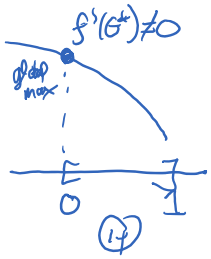
Lecture 5 - statistical decision theory

September 17, 2021 2:57 PM

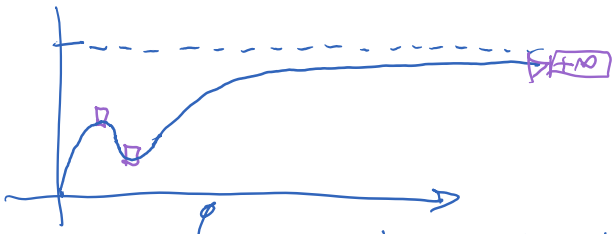
today: • MLE cf'ed
• decision theory

(*) be careful with boundary cases

i.e. $\theta^* \in \text{boundary}(\Theta)$



other example



example where the MLE "does not exist"

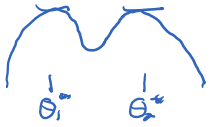
(*) Some notes about MLE

- does not always exist [$\theta^* \in \text{bd}(\Theta)$ but Θ is open] or when "$\theta^* = \infty$"

$\Theta =]0,1[$

- is not rec. unique [ie. multiple maxes]

e.g. mixture models



- is not "admissible" in general [see later]

$\exists$ strictly "better" estimators

example II: multinomial distribution

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

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

but instead, convenient to encode with unit basis vectors in $\mathbb{R}^K$ "one hot encoding"

ie $\Omega_{X_i} = \{e_1, \dots, e_K\}$ where $e_j \in \mathbb{R}^K$

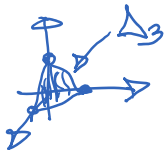
$e_j = (0, \dots, 0, \underbrace{1}_{j^{th} \text{ position}}, 0, \dots, 0)$

$(X = \sum_{i=1}^{n} X_i)$ multinomial

parameter for discrete R.V.s: $\pi \in \Delta_K$ $(\Theta = \Delta_K)$

$\Delta_K \triangleq \{\pi \in \mathbb{R}^K : \pi_j \geq 0 \ \forall j, \sum_{j=1}^{K} \pi_j = 1\}$

probability simplex on K choices



we will write $X_i \sim \text{Mult}(\pi)$ "multinoulli"

π: parameter

K x_i

parameter

pmf: $p(X_i = e_j) = \pi_j = \prod_{l=1}^{K} \pi_l^{x_{il}}$

"$X_i = j$"

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

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

"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:

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

$p(x|\pi) = \text{coeff.} \prod_{j=1}^{K} \pi_j^{(x)_j}$

$\binom{n}{(x)_1 (x)_2 \cdots (x)_K} \quad \binom{n}{n_1 n_2 \cdots n_K} \triangleq \frac{n!}{n_1! n_2! \cdots n_K!}$

multinomial coeff.

multinomial MLE:

log-likelihood: $\ell(\pi) = \log p(x|\pi) = \underbrace{\log \binom{n}{n_1 \cdots n_K}}_{\text{constant} \Rightarrow \text{ignore for MLE}} + \sum_{j=1}^{K} n_j \log \pi_j$

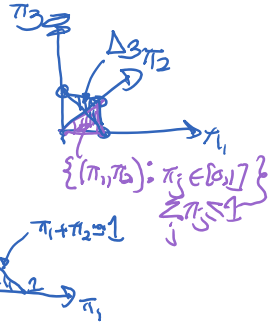
MLE: $\hat{\pi}_{ML}(x) = \underset{\substack{\pi \in \mathbb{R}^K \\ \text{s.t. } \pi \in \Delta_K}}{\text{argmax}} \ell(\pi)$ } constraint

two options:

a) reparameterize problem so that (H) is full dimensional

$\pi_1, \dots, \pi_{K-1} \in [0,1]$

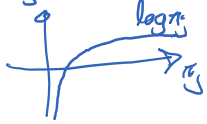
$\pi_K \triangleq 1 - \sum_{j=1}^{K-1} \pi_j$ with constraint $\sum_{j=1}^{K-1} \pi_j \leq 1$



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

can try unconstrained opt. on $\pi_1, \dots, \pi_{K-1}$ of $\ell(\pi_1, \dots, \pi_{K-1})$

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



[note: here $\ell(\pi)$ is a concave fct.]

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. $g(\pi) = 0$

$\sum_j \pi_j = 1 \Rightarrow g(\pi) \triangleq 1 - \sum_j \pi_j$

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

Lagrange multiplier

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

* necessary for local opt.

