Statistical estimation Problems

Maximum Likelihood estimation (MLE)

chance

$$\max_{\underline{\theta} \in \mathbb{R}^r} L(\underline{\theta}) = \max_{\underline{\theta} \in \mathbb{R}^r} \log p(\underline{y}, \underline{\theta})$$

$\underline{\theta}$: Parameter vector

$L(\underline{\theta})$: Log_Likelihood

$p(\underline{y}, \underline{\theta})$: probability density function (PDF)

$\underline{y}$: realization of some random vector

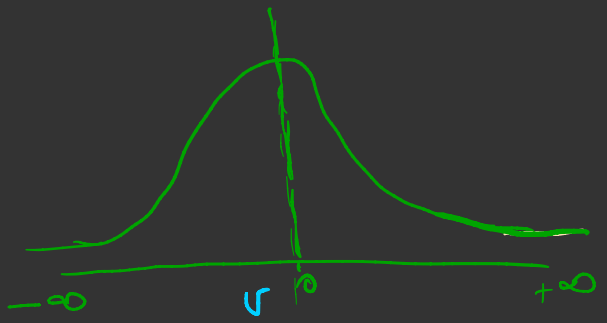
If p is a function of $\underline{\theta}$ (parameter vector) is log-concave, then the MLE problem is a convex optimation problem.

Example: linear measurement with i.i.d noise.

independent and identically distributed

$$y_i = \underbrace{a_i^T}_{\in \mathbb{R}^n} \underline{\theta} + \underbrace{v_i}_{\in \mathbb{R}}, \quad i = 1, \dots, m$$

v_i: measurement noise, random variable



We assum $v \in \mathbb{R}$
measurment noise (i.i.d)
has PDF on $\mathbb{R}$

$$y_i = \underline{a}_i^T \underline{\theta} + v_i \quad \Rightarrow \quad v_i = y_i - \underline{a}_i^T \underline{\theta}$$

independent

$$\Rightarrow \underline{\bar{p}}(\underline{y}, \underline{\theta}) = \prod_{i=1}^{m} p(y_i - \underline{a}_i^T \underline{\theta})$$

Joint PDF

$$\Rightarrow \log \tilde{p} = \sum_{i=1}^{m} \log p(y_i - a_i^T \theta)$$

MLE problem

$$\max_{\theta \in \mathbb{R}^n} \sum_{i=1}^{m} \log p(y_i - a_i^T \theta)$$

---

convex optimization problem, if p is log-concave in θ (parameter vector)

special case: Gaussian noise: $v_i \sim N(0, \sigma^2)$

$$\Rightarrow p(z) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{z^2}{2\sigma^2}\right)$$

$$l(\underline{\theta}) = -\left(\frac{m}{2}\right)\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|A\underline{\theta} - \underline{y}\|_2^2$$

A has rows $\underline{a}_1^T, \ldots, \underline{a}_m^T$

$$\theta_{ML}^{\star} = \arg\min \|A\underline{\theta} - \underline{y}\|_2^2$$

ordinary least square (OLS)

MLE is robust against Gaussian noise

Logistic regression :

Bernoulli a.k.a two class Logistic regression :

n expriment : $[\underline{X}_i, y_i]$, $y_i \in [0, 1]$

covid test : X_i : age, blood pressure, heart rate

n patients : y : $y_i = 0$ no, $y_i = 1$ yes

spam : X_i : capital letter + attentions

y_i : $y_i = 0$ no, $y_i = 1$ yes

$Y \in \{0, 1\}$ with probability mass function. $\mathbb{P}(Y=1) + \mathbb{P}(Y=0) = 1$

Bernoulli(p), $p \in (0,1)$, $\mathbb{P}(Y=1) = p$, $\mathbb{P}(Y=0) = 1-p$

Logistic function: $f_{\text{Logistic}}(z) = \dfrac{\exp(z)}{1+\exp(z)} \in (0,1)$, $z \in \mathbb{R}$

we want to model p as the composition of f_{Logistic}:

$$p := f_{\text{Logistic}}(\underline{\theta}^T \underline{x}) = \frac{\exp(\underline{\theta}^T \underline{x})}{1 + \exp(\underline{\theta}^T \underline{x})}$$

$\underline{x}$: Feature vector $\in \mathbb{R}^d$

$\underline{\theta} \in \mathbb{R}^d$ to be determined.

