Least Squares and the Normal Distribution

Prof. Dr. Christoph Lippert

Digital Health & Machine Learning

Regression

- Given a dataset $\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N}$, where $\mathbf{x}^{(i)}$ is D dimensional, fit parameters $\mathbf{w}$ of a regressor f with added Gaussian noise:

$$y^{(i)} = f(\mathbf{x}^{(i)}; \boldsymbol{\theta}) + \epsilon^{(i)} \quad \text{where} \quad p(\epsilon^{(i)} | \sigma^2) = \mathcal{N}\left(\epsilon^{(i)} \,\middle|\, 0, \sigma^2\right).$$

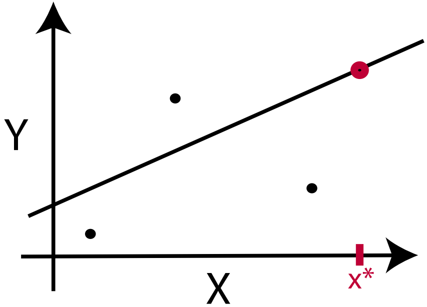
- Equivalent likelihood formulation:

$$p(\mathbf{y}|\mathbf{X}) = \prod_{i=1}^{N} \mathcal{N}\left(y^{(i)} \,\middle|\, f(\mathbf{x}^{(i)}; \boldsymbol{\theta}), \sigma^2\right)$$

Regression

- Choose f to be linear:

$$p(\mathbf{y}|\mathbf{X}) = \prod_{i=1}^{N} \mathcal{N}\left(y^{(i)} \,\middle|\, \mathbf{w}^\top \mathbf{x}^{(i)} + b, \sigma^2\right)$$



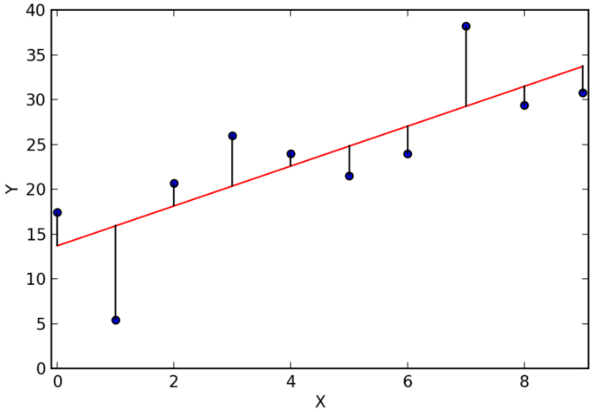
Maximum likelihood

- Taking the logarithm, we obtain

$$\begin{aligned}\ln p(\mathbf{y}|\mathbf{X};\mathbf{w},\sigma^2) &= \sum_{i=1}^{N} \ln \mathcal{N}\left(y^{(i)} \,\middle|\, \mathbf{w}^\top \mathbf{x}^{(i)}\ ,\ \sigma^2\right) \\ &= -\frac{N}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{N}(y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)})^2}_{\text{Sum of squares}}\end{aligned}$$

- The likelihood is maximized when the squared error is minimized.
- Least squares and maximum likelihood are equivalent.

Maximum Likelihood and Least Squares



$$\text{argmin}_{\boldsymbol{\beta}} \frac{1}{2} \sum_{n=1}^{N} (y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)})^2$$

- Derivative w.r.t a single weight entry β_j

$$\frac{\partial}{\partial w_j} \ln p(\mathbf{y}|\mathbf{X}\ ;\ \mathbf{w}, \sigma^2) = \frac{\partial}{\partial w_j} \left[-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)})^2 \right]$$

$$= \left[-\frac{1}{\sigma^2} \sum_{i=1}^{N} x_d^{(i)} (y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)}) \right]$$

- Set gradient w.r.t. $\mathbf{w}$ to zero [vector holding the partial derivatives $\forall w_j$]

$$\nabla_{\mathbf{w}} \ln p(\mathbf{y}|\mathbf{X}\ ;\ \mathbf{w}, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{N} \mathbf{x}^{(i)\top} (y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)}) = \mathbf{0} \qquad \text{(where } \mathbf{0} \text{ is a vector of 0s)}$$

$$\frac{1}{\sigma^2} \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) = \mathbf{0}$$

$$\Longrightarrow \boldsymbol{\beta}_{\text{ML}} = \underbrace{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}}_{\text{Pseudo inverse of } \mathbf{X}} \mathbf{y}$$

- Here, the matrix $\mathbf{X}$ is defined as $\mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1D} \\ \dots & \dots & \dots \\ x_{N1} & \dots & x_{ND} \end{bmatrix}$

Summary

- Linear Regression
  - Assumes normal distributed residuals
  - Maximum Likelihood Estimation
  - equivalent to Least Squares