Lecture #15

Linear Regression

Max. Likelihood: $y_i = a_i^T \underline{x} + v_i, \quad i = 1, \ldots, m$

- y_i: measurements
- a_i: known
- $\underline{x}$: unknown (deterministic) parameter to be estimated
- v_i: iid $\in \mathbb{R}$ noise.

$$\Leftrightarrow \quad \underline{y}_{m\times 1} = \underset{m\times n}{A}\, \underline{x} + \underline{v}_{m\times 1}$$

likelihood f^n: $\displaystyle\prod_{i=1}^{m} p(y_i - a_i^T \underline{x})$

$$\Leftrightarrow \text{log-likelihood } \ell(\underline{x}) = \sum_{i=1}^{m} \log p(y_i - a_i^T \underline{x})$$

what if $v_i \sim \text{Laplace}(\alpha)$, $\alpha > 0$

$$p_v(z) = \frac{1}{2\alpha} \exp\left(-|z|/\alpha\right) \quad \leftarrow \text{Laplace PDF}$$

$$\Leftrightarrow \sum_{i=1}^{m} \log p(y_i - a_i^T \underline{x}) = -\sum_{i=1}^{m} \left(2\alpha + \frac{|y_i - a_i^T \underline{x}|}{\alpha}\right)$$

$$\therefore \underline{x}^*_{ML} = \arg\max \, \ell(\underline{x})$$

$$= \underset{\underline{x} \in \mathbb{R}^n}{\arg\min} \, \|\underline{y} - A\underline{x}\|_1 \quad \text{(LP)}$$

Many PDFs in practice are log-concave

e.g. multivariate Gaussian, Laplace, Exponential, Uniform over convex set, Beta, Gamma, χ^2 etc.

log-concavity is preserved under

1. product
2. Marginalization
3. Convolution

$$\int p(x, y)\, dx = q(y)$$

$$\int p(x, y)\, dy = r(x)$$

Conversely, we can give statistical interpretation of any convex optimization problem of the form:

$$\underline{x}^* = \underset{\underline{x} \in \mathbb{R}^n}{\arg\min} \sum_{i=1}^{m} \phi(b_i - a_i^T \underline{x})$$

This is an ML problem where

$v_i \rightarrow$ i.i.d. with PDF $\rho_v(z) \propto \exp(-\phi(z))$

$\Updownarrow$

$$\rho_v(z) = \frac{\exp(-\phi(z))}{\int \exp(-\phi(z))\, dz}$$

ML with Poisson's distribution:

$$Y \in \mathbb{Z}_+ = \{0, 1, 2, \dots\}$$ (Poisson distribution)

↑ random variable → distribution $\text{Poiss}(\mu)$

$$\mathbb{P}(Y = k) = \frac{\exp(-\mu)\, \mu^k}{k!}.$$

$$\mu = a^T x + b$$

a, b: parameters

x: explanatory variable

Given (x_i, y_i)

x_i: explanatory variable measured.

y_i: observed/realized measurement

log-likelihood : $l(\underline{a}, b)$

$$= \sum_{i=1}^{m} \left\{ y_i \log(\underline{a}^T \underline{u}_i + b) - (a^T u_i + b) - \log(y_i!) \right\}$$

$$(\underline{a}^*_{ML}, b^*_{ML}) = \underset{\substack{\underline{a} \in \mathbb{R}^n \\ b \in \mathbb{R}}}{\operatorname{argmax}} \left\{ y_i \log(\underline{a}^T u_i + b) - (\underline{a}^T \underline{u}_i + b) \right\}$$

Max. Likelihood estimate of the parameters $\underline{a}$ & b

convex f^n

ML estimate of covariance of some Gaussian random vector.

$$\underline{y} \in \mathbb{R}^n, \quad \underline{y} \sim \mathcal{N}(\underline{0}, R), \quad R \in S^n_{++},$$

$$\therefore \, p(\underline{y}) = \frac{1}{\sqrt{(2\pi)^n \det(R)}} \exp\left(-\frac{1}{2}\underline{y}^T R^{-1} \underline{y}\right)$$

We want to estimate the matrix $R \succ 0$ based on N samples $\underline{y}_1, \ldots, \underline{y}_N \in \mathbb{R}^n$ drawn from this distribution.

