# Lecture 4 - scribbles

Tuesday, September 13, 2016
14:27

today
- MLE
- statistical decision theory
  bias/variance decomposition

## Maximum likelihood principle

parametric family $p(x;\theta)$ for $\theta \in \Theta$

how to estimate $\theta$?

choose $\hat{\theta}_{ML}(x)$ which maximizes $p(x;\theta)$

$$\hat{\theta} \in \underset{\theta \in \Theta}{\operatorname{argmax}}\ p(x;\theta) \triangleq g(\theta)$$

likelihood function (of $\theta$)

example: $n$ coin flips

we had $X = \sum_{i=1}^{n} X_i$ ($X_i$ Bernoulli(θ))  $X \sim Bin(n,\theta)$

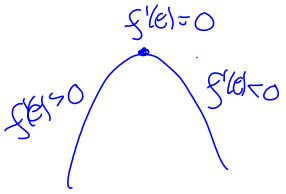
$$p(x;\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \qquad [x \in 0:n]$$

trick: $\log(\cdot)$ is strictly increasing fct.

ie. $a < b \Rightarrow \log(a) < \log(b)$

$$\Rightarrow \underset{\theta \in \Theta}{\operatorname{argmax}} \log(p(x;\theta)) = \underset{\theta \in \Theta}{\operatorname{argmax}}\ p(x;\theta)$$

log-likelihood: $\log p(x;\theta) = \underbrace{\text{cst.}}_{\text{constant (doesn't depend on } \theta)} + x\log\theta + (n-x)\log(1-\theta) = \ell(\theta)$




look for $\frac{\partial \ell(\theta)}{\partial \theta} = 0$

want $\frac{\partial \ell(\theta)}{\partial \theta} = \frac{x}{\theta} - \frac{(n-x)}{1-\theta} = 0$

$\nabla g(\theta) = 0$

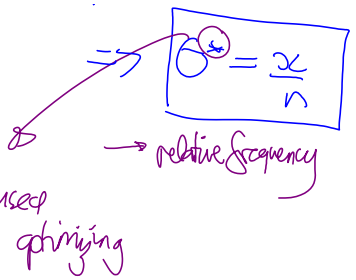
necessary condition if max is in the interior of $\Theta$

$\Rightarrow x - \theta x - n\theta + \theta x = 0$

$\Rightarrow \hat{\theta} = \frac{x}{n}$ ← estimator

$\hat{\theta}_{ML}$

necessary condition
if max is in the interior of $\Theta$



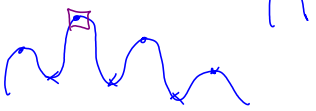

here $\hat{\theta}_{ML}(x) = \hat{\theta}^*(x) = \frac{x}{n}$

note: $\mathbb{E}[\hat{\theta}_{ML}] = \mathbb{E}[\frac{X}{n}] = \frac{n\theta}{n} = \theta$ → unbiased (here)

(with respect to $X \sim p(x; \theta_0)$)

Some optimization comments:

- $f'(\theta) = 0$ not sufficient for local max → also need to check $f''(\theta) < 0$ for max



- only local result in general

but if $f''(\theta) < 0 \quad \forall \theta \in \Theta$ function is "concave"

→ $f'(\theta) = 0$ is sufficient for global max

- careful with boundary cases $\theta^* \in \text{boundary}(\Theta)$

Some notes on MLE

- does not always exist [e.g. $\hat{\theta} \in \text{bd}(\Theta)$ but $\Theta$ is open]
- is not necessarily unique
- is not admissible in general (see end of class)

Example 2:

suppose $X_i$ is discrete R.V. on $k$ choices

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

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

i.e. $\Omega_{X_i} = \{e_1, \dots, e_k\}$

where $e_i \in \mathbb{R}^k \quad e_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$ ← $i^{th}$ position

parameter $\pi \in \Delta_k$ ← probability simplex on $k$ choices

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

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

[really $\text{Mult}(1, \pi)$]

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

then $X = \sum_{i=1}^n X_i \sim \text{Mult}(n, \pi)$ (analog of binomial but for $K$ choices)

$$\Omega_X = \{(n_1, \ldots, n_K) : \sum_k n_k = n, \ n_k \in \mathbb{N}\}$$

consider MLE for R. vector $\vec{X} = (X_1, \ldots, X_n)$

$p(\vec{x} \mid \pi)$ $\quad p(x_i \mid \pi) = \text{Mult}(x_i \mid \pi)$

$= \prod_{j=1}^K \pi_j^{x_{ij}}$ ← $j^{th}$ component of vector $x_i$

"switch starred"

