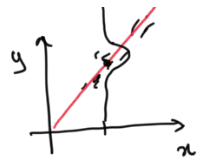


Linear regression

Setup: Given $\mathcal{D} := \{(x_1, y_1), \dots, (x_n, y_n)\}$, $x_i \in \mathbb{R}^d$, $y \in \mathbb{R}$

Workflow:

i) Model definition : parametrization, likehood, prior
$\hookrightarrow \theta \in \mathbb{R}^p$

ii) Training to identify $\underline{\theta^*}$ $\big(\underbrace{p(\theta|\mathcal{D})}_{\text{posterior}}\big)$

iii) Perform predictions $f_\theta(x^*)$? $\underbrace{p(f(x^*)|x^*, \mathcal{D})}_{\text{predictive posterior}}$

1) Model:

$$y = \underbrace{f_\theta(x)}_{\text{model form}} + \underbrace{\varepsilon}_{\text{noise}} \qquad \left\{\begin{array}{l} \cdot\ f_\theta(x) = w^T x, \quad w \in \mathbb{R}^d \\ \cdot\ \varepsilon \sim \mathcal{N}(0, \sigma_n^2) \end{array}\right\} \checkmark$$

Parameters: $\theta := \{w_1, \dots, w_d, \sigma_n^2\}$

Likehood: $y_i \overset{\text{i.i.d.}}{\sim} p(y_i | x_i, \theta) = \mathcal{N}(y_i | w^T x_i, \sigma_n^2)$

$$y_1, \dots, y_N \sim \mathcal{N}(y | Xw, \sigma_n^2 I)$$

$$\underset{n\times1}{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \underset{n\times d}{X} = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & & \\ x_{n1} & \cdots & x_{nd} \end{bmatrix}, \quad \underset{d\times1}{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$$

2.) Training

: Goal $\rightarrow$ estimate θ^* that "best" explain $\mathcal{D}$.

$$\theta_{MLE} = \underset{\theta}{\arg\max}\ p(y | X, w, \sigma^2)$$

Likelihood: $p(\mathcal{D}|\theta) = p(y_1, \dots, y_n | x_1, \dots, x_n, w_1, \dots, w_d, \sigma_n^2)$

$$\overset{\text{i.i.d.}}{=} \prod_{i=1}^{n} \mathcal{N}(y_i \mid w^T x_i, \sigma_n^2)$$

$$\mathcal{L}(w, \sigma_n^2) := -\log p(y \mid X, w, \sigma^2)$$

$$= \frac{n}{2}\log(2\pi\sigma_n^2) + \frac{1}{2\sigma_n^2}\overbrace{(y - Xw)^T(y - Xw)}^{\sum_{i=1}^{n}(y_i - w^T x_i)^2}$$

Assume that σ_n^2 is known, hence we will only estimate W.

- optimization $\begin{cases} \text{GD}: & w^{n+1} = w^n - \eta \nabla_w \mathcal{L}(w^n) \\ \text{Newton}: & w^{n+1} = w^n - \eta H_n^{-1} \nabla_w \mathcal{L}(w^n) \end{cases}$

$$\left.\begin{aligned} \nabla_w \mathcal{L}(w) &:= -X^T y + X^T X w \\ \nabla_w^2 \mathcal{L}(w) &= X^T X \end{aligned}\right\} \Longrightarrow \begin{aligned} &\text{GD}: w^{n+1} = w^n - \eta\{X^T X w^n - X^T y \\ &\text{Newton}: w^{n+1} = w^n - \eta(X^T X)^{-1}[X^T X w^n \end{aligned}$$

Solve for W_{MLE} analytically:

$$W_{MLE} = \arg\min_{w} -\log p(y \mid X, W) := \mathcal{L}(w)$$

$$\mathcal{L}(w) = \frac{n}{2}\log(2\pi\sigma_n^2) + \frac{1}{2\sigma_n^2}\underbrace{(y - Xw)^T(y - Xw)}_{\text{quadratic form}} \leftarrow$$

Identify critical points:

$$\nabla_w \mathcal{L}(w) = 0$$

- $\frac{1}{2}(y - Xw)^T(y - Xw) = \frac{1}{2}\left(y^T y - y^T X w - (Xw)^T y + (Xw)^T X w\right)$

$$= \frac{1}{2} y^T y - \underbrace{y^T X w}_{w^T X^T y} + w^T X^T X w \frac{1}{2} \}$$

$$\nabla_w \mathcal{L}(w) = -X^T y + X^T X w$$

Condition satisfied by critical points:

$$\nabla_w \mathcal{L}(w) = 0 \implies \boxed{w_{MLE} = \underbrace{(X^T X)^{-1}}_{\text{needs to be invertible.}} X^T y} \rightarrow$$ Solution a least squ regressi proble

Observe that $H := \nabla_w^2 \mathcal{L}(w) = \underbrace{X^T X}_{} \rightarrow$ symmetric positive-definite

$\implies \mathcal{L}(w)$ is strictly convex in W and W_{MLE} is a unique global minimizer.