regressor parameter

N experiment

Reorder the data

$\underline{x}_1, \dots, \underline{x}_{N_1} \Rightarrow y = 1$, $\underline{x}_{N_1+1}, \dots, \underline{x}_N \Rightarrow y = 0$

$$\ell(\underline{\theta}) = \left(\prod_{i=1}^{N_1} p_i \right) \left(\prod_{N_1+1}^{N} (1 - p_i) \right)$$

the likelihood function

where $p_i = f_{\text{Logistic}}(\underline{\theta}, \underline{x}_i)$

$$\log p(\underline{\theta}) = \sum_{i=1}^{N_1} \log \frac{\exp(\underline{\theta}, \underline{x}_i)}{1+\exp(\underline{\theta}, \underline{x}_i)} + \sum_{i=N_1+1}^{N_2} \log \frac{1}{1+\exp(\underline{\theta}, \underline{x}_i)}$$

$\arg\max_{\underline{\theta} \in \mathbb{R}^d} \log L(\underline{\theta})$ $\Rightarrow$ This is an unconstrained convex optimization problem in variable $\underline{\theta} \in \mathbb{R}^d$

(convexity by Hessian of obj)

The development above can be generalized for the multi_trial case.

categorical a.k.a. a multiclass logistic regression.

$\{(\underline{x}_i, y_i)\}$, $y \in \underbrace{\{0, 1, \dots, K\}}_{K+1}$ the recorded labels are the realization of a categorical random variable

PMF $\text{cat}(p_1, \dots, p_K) \Rightarrow p_0 = 1 - \sum_{n=1}^{K} p_n$

$\mathbb{P}(y=1) = p_1, \dots, \mathbb{P}(y=K) = p_K \Rightarrow P(y=0) = 1 - (p_1 + \dots + p_K)$

Regressor parameter $\underline{\theta} := \{\underline{\theta}_1, \dots, \underline{\theta}_K\}^T \in \mathbb{R}^{d \times K}$

The regression model

$$p_1 = \frac{\exp(\underline{\theta}_1, \underline{x})}{1 + \sum_{r=1}^{K} \exp(\theta_i, \underline{x})}, \dots, \quad P(x=0) = 1 - (p_1 + \dots + p_K)$$

$$= \frac{1}{1 + \sum_{r=1}^{K} \exp(\underline{\theta}_r, \underline{x})}$$

$$L(\theta) = \left(\prod_{i=1}^{N_1} p_{1i}\right) \cdots \left(\prod_{i=N_{k-1}+1}^{N_k} p_{ki}\right) \left(\prod_{i=N_k+1}^{N} (1 - (p_{1i} + \cdots p_{ki}))\right)$$

$$p_{ri} = \frac{\exp(\underline{\theta}_r, \underline{x}_i)}{1 + \sum_{r=1}^{k} \exp(\underline{\theta}_r, \underline{x}_i)}$$

MLE

$\arg\max \log L(\underline{\theta})$

$\underline{\theta} \in \mathbb{R}^d$

$$\sum_{i=1}^{N_1} \underline{\theta}_1^T \underline{x}_i + \sum_{i=N_1+1}^{N_2} \underline{\theta}_2^T \underline{x}_i + \cdots + \sum_{i=1}^{N} \log\left(1 + \sum_{r=1}^{k} \exp(\underline{\theta}_r^T \underline{x}_i)\right)$$

A different way to do statistical inference is MAP (Maximum a posteriori probability estimation)

after the fact

$\underline{X}$ ← to be estimated

$\underline{Y}$ ← observed variable

} Joint PDF $f(\underline{X}, \underline{Y})$

In MLE, $\underline{X}$ was deterministic (non-random)

But in MAP, both $\underline{X}$ and $\underline{Y}$ are random

Prior of X

before we observe Y

$$p_X(\underline{x}) = \int p(\underline{x}, \underline{y})\, d\underline{y}$$

$$p_Y(\underline{y}) = \int p(\underline{x}, \underline{y})\, d\underline{x}$$

conditional:

$$p_{Y|X}(\underline{x}, \underline{y}) = \frac{p(\underline{x}, \underline{y})}{p_X(\underline{x})} \quad \Big\} \text{ Bayes' rule}$$

$$p(\underline{x}, \underline{y}) = p_{Y|X}(\underline{x}, \underline{y})\, p_X(\underline{x})$$

$$p_{X|Y} = \frac{p(\underline{x}, \underline{y})}{p_Y(\underline{y})} = \frac{p_{Y|X}(\underline{x}, \underline{y})\, p_X(\underline{x})}{p_Y(\underline{y})}$$