Linear regression

MATH-412 - Statistical Machine Learning

Design matrix, etc

Given a training set

$$D_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\},$$

we consider

- the design matrix $\boldsymbol{X}$
- output vector $\boldsymbol{y}$

$$\boldsymbol{X} = \begin{bmatrix} \text{---} & \mathbf{x}_1^\top & \text{---} \\ \text{---} & \mathbf{x}_2^\top & \text{---} \\ & \vdots & \\ \text{---} & \mathbf{x}_n^\top & \text{---} \end{bmatrix} \quad \text{and} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Remark : remember that most of the time it is relevant to

- center the data : $\mathbf{x}_i^{\mathsf{c}} = \mathbf{x}_i - \bar{\mathbf{x}}$
- normalize via e.g. $x_{ij}^{\mathsf{s}} = x_{ij}^{\mathsf{c}} / \widehat{\sigma}_j$ or mapping $\mathbf{x}_{ij}^{c}$ to $[0, 1]$, etc

Linear regression

- We consider the OLS regression for the linear hypothesis space.
- We have $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \mathbb{R}$ and ℓ the square loss.

Consider the hypothesis space :

$$S = \{f_{\boldsymbol{w}} \mid \boldsymbol{w} \in \mathbb{R}^p\} \qquad \text{with} \qquad f_{\boldsymbol{w}} : \mathbf{x} \mapsto \boldsymbol{w}^\top \mathbf{x}.$$

Given a training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ we have

$$\widehat{\mathcal{R}}_n(f_w) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{w}^\top \mathbf{x}_i)^2 = \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2$$

with

- the vector of outputs $\boldsymbol{y}^\top = (y_1, \ldots, y_n) \in \mathbb{R}^n$
- the design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ whose ith row is equal to $\mathbf{x}_i^\top$.

Solving linear regression

To solve $\min_{\boldsymbol{w}\in\mathbb{R}^p} \widehat{\mathcal{R}}_n(f_{\boldsymbol{w}})$, we consider that

$$\widehat{\mathcal{R}}_n(f_w) = \frac{1}{2n}\left(\boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} - 2\,\boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{y} + \|\boldsymbol{y}\|^2\right)$$

is a differentiable convex function whose minima are thus characterized by the

Normal equations

$$\boxed{\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{0}}$$

If $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible, then there is a unique solution to the normal equations and and $\widehat{f}$ is given by :

$$\widehat{f} : \mathbf{x}' \mapsto \mathbf{x}'^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

Problem : $\boldsymbol{X}^\top \boldsymbol{X}$ is never invertible for $p > n$ and thus the solution is not unique.

Linear or affine regression ?

$$f_{\boldsymbol{w}}(\mathbf{x}) = \boldsymbol{w}^\top \mathbf{x} \qquad \text{vs} \qquad f_{\boldsymbol{w},b}(\mathbf{x}) = \boldsymbol{w}^\top \mathbf{x} + b = \widetilde{\boldsymbol{w}}^\top \widetilde{\mathbf{x}}$$

With

$$\widetilde{\boldsymbol{w}} = \begin{bmatrix} \boldsymbol{w} \\ b \end{bmatrix} \quad \text{and} \quad \widetilde{\mathbf{x}} = \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}$$

- ... an affine model in dimension p is a linear model in dimension $p+1$
- These two models are equivalent when we don't regularize, otherwise not because usually b is not regularized.
- Exercise : What is the value of $\hat{b}$ if the data is centered ?

Hat matrix and geometry of linear regression

If $\boldsymbol{X}$ has full column rank, then $\widehat{\boldsymbol{w}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$,
so that for the training data

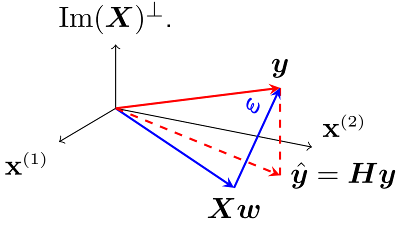
$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{w}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \qquad \text{with} \quad \boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top.$$

Let $r = \mathrm{rank}(\boldsymbol{X})$, and $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^\top$ be the reduced form of the eigenvalue decomposition of $\boldsymbol{X}\boldsymbol{X}^\top$ with

- $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ an orthonormal matrix
- $\boldsymbol{S} \in \mathbb{R}^{r\times r}$ a diagonal matrix with (strictly) positive entries.

then $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{U}^\top$ and $\boldsymbol{H}$ is the orthogonal projector on $\mathrm{Im}(\boldsymbol{X})$.

$$\boldsymbol{X} = [\mathbf{x}^{(1)}\mathbf{x}^{(2)}] \in \mathbb{R}^{n\times 2}$$



Optimality of least squares linear regression

Assume that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with

Full column rank design : $\text{rank}(\boldsymbol{X}) = p$

Decorrelated centered noise : $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$ and $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] = \sigma^2 \boldsymbol{I}$

Gauss-Markov Theorem :

Then $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ is the best linear unbiased estimator (BLUE)
that is that for any other *unbiased* estimator $\widetilde{\boldsymbol{\beta}}$ we have

$$\text{Cov}(\widetilde{\boldsymbol{\beta}}) = \text{Cov}(\widehat{\boldsymbol{\beta}}) + \boldsymbol{K}_{\widetilde{\boldsymbol{\beta}}} \quad \text{with } \boldsymbol{K}_{\widetilde{\boldsymbol{\beta}}} \text{ positive semi-definite.}$$

Remarks :

- Requires that the data is really generated from the linear model
- That the noise is decorrelated and homoscedastic.
- Compares only with *linear* and *unbiased* estimators.

Gaussian conditional model and least square regression

Modeling the conditional distribution of Y given X by

$$Y \mid X \sim \mathcal{N}(\boldsymbol{\beta}^\top X, \sigma^2)$$

Likelihood for one pair

$$p(y_i \mid \mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{1}{2}\frac{(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2}{\sigma^2}\right)$$

Negative log-likelihood

$$-\ell(\boldsymbol{\beta}, \sigma^2) = -\sum_{i=1}^n \log p(y_i|\mathbf{x}_i) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2}\sum_{i=1}^n \frac{(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2}{\sigma^2}.$$

Gaussian conditional model and least square regression

$$\min_{\sigma^2, \boldsymbol{\beta}} \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2}{\sigma^2}$$

The minimization problem in $\boldsymbol{w}$

$$\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

that we recognize as the usual linear regression.
Optimizing over σ^2, we find :

$$\widehat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{\boldsymbol{\beta}}^\top_{MLE} \mathbf{x}_i)^2$$

Properties if the model is well-specified

Assume that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ with

Full column rank *fixed* design : $\mathrm{rank}(\boldsymbol{X}) = p$ (which implies $n \geq p$).

I.i.d. centered **Gaussian** noise : $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$

then

- $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\beta}^*, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1})$
- $S^2 = \frac{1}{n-p} \|\hat{\boldsymbol{y}} - \boldsymbol{y}\|_2^2 \sim \frac{\sigma^2}{n-p} \chi^2_{n-p}$
- $\widehat{\boldsymbol{\beta}}$ and S^2 are independent

All of these are used for

- ANOVA, t-test and to construct confidence intervals
- Only valid if the data is Gaussian ($=$ model is well-specified)