Lecture 5 - statistical decision theory

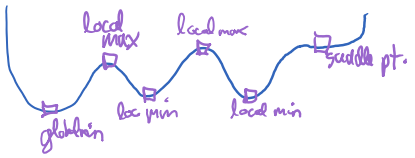
Tuesday, September 20, 2022 3:08 PM

today: • MLE ct'ed
• statistical decision theory

optimization comments about MLE

$\min_{\theta \in \Theta} f(\theta)$ ⊛ $\nabla f(\theta^*) = 0$ (if f is differentiable on Θ) is a necessary cond. for θ^* being a local min when θ^* is in the interior of Θ

"stationary pts."



→ also need to check that Hessian$(f)(\theta^*) \succ 0$ for a local min

$H \succ 0 \Leftrightarrow u^T H u > 0 \quad \forall u \neq 0 \in \mathbb{R}^d$

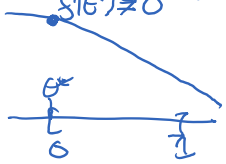
$(f''(\theta^*) > 0)$

⊛ only local results in general

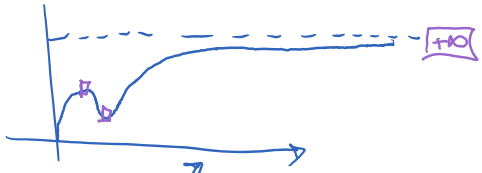
- but if Hessian$(f(\theta)) \succ 0 \ \forall \theta \in \Theta$, fct. is said to be "convex"
  then $\nabla f(\theta^*) = 0 \Rightarrow \theta^*$ is global min
- otherwise, for smooth fct., looking at zero gradient pts and boundary pts give you enough information to find global min

⊛ be careful with boundary cases

i.e. $\theta^* \in$ boundary (Θ) e.g.



other example :

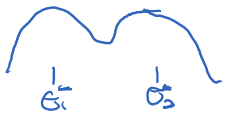


↑ example where MLE doesn't exist

* some notes about MLE

* does not always exist [$\theta^* \in$ bdd(Θ) but Θ is open] or when "$\theta^* = +\infty$" $\Theta =]0,1[$

• is not nec. unique [e.g. mixture models]



• is not "admissible" in general [see later]
∃ strictly "better" estimators

example II : multinomial distribution

suppose X_i is discrete R.V. on K choices "multinoulli"

(we could choose $\Omega_{X_i} = \{1, \dots, K\}$)

but instead, convenient to encode the k possibilities using unit basis in $\mathbb{R}^k$

ie $\Omega_{X_i} = \{e_1, \dots, e_k\}$ where $e_j \in \mathbb{R}^k$ "one hot encoding"

$e_j = (0, \dots, 0, \underbrace{1}_{j^{th}\text{ coordinate}}, 0, \dots, 0)$

parameter for discrete R.V.: $\pi \in \Delta_k$ $(\Theta = \Delta_k \subset \mathbb{R}^k)$

$\Delta_k \triangleq \{\pi \in \mathbb{R}^k : \pi_j \geq 0 \ \forall j ; \sum_{j=1}^k \pi_j = 1\}$

probability simplex on k choices

we will write $X_i \sim \text{Mult}(\pi)$ ← parameter

(*) consider $X_i \overset{iid}{\sim} \text{Mult}(\pi)$

then $X = \sum_{j=1}^n X_i \sim \text{Mult}(n, \pi)$

"multinomial distribution"

$X \in \mathbb{N}^k \quad \Omega_X = \{(n_1, \dots, n_k) : n_j \in \mathbb{N} ; \sum_{j=1}^k n_j = n\}$

pmf for X: $p(x|\pi) = \binom{n}{(x)_1, \dots, (x)_k} \prod_{j=1}^k \pi_j^{(x)_j}$ $\quad x = (n_1, \dots, n_k)$

multinomial coeff. $\binom{n}{n_1 \cdots n_k} \triangleq \frac{n!}{n_1! n_2! \cdots n_k!}$

15h58

multinomial MLE

log-likelihood $\ell(\pi) = \log p(x|\pi) = \underbrace{\log \binom{n}{n_1, \dots, n_k}}_{\text{constant} \to \text{ignore MLE}} + \sum_{j=1}^k n_j \log \pi_j$

