Given data $(x_1, y_1), \dots (x_n, y_n) \in \mathbb{R}^D \times \mathbb{R}^N$

Find a function

$f: \mathbb{R}^D \longrightarrow \mathbb{R}^N$ so that

1. $f(x_i) \approx y_i \quad 1 \leq i \leq n$
2. For each new data point (x, y), $f(x) \approx f(y)$

Train Accuracy : 100%
Test Accuracy: 10%

x_i = input data / attributes

y_i = labels / output variables

f_θ = model | depending on θ

Machine learning - goal is to learn θ

Algorithm 3.1

1. Select a model
2. Split data into training + test data
3. Learn parameters θ
4. Validation of θ_* Choice of parameter

Validation options

~ Choose parameter θ_* which minimizes

ERM • Empirical risk ~ $R(\theta) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, f_\theta(x_i))$

$l : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ loss function such as distance, squared distance, etc.

MLE • likelihood function ~ $\prod_{i=1}^{n} P_\theta(y_i | x_i)$

log-likelihood ~

$$l(\theta) := \sum_{i=1}^{n} \log P_\theta(y_i | x_i)$$

MAP : posteriori function: $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^{n \times 1}$ $y \in \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^{n \times N}$

$$\alpha(\theta) = P(\theta | X, Y)$$

When ERM = linear Regression

$$f_\theta(x) = x^T \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_D \end{pmatrix} + \theta_0 \qquad l(y, \hat{y}) = (y - \hat{y})^2$$

MLE and MAP – an overview

We will first focus on Linear MLE, then MAP.

For nonlinear versions see Thm 3.13, Prop 3.14.

Select a model $(x_1, y_1), \dots (x_n, y_n) \in \mathbb{R}^D \times \mathbb{R}$

Deterministic

- linear model $f_\theta : \mathbb{R}^D \to \mathbb{R}$

$$\theta = (\theta_0, \dots, \theta_D) \in \mathbb{R}^{D+1}$$

$$f_\theta(x) = \theta_0 + x^T \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_D \end{pmatrix}$$

(w/ quadratic loss called linear regression)

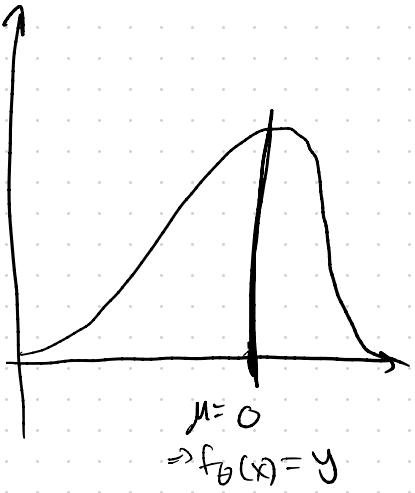
- linear probabalistic model

mean, variance

$$P_\theta(y|x) = \Phi(y \mid f_\theta(x), \sigma^2)$$

Φ: density function of normal distribution

$$= \frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-(f_\theta(x)-y)^2}{2\sigma^2}}$$



] linear model + random noise

$$\Rightarrow y = f_\theta(x) + \varepsilon \quad \text{for } \varepsilon \sim N(0, \sigma^2)$$

~ Julia notation

```
f(x::Number, θ) = θ[2]*x + θ[1]
f(X::Vector, θ) = [f(x,θ) for x ∈ X]
```

Types must be specified in first input because function f has 2 definitions & Julia needs to know which one to use.

- Given a model, we need to "learn" parameter θ, i.e. choose optimal value of θ with some measure

Deterministic version:

- Linear regression, Empirical Risk minimization

$$\theta_* = \arg\min_\theta R(\theta)$$

$$R(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i))$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f_\theta(x_i)\right)^2$$

$$\Omega = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \in \mathbb{R}^{n \times (D+1)}$$

feature matrix

Theorem 3.8 When $\text{rank}(\Omega) = r(\Omega) = D+1$ then $\underset{\theta}{\text{argmin}}\, R(\theta)$ has a unique solution

$$\theta^* = \Omega^+ y.$$

$$\Omega^+ = (\Omega^T \Omega)^{-1} \Omega^T$$

Julia note ~ in example we had

$$\theta\_ERM = \Omega \backslash y$$

↑ left-division

By Prop 1.8 (see HW 1 exercise)

Since $\Omega \in \mathbb{R}^{n \times (D+1)}$ has rank $D+1$, it is left-invertible, so can compute left division.

Thm 3.8 pf

$$R(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(f_\theta(x_i) - y_i \right)^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_i^T \theta' + \theta_0 - y_i \right)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(x_i^T \theta' + \theta_0 - y_i \right)\left(x_i^T \theta' + \theta_0 - y_i \right)$$

Notice

$$\Omega\theta = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_D \end{bmatrix} = \theta_0 + x_i^T \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_D \end{pmatrix}$$

$n \times D+1$ $\quad$ $(D+1) \times 1$

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^{n \times D}$$

$$= \frac{1}{n} \left(\Omega\theta - Y \right)^T \left(\Omega\theta - Y \right)$$

$(1 \times n) \times (n \times 1)$

Recall Pseudoinverse

$\Omega\theta = Y$ solved by $\Omega^+ Y = \tilde{\theta}$

$\tilde{\theta} = \mathrm{Im}(\Omega^T)$, $\Omega\tilde{\theta} = Y_0$ and

$$Y_0 = \underset{w \in \mathrm{Im}(\Omega)}{\operatorname{argmin}} \left\{ \| w - Y \|^2 \right\}$$

$(w-Y)^T (w-Y)$

Regression for (linear) probabilistic models

Same idea as before w/ randomized term, no normalization by $1/n$.

