Regression
2. Linear Least Squares
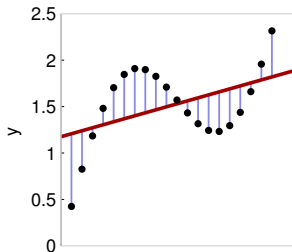
Olivier Sigaud

Sorbonne Université
http://people.isir.upmc.fr/sigaud

Linear
least squares

Linear Least Squares

Black dots represent 20 training examples, and the thick red line is the learned model $\hat{f}(\mathbf{x})$. Vertical lines represent residuals $\|y - \hat{f}(\mathbf{x})\|$.

- ▶ We want to minimize the squared sum of the residuals
- ▶ In matrix form, we want to find: $min(\mathbf{y} - \hat{f}(\mathbf{X}))^2$.

Adrien Marie Legendre (1805) *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot.

Carl Friedrich Gauss (1809) *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, volume 7. Perthes et Besser.

Offset trick

▶ In the linear case, we want $\mathbf{y} = \hat{f}(\mathbf{X}) = \mathbf{X}\hat{\boldsymbol{\theta}} + \mathbf{b}$

▶ We can remove the offset $\mathbf{b}$ by increasing $\mathbf{X}$ with a row of ones.

▶
$$\hat{f}(\mathbf{X}) = \begin{pmatrix} \mathbf{X} & 1 \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\theta}} \\ \mathbf{b} \end{pmatrix}$$

▶
$$\bar{\mathbf{X}} = \begin{pmatrix} \mathbf{x}_{1,1} & \mathbf{x}_{1,2} & \cdots & \mathbf{x}_{1,D} & 1 \\ \mathbf{x}_{2,1} & \mathbf{x}_{2,2} & \cdots & \mathbf{x}_{2,D} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{x}_{N,1} & \mathbf{x}_{N,2} & \cdots & \mathbf{x}_{N,D} & 1 \end{pmatrix}, \boldsymbol{\theta} = \begin{pmatrix} \hat{\boldsymbol{\theta}} \\ \mathbf{b} \end{pmatrix}$$

▶ If we rewrite $\mathbf{y} = \hat{f}(\mathbf{X}) = \bar{\mathbf{X}}\boldsymbol{\theta}$, $\boldsymbol{\theta}$ is a vector of weights

▶ We minimize the residuals, thus

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \underbrace{\|\mathbf{y} - \bar{\mathbf{X}}\boldsymbol{\theta}\|^2}_{L(\boldsymbol{\theta})}$$

ISIR
INSTITUT
DES SYSTÈMES
INTELLIGENTS
ET DE ROBOTIQUE

Optimal linear model

$$L(\boldsymbol{\theta}) = \|\mathbf{y} - \bar{\mathbf{X}}\boldsymbol{\theta}\|^2 \tag{1}$$

$$= (\mathbf{y} - \bar{\mathbf{X}}\boldsymbol{\theta})^\intercal(\mathbf{y} - \bar{\mathbf{X}}\boldsymbol{\theta}) \tag{2}$$

$$= \mathbf{y}^\intercal\mathbf{y} - \mathbf{y}^\intercal\bar{\mathbf{X}}\boldsymbol{\theta} - (\bar{\mathbf{X}}\boldsymbol{\theta})^\intercal\mathbf{y} + (\bar{\mathbf{X}}\boldsymbol{\theta})^\intercal(\bar{\mathbf{X}}\boldsymbol{\theta}) \tag{3}$$

$$= \mathbf{y}^\intercal\mathbf{y} - 2(\bar{\mathbf{X}}\boldsymbol{\theta})^\intercal\mathbf{y} + (\bar{\mathbf{X}}\boldsymbol{\theta})^\intercal(\bar{\mathbf{X}}\boldsymbol{\theta}) \tag{4}$$

At a minimum of a function, its derivative is null

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 2(\bar{\mathbf{X}}^\intercal\bar{\mathbf{X}}\boldsymbol{\theta} - \bar{\mathbf{X}}^\intercal\mathbf{y})$$

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0 \rightarrow \bar{\mathbf{X}}^\intercal\bar{\mathbf{X}}\boldsymbol{\theta} = \bar{\mathbf{X}}^\intercal\mathbf{y}$$