$x = (n_1, \dots, n_k)$

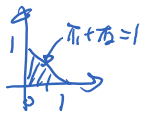
MLE: $\hat{\pi}_{MLE}(x) = \underset{\pi \in \mathbb{R}^k}{\text{argmax}} \ \ell(\pi)$ s.t. $\pi \in \Delta_k$ } constraint

two options:

a) reparameterize problem so that Θ is full dimensional

$\pi_k \triangleq 1 - \sum_{j=1}^{k-1} \pi_j$

→ $\pi_1, \dots, \pi_{k-1} \in [0,1]$ with constraint $\sum_{j=1}^{k-1} \pi_j \leq 1$



hope magic is that $\log \pi_j$ acts as a barrier fct. away from $\pi_j = 0$

can try unconstrained optimization on $\pi_1, \dots, \pi_k$

of $\ell(\pi_1, \dots, \pi_{k-1})$

hoping soln is in the interior of constraint set (and it usually will)

b) use Lagrange multiplier approach to handle equality constraints on Δ_k

b) use Lagrange multiplier approach to handle equality constraints on Δ_K

[and still ignore $\pi_j \in [0,1]$]

$\max f(\pi)$ s.t. $g(\pi)=0$

$[\,1-\sum_{j=1}^{K}\pi_j=0\,]$ $\triangleq g(\pi)$

$J(\pi,\lambda) = f(\pi) + \lambda g(\pi)$

λ: Lagrange multiplier

method: look at stationary pts of $J(\pi,\lambda)$ (0-gradient)

i.e. $\nabla_\pi J(\pi,\lambda)=0$

necessary cond. for local opt.

$\nabla_\lambda J(\pi,\lambda)=0 \Leftrightarrow g(\pi)=0$

(check "bordered Hessian" to get local min or max)

$\ell(\pi) = \sum_j n_j \log \pi_j$

(strictly concave fct. in π_j)

$\frac{\partial J}{\partial \pi_j} \overset{\text{want}}{=} 0$

$\frac{n_j}{\pi_j} - \lambda \overset{\text{want}}{=} 0 \Rightarrow \pi_j^* = \frac{n_j}{\lambda}$ scaling constant

(see Wikipedia)

want $g(\pi^*)=0$ i.e. $\sum_j \pi_j^* = 1$

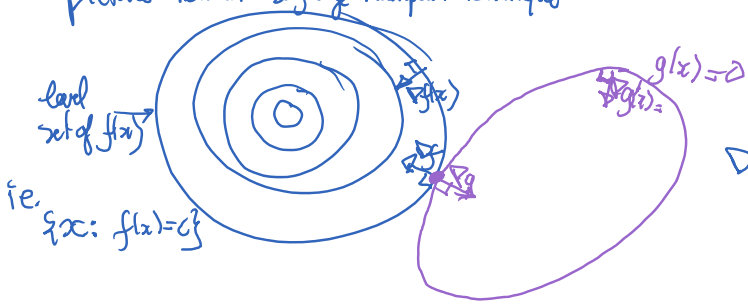
$\sum_j \frac{n_j}{\lambda} = 1$

$\Rightarrow \lambda^* = \sum_j n_j = n$

notice: $\pi_j^* = \frac{n_j}{n} \in [0,1]$

$\boxed{\pi_j^* = \frac{n_j}{n}}$ MLE for multinomial

picture behind Lagrange multiplier technique



$\nabla f(x^*) = \lambda \nabla g(x^*)$

i.e. ∇f & ∇g are parallel

Statistical decision theory

A) Bias-variance decomposition for squared loss

estimator: fct. from data (observations) to parameter

MLE: $\hat{\theta}_{MLE}(x) = \underset{\theta\in\Theta}{\operatorname{argmax}}\ p(x|\theta)$

MAP: $\hat{\theta}_{MAP}(x) = \underset{\theta\in\Theta}{\operatorname{argmax}}\ p(\theta|x) = \underset{\theta}{\operatorname{argmax}}\ p(x|\theta)\cdot p(\theta)$ (likelihood, prior)

* how do we evaluate these estimators?

estimator $S: \Omega \to \Theta$

$\hat{\theta} = S(X)$

most standard tool: frequentist risk of an estimator

$R(\theta, S) \triangleq \mathbb{E}_x[L(\theta, S(X))]$

(statistical) loss fct.

