbf{X}\mathbf{w}^\star\|\cos(\theta)$$

- c_k is the scalar product between the feature k and the residuals $\epsilon = \mathbf{y} - \mathbf{X}\mathbf{w}^\star$.
- θ is the angle between the two vectors.
- $c_k = 0, \ \forall k$ for Least Squares regression (optimality condition).

Effect of the regularization parameter λ

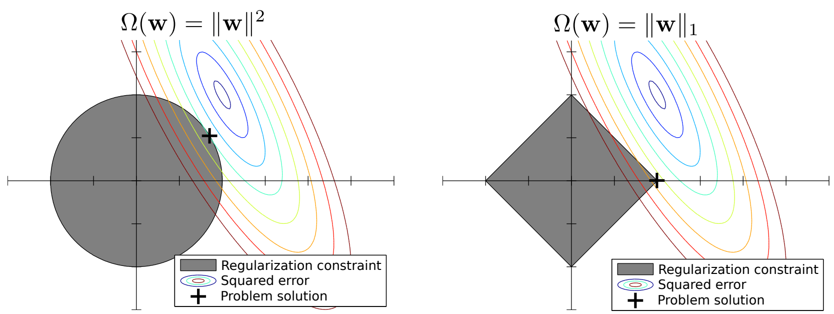
- $\lambda = 0$ boils down to Least Squares (no sparsity).
- If λ is small we have $w_k = 0$ only for variable k where

$$|\mathbf{X}_k'^T(\mathbf{y} - \mathbf{X}'\mathbf{w}^\star)| \leq \lambda$$

- If λ is very large at some point we have for all k,

$$|\mathbf{X}_k'^T\mathbf{y}| \leq \lambda \quad \text{which means} \quad w_k = 0, \ \forall k.$$

L2 VS L1 regularization



Tikhonov regularization

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{X}'\mathbf{w}\|^2 + \lambda\Omega(\mathbf{w})$$

$$\Leftrightarrow$$

Ivanov regularization

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{X}'\mathbf{w}\|^2 \quad \text{s.t.} \quad \Omega(\mathbf{w}) \leq \mu$$

The two optimization problems are equivalent for a strictly convex function Ω.

Solving the optimization problem

Least-angle regression (LARS)

- Algorithm developed by B. Efron, T. Hastie, I. Johnstone and R. Tibshirani.
- Allows to find efficiently the whole regularization path (all solution for all λ)
- Potiential problems with highly correlated variables.

Proximal gradient descent (PGD)

- Subgradientd ecsent is known to converge slowly.
- Proximal gradient descent allows for acceleration of the resolution.
- Can be seen a a Majoration-Minimization method.
- Each iteration is a simple soft thresholding of the parameter.
- Can be coupled for active sets to speedup sparse solutions.

Coordinate descent algorithm

- Optimize each components of $\mathbf{w}$ independently until conergence.
- Very fast for sparse solutions.

Coordinate descent for the Lasso

Algorithm

- Select an initial vector $\mathbf{w}$ (usually $\mathbf{w} = \mathbf{0}$).
- For all k : $w_k \leftarrow min_{w_k} \quad J_{lasso}(\mathbf{w})$ with all $w_j, j \neq k$ fixed
- Repeat until optimality conditions are statisfied.

Iteration for w_k

$$\begin{array}{ll} \min_{w_k} & \frac{1}{2}\|\mathbf{y} - \mathbf{X}'\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1 \\ \min_{w_k} & \frac{1}{2}\|\mathbf{y} - \sum_{j\neq k}\mathbf{X}'_i\mathbf{w}_i - \mathbf{X}'_k\mathbf{w}_k\|_2^2 + \lambda|w_k| \\ \min_{w_k} & \frac{1}{2}\|\mathbf{s} - \mathbf{X}'_k\mathbf{w}_k\|_2^2 + \lambda|w_k| \end{array}$$

were $\mathbf{s} = \mathbf{y} - \sum_{j\neq k}\mathbf{X}'_i\mathbf{w}_i$ is the residual wrt w_k. The last problem is a Lasso with only one variable, its solution is

$$w_k^\star = \text{sign}(\mathbf{X}_k'^T\mathbf{s})(|\mathbf{X}_k'^T\mathbf{s}| - \lambda)_+$$

This operator is called the soft thresholding.

Regularized linear regression

General problem formulation:

$$\min_{\mathbf{w},b} \quad \sum_{i=1}^{n} L(y_i, \mathbf{w}^\top\mathbf{x}_i + b) + \lambda\Omega(w) \tag{20}$$

With

- $L(\cdots)$ a loss function.
- $\Omega(\cdot)$ a regularization term.

Examples:

Loss function $L(y, \hat{y})$

- $(y - \hat{y})^2$, quadratic (this course).
- $|y - \hat{y}|$, absolute value.
- $\min(0, |y - \hat{y}| - \epsilon)$ epsilon insensitive

Regularizations $\Omega(\mathbf{w})$

- $\|\mathbf{w}\|_2^2$, quadratic.
- $\|\mathbf{w}\|_1$, ℓ_1 norm.
- $\mathbf{w}^\top\mathbf{\Sigma}\mathbf{w}$, Mahalanobis.