Log-likelihood :

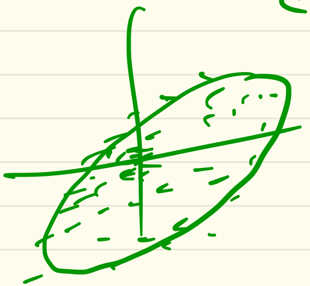
$$l(R) = \log p_R(\underline{y}_1, \ldots, \underline{y}_N)$$

$$\Rightarrow \ell(R) = -\left(\frac{Nn}{2}\right)\log(2\pi) - \left(\frac{N}{2}\right)\log\det(R) - \frac{1}{2}\sum_{k=1}^{N} \underline{y}_k^T R^{-1} \underline{y}_k$$

$$= -\frac{N}{2}\,\mathrm{tr}(R^{-1}Y)$$

where $Y := \frac{1}{N}\sum_{k=1}^{N} \underline{y}_k \underline{y}_k^T$ (sample covariance matrix)

$$\mathrm{covar}(X) = \mathbb{E}\left((X - \mathbb{E}(X))(X - \mathbb{E}(X))^T\right)$$



Can show:

$\ell(R)$ is concave in R (only on a subset of $\mathbb{S}^n_{++}$)

when $R \preceq 2Y$

Change-of-variable:

$$S := R^{-1} \succ 0$$

(Recall that matrix inversion preserves pos. definiteness)

S: Information matrix; R: cov. matrix

$$\therefore \ell(S) = -\left(\frac{Nn}{2}\right)\log(2\pi) + \left(\frac{N}{2}\right)\underbrace{\log\det(S)}_{\text{concave}} - \left(\frac{N}{2}\right)\underbrace{\mathrm{tr}(SY)}_{\text{concave}}$$

Concave in S

$$\therefore S^{*}_{ML} = \underset{S \in \mathbb{S}^n_{++}}{\arg\max}\ \log\det(S) - \mathrm{tr}(SY)$$

$$= \underbrace{\underset{S \succ 0}{\arg\min}\ -\log\det(S) + \mathrm{tr}(SY)}_{\text{convex optimization problem}}$$

∴ Analytical Sol^n :

$$S^*_{ML} = Y^{-1}$$

$$\Rightarrow (R^*_{ML})^{-1} = Y^{-1} \Leftrightarrow \boxed{R^*_{ML} = Y}$$

(ML estimate of ensemble covariance → R^*_{ML}; sample covariance → Y)

We can handle additional constraints on R or S

e.g. $L \preceq R \preceq U \Leftrightarrow U^{-1} \preceq S \preceq L^{-1}$

(Operator monotonicity, HW 3, p 2(e))

$$\Rightarrow \boxed{\begin{array}{l} \max\ \ell(S) \\ S \succ 0 \\ \&\ \ U^{-1} \preceq S \preceq L^{-1} \end{array}}$$ ← LMI

e.g.

$$\frac{\lambda_{max}(R)}{\lambda_{min}(R)} \le K_{max}$$

} condition number constraint on matrix R

$$\Leftrightarrow \exists\ u > 0 \text{ s.t. } uI \preceq S \preceq K_{max}\, u I$$

$$\therefore \; S^*_{ML} = \underset{S \succ 0,\, u \in \mathbb{R}}{\operatorname{argmax}} \, l(S)$$

$$\text{s.t.} \quad uI \preceq S \preceq k_{max} u I$$

} SDP

MAP (Maximum a posteriori Probability estimate)

$\underline{x} \leftarrow$ to be estimated
$\underline{y} \leftarrow$ observed variable

} Joint PDF $f(\underline{x}, \underline{y})$

(In ML estimation: $\underline{x}$ is deterministic, not a random variable)

Prior of $\underline{x}$ (before we observe $\underline{y}$):

$$P_X(\underline{x}) = \int P(\underline{x}, \underline{y}) \, d\underline{y}$$

$$P_Y(\underline{y}) = \int P(\underline{x}, \underline{y}) \, d\underline{x}$$

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

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

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

$$\hat{x}^*_{MAP} = \underset{\underline{x} \in \mathbb{R}^n}{\operatorname{argmax}}\; p_{X|Y}(\underline{x}, \underline{y})$$

$$= \underset{\underline{x} \in \mathbb{R}^n}{\operatorname{argmax}}\; \underline{p_{Y|X}(\underline{x}, \underline{y})\, p_X(\underline{x})}$$

$$= \underset{\underline{x} \in \mathbb{R}^n}{\operatorname{argmax}}\; \underbrace{p(\underline{x}, \underline{y})}_{\text{Joint}}$$

