Lecture 5 - scribbles

Tuesday, September 17, 2019 14:29

today : • finish MLE for multinomial
• statistical decision theory + properties of estimator

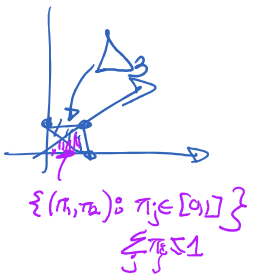
multinomial MLE :

$X \sim \text{Mult}(n, \pi) \quad \pi \in \Delta_K$

log-likelihood : $\ell(\pi) = \log p(x|\pi) = \underbrace{\log \binom{n}{n_1, \dots, n_K}}_{\text{const. with respect to } \pi \rightarrow \text{ignore for MLE}} + \sum_{j=1}^{K} n_j \log \pi_j$

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

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



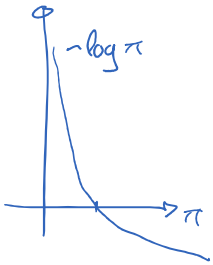
two options :

a) reparameterize problem to be full dimensional

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

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

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



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

can try unconstrained opt. on $\pi_1, \dots, \pi_{K-1}$
hoping solution is in interior of constraint set

b) use Lagrange multiplier approach to handle equality constraint on Δ_K
[and still not worry about $\pi_j \in [0,1]$]

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

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

method : look at stationary pts, (0-gradient) of

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

Lagrange multipliers

necessary but not sufficient conditions for (π^*, λ^*) to be optimal

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

$\ell(\pi) = \sum_j n_j \log \pi_j$
(concave fct. in π_j)

$\frac{\partial J}{\partial \pi_j} = 0 \Rightarrow \frac{n_j}{\pi_j} - \lambda = 0$

$\Rightarrow \pi_j^* = \frac{n_j}{\lambda}$] scaling constant

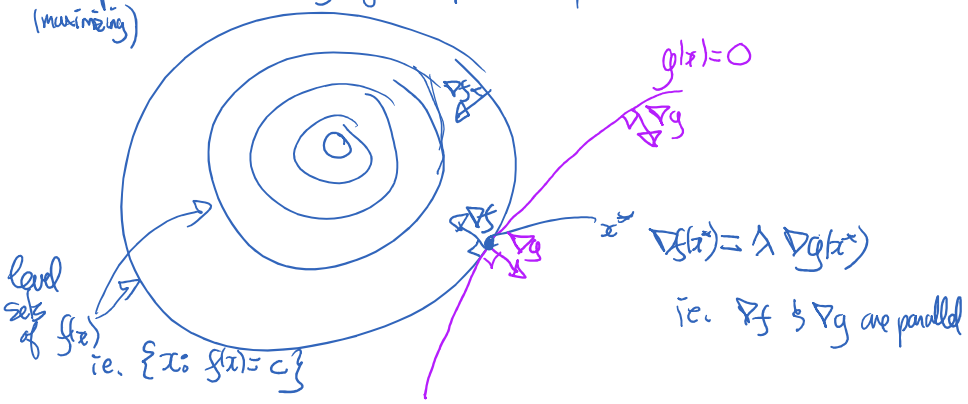
want $g(\pi^*)=0$ i.e. $\sum_j \pi_j^* = 1$

$\Rightarrow \frac{1}{\lambda}\sum_j n_j = 1 \Rightarrow \lambda^* = \sum_j n_j = n$

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

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

picture behind Lagrange multiplier technique:



i.e. ∇f & ∇g are parallel

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

* how do we evaluate these estimators?

estimator $S: \Omega \to \Theta$ $\quad \hat{\theta} = S(X)$

most standard tool: frequentist risk of an estimator

$$\mathbb{E}_X[L(\theta, S(x))] = R(\theta, S)$$

(Statistical) loss function

average over data

squared loss: $L(\theta,\theta') \triangleq \|\theta-\theta'\|_2^2$

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

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

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

$\|a+b\|^2 = \langle a+b, a+b\rangle = \|a\|^2 + \|b\|^2 + 2\langle a,b\rangle$

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

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

($\theta - \mathbb{E}[\hat{\theta}]$: constant)

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

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

$$\text{bias} \triangleq \|\theta-\mathbb{E}_X[\hat{\theta}]\|_2$$

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 $\text{variance}(\hat{\theta}_n) \xrightarrow{n\to\infty} 0$ $\Rightarrow R(\theta,\hat{\theta}_n) \xrightarrow{n\to\infty} 0 \Rightarrow \hat{\theta}_n$ is consistent

$[\hat{\theta}_n \xrightarrow{P} \theta]$

15h42

Statistical decision theory

formal setup:

- a random observation $D \sim P$ (unknown distribution which models the world) (often P_θ)
- action space $\mathcal{A}$
- loss $L(P,a)$ = loss of doing action $a\in\mathcal{A}$ when "the world" is P } describe the goal/task

(if have a parametric model of world, often write $L(\theta,a)$ where θ is s.t. $P=P_\theta$)

- $\delta: \mathcal{D} \to \mathcal{A}$ "decision rule"

$\mathcal{D}$

examples: a) parameter estimation:

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

δ is 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 used, $KL(P_\theta \| P_a)$

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

δ describes a statistical test

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

c) prediction in machine learning: learn a prediction function in supervised learning (function estimation)

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 $P = P_\theta^{\otimes n} = \underbrace{P_\theta \otimes P_\theta \cdots \otimes P_\theta}_{n \text{ times}}$

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

$|\mathcal{X}|^{|\mathcal{Y}|}$

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

$(X,Y) \sim P_\theta$ prediction loss "classification error"

* decision rule $\hat{f} = \delta(D)$ (training data)
  - prediction fct. / classifier / etc.
  - "learning algorithm"

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

in ML, is often called the "risk"

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

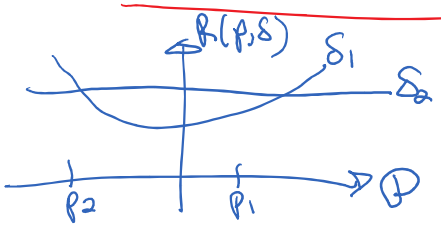
~~frequentist risk~~ $\mathbb{E}_D[L(P_\theta, \delta(D))]$

comparing procedures?

δ_1 vs. δ_2

property R: $\boxed{\text{(frequentist) risk } R(P, \delta) \triangleq \mathbb{E}_{D \sim P}[L(P, \delta(D))]}$

"risk profiles"



$\delta : \mathcal{D} \to \mathcal{A}$

incomparable functions? $R(\cdot, \delta_1)$ vs. $R(\cdot, \delta_2)$

* transform to scalar:
  - "minimax" analysis: $\max_{P \in \mathcal{P}} R(P, \delta)$ "worst case"
  - weighted average $\int_\Theta R(\theta, \delta)\pi(\theta)\, d\theta$