- Likelihood function (error from yesterday)

$$L(\theta) = \prod_{i=1}^{n} P_\theta(y_i | x_i)$$

- log-likelihood function $\ell(\theta) = \sum_{i=1}^{n} \log\left(P_\theta(y_i | x_i)\right)$

$\log(L(\theta)) = \ell(\theta)$, minimizing $\ell(\theta)$ & $L(\theta)$ equivalent since logarithm is non-decreasing.

Fix σ^2, $P_\theta(y_i | x_i) = \Phi\left(y | f_\theta(x), \sigma^2\right)$

Julia note $\Phi(x::\text{Number}, \theta) = \sigma * \text{randn}() + f(x, \theta)$

$\text{randn}() \sim N(0, 1)$

Theorem (3.12) When $r(\Omega) = D+1$, MLE is solved by $\theta_{ML} = \Omega^+ y$

pf $\frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-(f_\theta(x)-y)^2}{2\sigma^2}} = p_\theta(y \mid f_\theta(x), \sigma^2)$

$$\log p_\theta(y_i \mid f_\theta(x_i), \sigma^2) = \log\left(\frac{e^{\frac{-(f_\theta(x)-y)^2}{2\sigma^2}}}{\sqrt{2\pi}\,\sigma}\right)$$

$$= \log\left(e^{\frac{-(f_\theta(x)-y)^2}{2\sigma^2}}\right) - \log\left(\sqrt{2\pi}\,\sigma\right)$$

$$= \frac{-(f_\theta(x)-y)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$$

So log-likelihood function is

$$\ell(\theta) = \sum_{i=1}^{n} \log\left(p_\theta(y_i \mid f_\theta(x_i), \sigma^2)\right)$$

$$= -\sum_{i=1}^{n} \frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i^T\begin{pmatrix}\theta_1\\ \vdots \\ \theta_D\end{pmatrix} + \theta_0 - y_i\right)^2$$

From linear term

$$= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|\Omega\theta - y\|^2$$

independent of θ $\Rightarrow$ Unique solution follows from linear case.

Regularization → used to prevent overfitting.

ERM tends towards overfitting

Deterministic linear → Tikhonov regularization $\lambda \in \mathbb{R}$

$$\ell(y_i, f_\theta(x_i)) = (f_\theta(x_i) - y_i)^2 + \lambda \|\theta\|^2$$

$$\|\theta\|^2 = \theta_0^2 + \theta_1^2 + \cdots + \theta_d^2$$

Corresponding regression problem called ridge regression

* See Theorem 3.11

Probabilistic linear ~ MLE tends towards overfitting.

Instead of adding $\lambda\|\theta\|^2$, we put a distribution on θ called a prior.

posteriori function is then $\alpha(\theta) = P(\theta \mid X, Y)$

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^{n \times d} \qquad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$$

Want to view X, Y as random variables.

Bayes' Theorem $P(\theta \mid X, Y) = P(\theta) \dfrac{P(Y \mid X, \theta)}{P(Y \mid X)}$

Taking logarithm

constant w.r.t. θ

$$\log\left[P(\theta|X,Y)\right] = \log P(\theta) + \log P(Y|X,\theta) - \log(Y|X)$$

$\log P(\theta)$: distribution of θ = regularizing term

$\log P(Y|X,\theta)$: MLE case

★ Special case when $D=1$ (Thm 3.15 for general)

$$\theta = \begin{pmatrix}\theta_0\\ \theta_1\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \underbrace{\begin{bmatrix}\sigma & 0\\ 0 & \sigma\end{bmatrix}}_{\Sigma}\right)$$

$$\Rightarrow P(\theta) = \frac{1}{\sqrt{(2\pi)^2(\sigma^2)}} \exp\left(\frac{-1}{2\sigma^2}\|\theta\|^2\right)$$

Also assume as in MLE case w/ independent variables

$$P(Y|X,\theta) = \prod_{i=1}^{n} \Phi(y_i|f_\theta(x_i),\sigma^2)$$

$$= \frac{1}{\sqrt{(2\sigma^2)^n}} \exp\left(\frac{-1}{2\sigma^2}\|Y-X\theta\|^2\right)$$

θ_{MAP} maximizes $\log(P(\theta|X,Y))$

~ Maximize by taking derivative (gradient) w.r.t. θ.

$$\log\left[P(\theta|X,Y)\right] = \log P(\theta) + \log P(Y|X,\theta) - \log(Y|X)$$

$$= \frac{-1}{2\sigma^2}\|\theta\|^2 - \log\left(\sqrt{2\pi}^2\sigma^2\right)$$

$$-\frac{1}{2\sigma^2}\|Y-X\theta\|^2 - \log\left((\sqrt{2\sigma^2})^N\right)$$

$$+ \text{constant w.r.t. } \theta$$

Taking derivative, green goes away we have

$$\frac{d}{d\theta}\log(P[\theta|X,Y)) = \frac{-1}{\sigma^2}\theta - \frac{1}{\sigma^2}X^T(Y-X\theta)$$

$$\|\theta\|^2 = \theta_0^2 + \theta_1^2 + \cdots + \theta_0^2$$

$$\frac{d}{d\theta} = [2\theta_0, 2\theta_1, \cdots, 2\theta_2] = 2\theta$$

$$(y_1 - X\theta_1)^2$$

$$2\left[y_1 - X\theta_1\right] \cdot [1 \; x_1]$$

Set derivative = 0

$$\theta = -\Omega^T(\Omega\theta - Y)$$

$$(\Omega^T\Omega - Id)\theta = \Omega^T Y$$

$$\theta_{MAP} = (\Omega^T\Omega - Id_2)^{-1}\Omega^T Y$$

This is Theorem 3.15 in a special case $\Sigma = \sigma \cdot Id_2$ $\mu = \vec{0}$ (