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

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

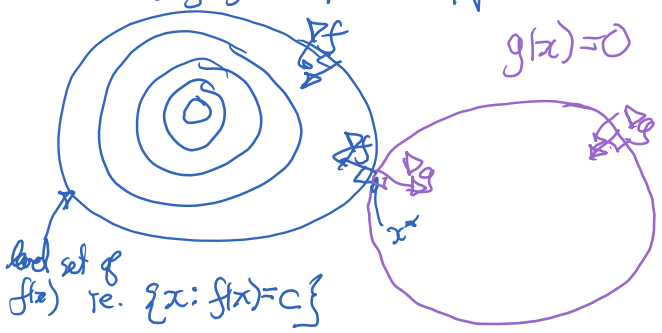
$\ell(\pi)=\sum_{j=1}^{K} n_j \log \pi_j$ (strictly concave fct. of π)

want $\frac{\partial J}{\partial \pi_j}=0 \Rightarrow \frac{n_j}{\pi_j}-\lambda=0 \Rightarrow \pi_j^*=\frac{n_j}{\lambda}$] scaling constant

want $g(\pi^*)=0$ i.e. $\sum_j \pi_j^*=1 \Rightarrow \sum_j \frac{n_j}{\lambda}=1$

notice: $\pi_j^*=\frac{n_j}{n} \in [0,1]$ $\boxed{\pi_j^*=\frac{n_j}{n}}$ MLE for multinomial $\Rightarrow \lambda^*=\sum_j n_j=n$

picture behind Lagrange multiplier technique



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

i.e. ∇f & ∇g are parallel

16h10

Statistical decision theory

A) Bias-variance decomposition for squared loss

estimator: function from data (observation) to parameters

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

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

(*) how do we evaluate these estimators? estimator $\delta:\Omega\to\Theta$, $\hat{\theta}=\delta(X)$

most standard tool: ~~frequentist~~ risk of estimator

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

(statistical) loss fct.; average over possible datasets

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

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

$$= \mathbb{E}\left[\|\theta - \mathbb{E}[\hat\theta]\|^2\right] + \mathbb{E}\left[\|\hat\theta - \mathbb{E}[\hat\theta]\|^2\right] + \underbrace{2\,\mathbb{E}\left[\langle \theta - \mathbb{E}[\hat\theta],\ \mathbb{E}[\hat\theta] - \hat\theta\rangle\right]}_{\text{constant}}$$

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

$$R(\theta, \delta) = \mathbb{E}_X\left[\|\theta - \hat\theta\|_2^2\right] = \underbrace{\|\theta - \mathbb{E}[\hat\theta]\|_2^2}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[\|\hat\theta - \mathbb{E}[\hat\theta]\|^2\right]}_{\text{variance}}$$

where $X \sim p_\theta$

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

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

* consistency: informally "do right thing as $n \to \infty$"

where n is training set size

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

$\hat\theta_n$ $\hat\theta$ (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 ($\hat\theta_n \xrightarrow{P} \theta$)

formal setup for statistical decision theory

- a random observation $D \sim P$ — unknown distribution which models the world / phenomena (often related to p_θ)
- action space $\mathcal{A}$
- (statistical) loss $L(P, a)$ = statistical loss of doing action $a \in \mathcal{A}$ when the world is P } describe the goal / task
- $\delta : \Omega_D \to \mathcal{A}$ "decision rule"

often write $L(\theta, a)$ if we have a parametric model of world ie. P has pmf/pdf p_θ for some $\theta \in \Theta$

examples: a) parameter estimation

$\mathcal{A} = \Theta$ for a parametric family $\mathcal{P}_\Theta$

δ is a parameter estimator from data

typically $D = (X_1, \dots, X_n)$

[usually, $X_i \overset{iid}{\sim} p_\theta$, θ unknown]

typical loss $L(\theta, a) = \|\theta - a\|_2^2$, $a \in \Theta$ "squared loss"

but other losses are used eg. $KL(p_\theta \| p_a)$

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

& describes a statistical test

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

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

have $D = (X_i, Y_i)_{i=1}^n$ $X_i \in \mathcal{X}$ (input space) $Y_i \in \mathcal{Y}$ (output space)

$\mathcal{Y} = \{0,1\}$ → binary classification

$\mathcal{Y} = \mathbb{R}$ → regression

P_θ gives joint (X,Y)

$D \sim P$ where $P = P_\theta \otimes P_\theta \cdots \otimes P_\theta = P_\theta^{\otimes n}$ (n times)

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

in machine learning $L(P_\theta, f) \triangleq \mathbb{E}_{(X,Y) \sim P_\theta}[\ell(Y, f(X))]$ ("generalization error" / "classification error")

prediction loss

in ML is often called "risk"

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

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

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

"learning algorithm"

prediction fct. / classifier etc...

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