Thus min reached where $\boldsymbol{\theta}^* = (\bar{\mathbf{X}}^\intercal\bar{\mathbf{X}})^{-1}\bar{\mathbf{X}}^\intercal\mathbf{y}$

Regularized Least Squares

- Potential singularities in $\bar{\mathbf{X}}^\intercal\bar{\mathbf{X}}$ can generate very large $\boldsymbol{\theta}^*$ weights
- Regularized Least Squares (Ridge Regression, RR): penalize large weights
- Optimize with lower weights (sacrifice optimality):
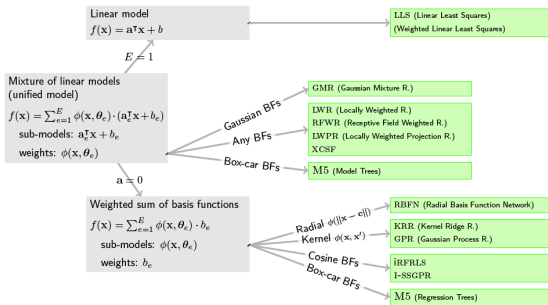- $$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}}\frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 + \frac{1}{2}\|\mathbf{y} - \bar{\mathbf{X}}\boldsymbol{\theta}\|^2, \tag{5}$$
- Analytical solution:
$$\boldsymbol{\theta}^* = (\lambda\boldsymbol{I} + \bar{\mathbf{X}}^\intercal\bar{\mathbf{X}})^{-1}\bar{\mathbf{X}}^\intercal\mathbf{y}. \tag{6}$$
- Tikhonov regularization $=$ Ridge regression

Next lessons: non-linear regression

Linear model
$f(\mathbf{x}) = \mathbf{a}^{\mathsf{T}}\mathbf{x} + b$

LLS (Linear Least Squares)
(Weighted Linear Least Squares)

$E = 1$

Mixture of linear models
(unified model)
$f(\mathbf{x}) = \sum_{e=1}^{E} \phi(\mathbf{x}, \boldsymbol{\theta}_e) \cdot (\mathbf{a}_e^{\mathsf{T}}\mathbf{x} + b_e)$
sub-models: $\mathbf{a}_e^{\mathsf{T}}\mathbf{x} + b_e$
weights: $\phi(\mathbf{x}, \boldsymbol{\theta}_e)$

Gaussian BFs

GMR (Gaussian Mixture R.)

LWR (Locally Weighted R.)
RFWR (Receptive Field Weighted R.)
LWPR (Locally Weighted Projection R.)
XCSF

Any BFs

Box-car BFs

M5 (Model Trees)

$\mathbf{a} = 0$

Weighted sum of basis functions
$f(\mathbf{x}) = \sum_{e=1}^{E} \phi(\mathbf{x}, \boldsymbol{\theta}_e) \cdot b_e$
sub-models: $\phi(\mathbf{x}, \boldsymbol{\theta}_e)$
weights: b_e

Radial $\phi(\|\mathbf{x} - \mathbf{c}\|)$
Kernel $\phi(\mathbf{x}, \mathbf{x}')$

RBFN (Radial Basis Function Network)

KRR (Kernel Ridge R.)
GPR (Gaussian Process R.)

Cosine BFs
Box-car BFs

iRFRLS
I-SSGPR

M5 (Regression Trees)

▶ Two different approaches:
  ▶ Multiple local and weighted least square regressions (shown with LWR)
  ▶ Projecting the input space into a feature space using non-linear basis functions (shown with RBFNs)
▶ We provide unifying views of algorithms from each family
▶ Then we highlight the similarity between both approaches

Stulp, F. and Sigaud, O. (2015). Many regression algorithms, one unified model: A review. *Neural Networks*, 69:60–79.

Any question?

Send mail to: `Olivier.Sigaud@upmc.fr`

Gauss, C. F. (1809).
Theoria motus corporum coelestium in sectionibus conicis solem ambientium, volume 7.
Perthes et Besser.

Legendre, A. M. (1805).
Nouvelles méthodes pour la détermination des orbites des comètes.
F. Didot.