Taking $\log(\cdot)$

$$\hat{x}^*_{MAP} = \underset{\underline{x} \in \mathbb{R}^n}{\arg\max} \left\{ \underbrace{\log p_{Y|X}(\underline{x}, \underline{y})}_{\text{log-likelihood of conditional}} + \underbrace{\log p_X(\underline{x})}_{\text{penalizing unlikely } \underline{x}} \right\}$$

e.g. linear regression (with MAP) with i.i.d. noise

$$y_i = \underline{a}_i^T \underline{x} + v_i, \quad i = 1, \dots, m$$

(v_i i.i.d. with PDF p_v on $\mathbb{R}$)

This time $\underline{x} \sim p_X(\underline{x})$ on $\mathbb{R}^n$ (Prior PDF)

$$\therefore \text{Joint } p(\underline{x}, \underline{y}) = p_X(\underline{x}) \prod_{i=1}^{m} p_v(y_i - a_i^T x)$$

$$\therefore \ \underline{\hat{x}}^*_{MAP} = \arg\max_{\underline{x} \in \mathbb{R}^n} \left\{ \log \rho_X(\underline{x}) + \sum_{i=1}^{m} \log \rho_V(y_i - a_i^T \underline{x}) \right\}$$

extra term compared to ML

e.g.

If $V_i \sim \text{Uniform}([-a, a])$

& $\underline{x} \sim \mathcal{N}(\underline{\mu}, \Sigma)$

then this reduces to

$$\underline{\hat{x}}^*_{MAP} = \arg\min_{\underline{x} \in \mathbb{R}^n} (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu})$$

$$\text{s.t.} \ \| A\underline{x} - \underline{y} \|_\infty \leq a$$

} QP

Non-parametric Estimation of Prob. Distribution over a finite set $\{\alpha_1, \dots, \alpha_n\} \subseteq \mathbb{R}$

Probability (discrete) distribution $\underline{p} \in \underbrace{S}_{\text{Probability simplex}}$

$$S = \{ \underline{p} \in \mathbb{R}^n_{\geq 0} \mid \mathbb{1}^T \underline{p} = 1 \}$$

Notice that given $f : \mathbb{R} \mapsto \mathbb{R}$, (any nonlinear function f)

$$\mathbb{E}[f(X)] = \underbrace{\sum_{i=1}^{n} p_i f(\alpha_i)}_{\text{linear in } \underline{p}}.$$

e.g. $\mathbb{P}(X \in \mathcal{C}) = \boxed{c^T p}$ where $c_i = \begin{cases} 1 & \text{if } x_i \in \mathcal{C} \\ 0 & \text{otherwise.} \end{cases}$

linear in p

e.g. $\mathbb{E}[X] = a$ (given) $\quad \mathbb{E}[X^2] = b$ (given)

$$\Updownarrow$$

$$\boxed{\sum_{i=1}^{n} x_i p_i = a} \qquad \Leftrightarrow \boxed{\sum_{i=1}^{n} x_i^2 p_i = b}$$

$\mathbb{P}(X \geq 0) \leq 0.3$

$$\Leftrightarrow \boxed{\sum_{x_i \geq 0} p_i \leq 0.3}$$

→ Linear in $\underline{p}$

e.g. Nonlinear in vector $\underline{p}$

$$\text{Var}(X) = c$$

$$\| $$

$$\mathbb{E}[X^2] - (\mathbb{E}X)^2$$

$$= \sum_{i=1}^{n} \alpha_i^2 p_i - \left(\sum_{i=1}^{n} \alpha_i p_i\right)^2$$

quadratic in p

$$\text{Var}(X) := \mathbb{E}\left[(X - \mathbb{E}(X))^2\right]$$

$$= \mathbb{E}\left[X^2 - 2X\,\mathbb{E}[X] + (\mathbb{E}[X])^2\right]$$

$$= \mathbb{E}[X^2] - 2(\mathbb{E}[X])^2 + (\mathbb{E}[X])^2$$

Max. Entropy Problem

$$\max_{p \in \mathcal{S}} \text{Entropy} \Rightarrow \max_{p \in \mathcal{S}} -\sum_{i=1}^{n} p_i \log p_i$$

← concave

$$\Updownarrow$$

$$\min_{p \in \mathcal{S}} \sum_{i=1}^{n} p_i \log p_i$$

convex.

s.t. $p \in \mathcal{C}$ ← some convex set

Min. KL. Divergence Problem:

Kullback-Leibler (KL) divergence

$$\min_{p \in \mathcal{S}} \sum_{i=1}^{n} p_i \log(p_i / q_i) = D_{KL}(p \| q)$$

s.t. $p \in \mathcal{C}$ ← convex set

If $q = \frac{1}{n}\mathbb{1}$ (uniform distribution)

then $D_{KL}(p \| q) = \sum_{i=1}^{n} p_i \log p_i$