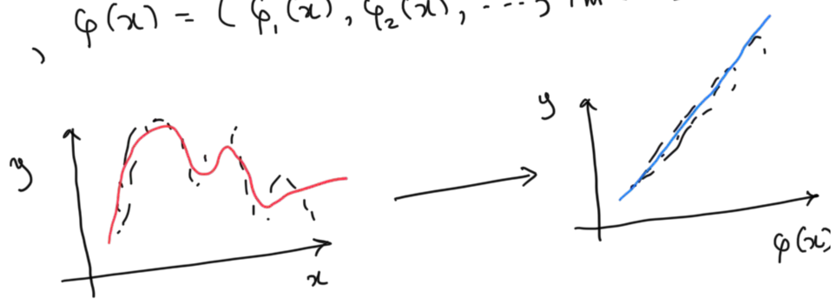
Linear regression with basis functions:

Model setup: $y = f_\theta(x) + \varepsilon$ $\begin{cases} \cdot\ f_\theta(x) = w^T \varphi(x) \\ \cdot\ \varepsilon \sim \mathcal{N}(0, \sigma_n^2) \end{cases}$

basis functions

$\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^m$, $\varphi(x) = (\varphi_1(x), \varphi_2(x), \ldots, \varphi_m(x))$

φ: features mapping; $\mathbb{R}^d$: input space; $\mathbb{R}^m$: features space



$$\underset{n\times 1}{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \underset{n\times m}{\Phi} = \begin{bmatrix} \varphi_1(x_1) & \cdots & \varphi_m(x_1) \\ \vdots & & \\ \varphi_1(x_n) & \cdots & \varphi_m(x_n) \end{bmatrix}, \quad \underset{m\times 1}{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix}$$

2.) MLE for w:

$$\underset{m\times 1}{w_{MLE}} = \underbrace{(\underset{m\times n}{\Phi^T} \underset{n\times m}{\Phi})^{-1}}_{m\times m} \underbrace{\underset{m\times n}{\Phi^T} \underset{n\times 1}{y}}_{m\times 1}$$

Maximum a-posteriori estimation (MAP):

Setup: Given $\mathcal{D} := \{(x_1, y_1), \ldots, (x_n, y_n)\}$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

Model : $y = f_\theta(x) + \varepsilon \iff p(y|x,\theta)$: likelihood

assume a prior $\theta \sim p(\theta)$: prior

Bayes rule : $\underbrace{p(\theta|D)}_{\text{posterior}} = \dfrac{\overbrace{p(D|\theta)}^{\text{likelihood}}\,\overbrace{p(\theta)}^{\text{prior}}}{p(D)}$; $p(D)$: marginal likelihood / model evid

$\hookrightarrow \int p(D|\theta)\,p(\theta)\,d\theta$

Goal : $\theta_{MAP} = \arg\max_\theta p(\theta|D)$ (vs $\theta_{MLE} = \arg\max_\theta p(D|\theta)$)

Recall, for linear regression:

$$\underline{p(w|X,y)} \overset{\text{Bayes}}{\propto} \underbrace{p(y|X,w)}\cdot\underbrace{p(w)}$$

(omitting $p(y)$ since it does not depend on θ)

$$\theta_{MAP} = \arg\min_\theta \; -\underbrace{\log p(w|X,y)}$$

(*) We need to assume a prior for $w \sim p(w)$. The simplest choice is to assume $\boxed{p(w) = \mathcal{N}(0, \lambda^{-1}}$

$$-\log p(w|X,y) = -\underbrace{\log p(y|X,w)} - \log p(w)$$

$w^Tw := \|w\|_2^2$

$$= \cancel{\frac{n}{2}\log(2\pi\sigma_n^2)} + \underbrace{\frac{1}{2\sigma_n^2}(y-Xw)^T(y-Xw)}_{\text{likelihood}} + \underbrace{\frac{\lambda}{2}w^Tw}_{\text{prior}} := \underline{\mathcal{L}(w)}$$

Critical points :

$$\nabla_w \mathcal{L}(w) = 0 \Rightarrow w_{MAP} = (X^TX + \underline{\lambda I})^{-1}X^Ty$$

vs

$$w_{MLE} = (X^T X)^{-1} X^T y$$

Comments on MLE vs MAP :

Pros of MAP : • easy to compute and interpetable (interpolates between th MLE and the prio

• It is more resilient against overfitting

• Tends to look like the MLE assymptotically $(n \to \infty)$

Cons of MAP :

• It is just a point-estimate (no quantification of uncertainty)

• Unlike the MLE, the MAP is not invariant to re-parametrization.

• Must assume an appropriate prior for θ, possible choices :

$$\begin{cases} \|\theta\|_2 \longleftarrow p(\theta) \sim \mathcal{N}(0, b^{-1}) \longrightarrow \text{promote "simple" models} \\ \|\theta\|_1 \longleftarrow p(\theta) \sim \mathcal{L}ap(b) \longrightarrow \text{promote "sparsify"} \end{cases}$$