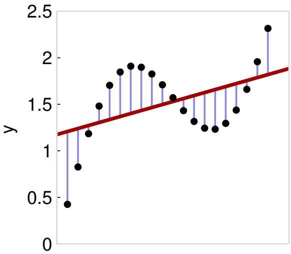
Regression

2. Linear Least Squares

Olivier Sigaud

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

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: $min(\mathbf{y} - \hat{f}(\mathbf{X}))^2$.

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 $\mathbf{y} = \hat{f}(\mathbf{X}) = \boldsymbol{\theta}^\intercal \mathbf{X} + \mathbf{b}$
- We can remove the offset $\mathbf{b}$ by increasing $\mathbf{X}$ with a row of ones.
- $$\hat{f}(\mathbf{X}) = \begin{pmatrix} \boldsymbol{\theta} \\ \mathbf{b} \end{pmatrix}^\intercal \begin{pmatrix} \mathbf{X} & 1 \end{pmatrix}$$
- $$\mathbf{X} = \begin{pmatrix} \mathbf{x}_{1,1} & \mathbf{x}_{1,2} & \cdots & \mathbf{x}_{1,D} & 1 \\ \mathbf{x}_{2,1} & \mathbf{x}_{2,2} & \cdots & \mathbf{x}_{2,D} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{x}_{N,1} & \mathbf{x}_{N,2} & \cdots & \mathbf{x}_{N,D} & 1 \end{pmatrix}$$
- If we rewrite $\mathbf{y} = \hat{f}(\mathbf{X}) = \boldsymbol{\theta}^\intercal \mathbf{X}$, $\boldsymbol{\theta}$ is a vector of weights
- We minimize the residuals, thus

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

Optimal linear model

$$L(\boldsymbol{\theta}) = \|\mathbf{y} - \boldsymbol{\theta}^\intercal \mathbf{X}\|^2 \quad (1)$$
$$= (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\intercal (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \quad (2)$$
$$= \mathbf{y}^\intercal \mathbf{y} - \mathbf{y}^\intercal \mathbf{X}\boldsymbol{\theta} - (\mathbf{X}\boldsymbol{\theta})^\intercal \mathbf{y} + (\mathbf{X}\boldsymbol{\theta})^\intercal (\mathbf{X}\boldsymbol{\theta}) \quad (3)$$
$$= \mathbf{y}^\intercal \mathbf{y} - 2(\mathbf{X}\boldsymbol{\theta})^\intercal \mathbf{y} + (\mathbf{X}\boldsymbol{\theta})^\intercal (\mathbf{X}\boldsymbol{\theta}) \quad (4)$$

At a minimum of a function, its derivative is null

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

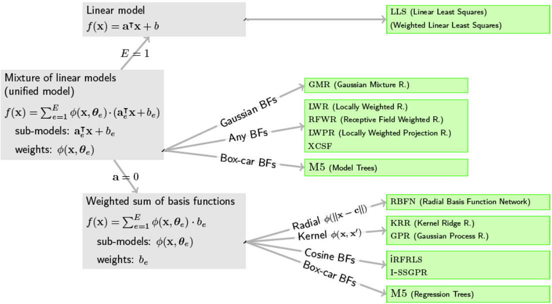
$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0 \rightarrow \boldsymbol{\theta}\mathbf{X}^\intercal \mathbf{X} = \mathbf{X}^\intercal \mathbf{y}$$

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

Regularized Least Squares

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

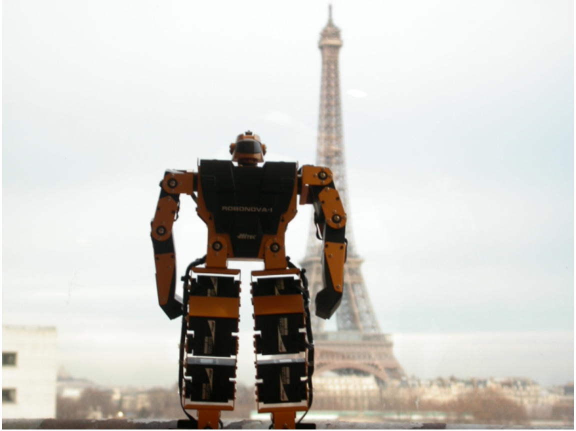
Next classes: regression for robotics



- 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`