[e.g. if $x_i = e_j$, then $p(x_i \mid \pi) = \pi_j$]

$$p(\vec{x} \mid \pi) = \prod_{i=1}^n p(x_i \mid \pi) = \prod_{i=1}^n \left( \prod_{j=1}^K \pi_j^{x_{ij}} \right)$$

by indep.

$$= \prod_{j=1}^K \pi_j^{\left(\sum_{i=1}^n x_{ij}\right)} \quad \text{note } n_j(\vec{x})$$

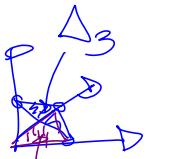
$n_j$

log-likelihood: $\ell(\pi) = \log p(\vec{x} \mid \pi) = \sum_{j=1}^K n_j \log \pi_j$

we want $\max_{\pi \in \Delta_K} \ell(\pi)$ } constraints

could reparameterize with $\pi_1, \ldots, \pi_{K-1} \in [0,1]$

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




here instead demonstrate simple Lagrange multiplier approach for equality constraint

look for stationary points (0 gradient) of

$$J(\pi, \lambda) \overset{e}{=} \ell(\pi) + \lambda \left(1 - \sum_{j=1}^K \pi_j\right)$$

$\lambda$ ← Lagrange

equality constraint $a(\pi) = 0$

multiplier

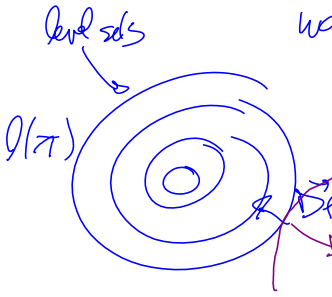
want $\nabla_\pi J(\pi, \lambda) = 0$

and

$\nabla_\lambda J(\pi, \lambda) = 0 \longrightarrow$ equivalent to each $g(\pi) = 0$

$\partial_{\pi_j} J = 0 \Rightarrow \frac{n_j}{\pi_j} - \lambda = 0 \quad \forall j$

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




want $\sum_j \pi_j^* = 1 \Rightarrow \lambda = \sum_j n_j = n$

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

$-\lambda \nabla g = \nabla \ell \Rightarrow \nabla \ell + \lambda \nabla g = 0$

## Statistical decision theory (frequentist)

general setup:

- $D \sim P$ ← unknown distribution which model the "world"
  - $D$ random data
- $A$ action space
- $L(P, a)$ ← loss of doing action $a$ when "true world" is $P$ } describes goal/task

[if have parametric model; often write $L(\theta, a)$ where $P = P_\theta$]

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

Examples: a) $A = \Theta$ $\delta$ is then parameter estimator from data

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

b) $A = \{0, 1\}$ ← hypothesis test

c) say $D = ((x_i, y_i)_{i=1}^n)$ $x_i \in X$, $y_i \in Y$

If $P_0$ gives joint on $(x_i, y_i)$; then $P = P_0^{\otimes n}$

$A = Y^X$ (set of fcts from $X \to Y$)

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

in M.L., $L(P, f) = \mathbb{E}_{P_0}[\ell(Y, f(X))]$

- $\ell$: prediction loss
- $(X,Y) \sim P_0$
- "generalization error"

decision rule: $f = \delta(D)$

- $f$: prediction fct.
- $D$: training dataset
- $\delta$: "learning algorithm"

How to compare procedures? (given $\delta$)

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

- $D \sim P$: observation is random

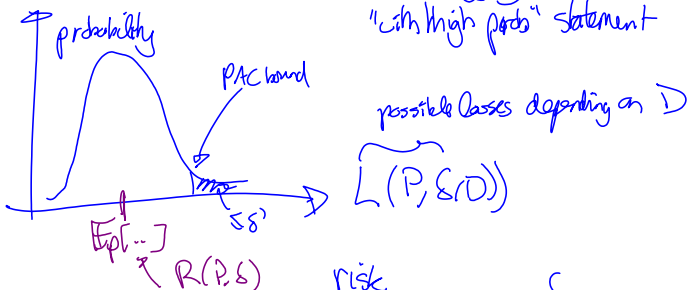
"how well does it do in average"?

vs. PAC analysis (learning theory) → look at tail bound small

$P\{L(P, \delta(D)) > \text{stuff}\} \leq \delta'$

"with high prob" statement

cartoon:




⊛ now $R(P, \delta)$

depends P/θ unknown