average over possible data

squared loss: $L(\theta,\theta') \triangleq \|\theta-\theta'\|_2^2$ $\quad \hat{\theta}=\delta(x)$

$$\mathbb{E}_x[\|\theta-\hat{\theta}\|_2^2] = \mathbb{E}[\|\theta - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \hat{\theta}\|_2^2]$$

$$= \mathbb{E}[\|\theta-\mathbb{E}[\hat{\theta}]\|^2] + \mathbb{E}[\|\hat{\theta}-\mathbb{E}[\hat{\theta}]\|_2^2] + 2\,\mathbb{E}[\langle \theta-\mathbb{E}[\hat{\theta}], \mathbb{E}[\hat{\theta}]-\hat{\theta}\rangle]$$

(by linearity; $\theta-\mathbb{E}[\hat{\theta}]$ is constant)

$$2\langle \theta-\mathbb{E}[\hat{\theta}], \mathbb{E}[\hat{\theta}]-\mathbb{E}[\hat{\theta}]\rangle = 0$$

$$R(\theta,\delta) = \mathbb{E}_x[\|\theta-\hat{\theta}\|_2^2] = \|\theta-\mathbb{E}[\hat{\theta}]\|_2^2 + \mathbb{E}[\|\hat{\theta}-\mathbb{E}[\hat{\theta}]\|^2]$$

bias^2 variance

bias $\triangleq \|\theta-\mathbb{E}[\hat{\theta}]\|$

(Sq.) risk for squared loss = bias^2 + variance (bias-variance decomposition "tradeoff")

* consistency is informally "do right thing as $n\to\infty$" where n is training set size

$X \rightsquigarrow (X_i)_{i=1}^n$

$\hat{\theta}_n$ $\hat{\theta}_n$ (data of size n)

assignment: if $\text{bias}(\hat{\theta}_n)\xrightarrow{n\to\infty}0$ and $\text{variance}(\hat{\theta}_n)\to 0$ $\Rightarrow R(\theta,\hat{\theta}_n)\xrightarrow{n\to\infty}0$ $\Rightarrow \hat{\theta}_n$ is consistent $(\hat{\theta}_n \xrightarrow{P} \theta)$

Statistical decision theory — formal setup

- a random observation $D \sim P$ (unknown distribution which models the world/phenomenon)

(often P_θ)

- action space $\mathcal{A}$
- loss (statistical) $L(P,a)$ = statistical loss of doing action $a\in\mathcal{A}$ when the world is P } describe the goal/task
  - often write $L(\theta,a)$ if we have a parametric model of world i.e. P is def/param by $\theta\in\Theta$
- $\delta : \mathcal{D} \to \mathcal{A}$ "decision rule" ($\mathcal{D}$: Ω_D)

examples: a) parameter estimation:

$\mathcal{A}=\Theta$ for parametric family P_θ

δ is a parameter estimator from data $D \overset{\text{typically}}{=} (X_1,\dots,X_n)$

typical loss $L(\theta,a)=\|\theta-a\|_2^2$ [usually $X_i \overset{iid}{\sim} P_\theta$, θ unknown]

"squared loss"

but other losses are $KL(p_\theta \| p_{\theta'})$

b) $\mathcal{A} = \{0,1\}$; this is hypothesis testing

δ describes a statistical test

loss → usually 0-1 loss $L(\theta, a) = \mathbb{1}\{\theta \neq a\}$

c) prediction in ML: (function estimation) learn a prediction fct. in supervised learning

here $D = (X_i, Y_i)_{i=1}^n$ $X_i \in \mathcal{X}$ (input space) $Y_i \in \mathcal{Y}$ (output space) $\mathcal{Y} = \{0,1\}$ → classification, $\mathcal{Y} = \mathbb{R}$ → regression

p_θ gives joint on (X, Y)

$D \sim P$ where $P = p_\theta \otimes p_\theta \otimes \cdots p_\theta$ n times

$\mathcal{A} = \mathcal{Y}^{\mathcal{X}}$ (set of fct.'s from $\mathcal{X}$ to $\mathcal{Y}$)

in ML $L(p_\theta, f) \triangleq \mathbb{E}_{p_\theta}[\ell(Y, f(X))]$ ("generalization error" / "classificat. error")

$(X,Y) \sim p_\theta$; ℓ = prediction loss

in ML is often called the "risk"

Vapnik calls it "Vapnik risk" to distinguish it from frequentist risk

eg classification $\ell(Y, f(X)) = \mathbb{1}\{Y \neq f(X)\}$ 0-1 error

frequentist risk $\mathbb{E}_p[L(p_\theta, \delta(D))]$

* decision rule $\hat{f} = \delta(D)$ ("learning algorithm")

$\hat{f}$: prediction fct. / classifier / etc.