Lecture 5 - scribbles

Tuesday, September 18, 2018 14:40

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

multinomial (cted from last class)

$$X_i \sim \text{Mult}(\pi)$$

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

pmf for X:

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

$$p((x_1, \dots, x_n)|\pi)$$

$$= \prod_{i=1}^{n} p(x_i|\pi) = \prod_{j=1}^{n} \left(\prod_{j=1}^{k} \pi_j^{x_{ij}} \right)$$

$$= \prod_{j=1}^{k} \pi_j^{\sum_{i=1}^{n} x_{ij}} \qquad = \prod_{j=1}^{k} \pi_j^{n_j}$$

$\sum_{i=1}^{n} x_{ij} = n_j = (x)_j$

$$p(x_i|\pi) = \text{Mult}(x_i|\pi) = \prod_{j=1}^{k} \pi_j^{x_{ij}}$$

x_{ij} ← j^{th} component of vector x_i

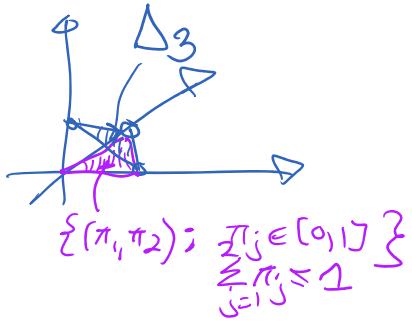
log-likelihood: $\ell(\pi) = \log p(x|\pi) = \underbrace{\log \binom{n}{n_1, \dots, n_k}}_{\text{const. with respect to } \pi} + \sum_{j=1}^{k} n_j \log \pi_j$ (can ignore for MLE)

we want: $\max_{\pi \in \mathbb{R}^K} \ell(\pi)$ s.t. $\pi \in \Delta_K$ } constraints

two options: a) would reparameterize $\pi_K \triangleq 1 - \sum_{j=1}^{K-1} \pi_j$

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

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



and then do unconstrained optimization 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 ← necessary but not sufficient conditions

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

$\sum_{j=1}^{K} \pi_j = 1$

$1 - \sum_{j=1}^{K} \pi_j = 0$

$\triangleq g(\pi)$

look for stationary points (0-gradient) of

$J(\pi, \lambda) \triangleq f(\pi) + \lambda \underbrace{g(\pi)}_{1 - \sum_{j=1}^{K} \pi_j}$

λ: Lagrange multiplier

ie. want $\nabla_\pi J(\pi, \lambda) = 0$
and
$\nabla_\lambda J(\pi, \lambda) = 0 \rightarrow$ equivalent to $g(\pi) = 0$

$\ell(\pi) = \sum_{i=1}^{K} n_i \log \pi_i$ $\quad \partial_i J = 0 \Rightarrow n_i - \lambda = 0$

$\ell(\pi) = \sum_{j=1}^{K} n_j \log \pi_j$

concave in π

$\frac{\partial^2 \ell}{\partial \pi_i \pi_j} = \delta_{ij} \left(\frac{-1}{\pi_j^2}\right)$

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

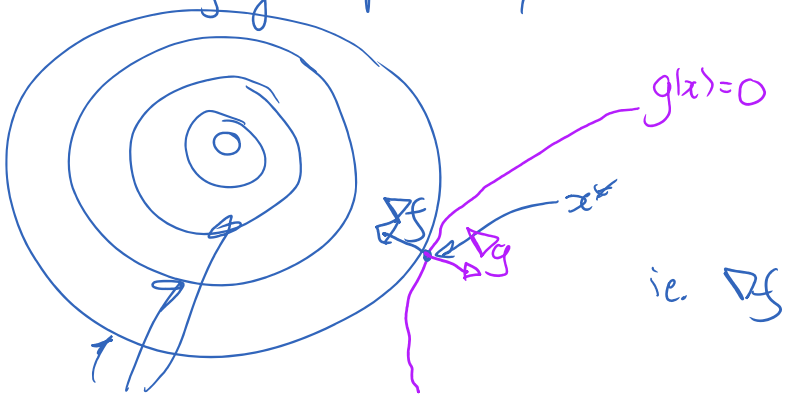
$\Rightarrow \pi_j^* = \frac{n_j}{\lambda}$] scaling constant

want $\sum_j \pi_j^* = 1 \Rightarrow \frac{1}{\lambda} \sum_j n_j = 1 \Rightarrow \lambda^* = \sum_{j=1}^{K} n_j = n$

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

sidenote: $\pi_j^* = \frac{n_j}{n} > 0$

picture behind Lagrange multiplier technique:



level sets of $f(x)$ [ie. $\{x : f(x) = c\}$]

$a^T x \le b$

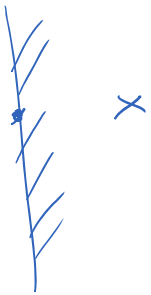
$J(x, \lambda) = f(x) + \lambda g(x)$

$\nabla_x J(x, \lambda) = 0$

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

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

i.e. ∇f & ∇g are parallel



Bias-variance decomposition for squared loss

estimator: function from data to parameters

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

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

* analyzing property of estimator:

frequentist risk of an estimator

estimator $\delta: \mathcal{X} \to \Theta$ $\quad \hat{\theta} = \delta(X)$

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

$\mathbb{E}_X$: an average over data; L: loss function

squared loss: $L(\theta, \delta(X)) = \|\theta - \delta(X)\|_2^2$

$\delta(X) = \hat{\theta}$ $\quad (\hat{\theta}(x))$

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

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

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

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

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

$\mathbb{E}[\hat{\theta}] - \mathbb{E}[\hat{\theta}] = 0$

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

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

risk for squared loss = bias^2 + variance

bias-variance decomposition "tradeoff"

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

here: training set of size n; $\hat{\theta}_n$ is $\hat{\theta}$ (data of size n)

assignment: $\text{bias}(\hat{\theta}_n) \xrightarrow{n\to\infty} 0$ & $\text{variance}(\hat{\theta}_n) \xrightarrow{n\to\infty} 0$ $\Rightarrow$ $\hat{\theta}_n$ is consistent $\left[\hat{\theta}_n \xrightarrow{?} \theta\right]$

(frequentist) statistical decision theory

, unknown distribution which models the "world"

unknown distribution which models the "world"

formal setup:
- a random observation $D \sim P$ (perhaps P_θ)
- action space $\mathcal{A}$
- loss $L(P, a)$ = loss of doing action $a \in \mathcal{A}$ when "the world" is P } describes 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"
  - $\mathcal{D}$: P, Ω_D

examples: a) estimation of parameters:

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

δ is then parameter estimator from data

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

[more specifically, often $D = (X_1, \dots, X_n)$ where $X_i \overset{iid}{\sim} P_\theta$ (θ is unknown)

$\delta(D) = \hat{\theta}$ $L(\theta, \delta(D)) = \|\theta - \hat{\theta}\|_2^2$]

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

here δ describes a statistical test

c) prediction in machine learning = learn a prediction function in supervised learning

here $D = ((x_i, y_i))_{i=1}^n$ $\quad x_i \in \mathcal{X}$ (input space), $y_i \in \mathcal{Y}$ (output space)

$\mathcal{Y} = \{0,1\}$ → classification
$\mathcal{Y} = \mathbb{R}$ → regression
...

P_e gives joint on (X,Y)

$D \sim P$ where $P = P_e^{\otimes n} = \underbrace{P_e \otimes P_e \cdots P_e}_{n \text{ times}}$

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

prediction loss

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

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

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

δ: "learning algorithm"

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

$(X,Y) \sim P_e$

in M.L., often called "risk"

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