Lecture 17 - scribbles

Tuesday, November 7, 2017
14:26

today: · exponential family
· sampling

Exponential family

a (flat/canonical) exponential family on $\mathcal{X}$

is a parametric family of distributions defined by two quantities

I) $h(x)\,d\mu(x)$ → reference measure on $\mathcal{X}$

reference density, base measure — counting (discrete R.V.), Lebesgue (cts. R.V.)

II) $T: \mathcal{X} \to \mathbb{R}^p$ called "sufficient statistics" vector (aka feature vector)

members of family will have dist.:

$$p(x;\eta)\,d\mu(x) = \exp\left(\eta^T T(x) - A(\eta)\right) h(x)\,d\mu(x)$$

$T(x)$: defining pieces (+ Ω_X)

η: "canonical parameter"

$A(\eta)$: log normalizer, log partition function, or cumulant generating function

if Ω_X is discrete, then $p(x;\eta)$ is a pmf

" " cts., " " " a pdf

* want $1 = \int_{\mathcal{X}} p(x;\eta)\,d\mu(x) = \int_{\mathcal{X}} \exp(\eta^T T(x))\, e^{-A(\eta)}\, h(x)\,d\mu(x)$

$$\Rightarrow \boxed{A(\eta) \triangleq \log \underbrace{\left(\int_{\mathcal{X}} \exp(\eta^T T(x)) h(x) \, d\mu(x) \right)}_{Z(\eta)}}$$

domain $\Omega \triangleq \{ \eta \in \mathbb{R}^p \mid A(\eta) < \infty \}$
(set of valid canonical parameters)

(*) more generally, consider a reparameterization of a subset of the family by defining the mapping $\eta : \Theta \to \Omega$ (new set of parameters)

consider $p(x; \theta) \triangleq p(x; \eta(\theta))$ for $\theta \in \Theta$

(get a "curved exponential family" if $\eta(\Theta)$ is a curved manifold in Ω)

* two examples not an exponential family:
  - mixture of Gaussians
  - Unif$(0, \theta)$

but most standard distributions are (Poisson, Gamma, Gaussian, etc...)

Example: (Multinoulli)

$X \sim \text{Mult}(\pi)$ $\quad \mathcal{X} = \{0,1\}^K$

$\Omega_X = \Delta_K \cap \mathcal{X}$

parameter space $\pi \in \Delta_K$ $\quad$ suppose $\pi_i > 0 \ \forall i$

$$p(x;\pi) = \prod_{j=1}^{K} \pi_j^{x_j} \overset{b}{=} \exp\left(\sum_{j=1}^{K} x_j \log \pi_j\right)$$

think of this as a "θ"

$$= \exp\left(\eta(\pi)^T x - 0\right)$$

we have $\eta_j(\pi) = \log \pi_j$

$T(x) = x$

$d\mu(x) =$ counting measure on $\mathcal{X}$

$h(x) = \mathbb{1}\{x \in \Omega_X\} = \mathbb{1}\{x$ has exactly one entry equal to $1\}$

$\Theta = \text{int}(\Delta_K)$

here, $A(\eta(\pi)) = 0 \quad \forall \pi \in \Theta$

but here $\Omega = \mathbb{R}^K$ remark: $\Theta \to$ dimension $K-1$

$\eta(\Theta) \to$ " $K-1$

$\Omega \to$ dimension K

(Ω domain, not be confused with Ω_X)

note:

for any x s.t. $h(x) \neq 0$

$$\sum_{j=1}^{K} T_j(x) = 1$$

affine linear dep. between components of T

$\Rightarrow$ multiple η's map to same distribution "over parameterization"

$\hookrightarrow$ not a minimal exponential family

* for multinoulli, minimal exp family

$$T(x) = \begin{pmatrix} x_1 \\ \vdots \\ x_{K-1} \end{pmatrix}$$

$$Z(\eta) = \sum_{x \in \Omega_x} \exp(\eta^T T(x)) = \sum_{j=1}^{K-1} e^{\eta_j} + 1$$

$$p(x;\eta) = \exp\Big(\sum_{j=1}^{K-1} \eta_j x_j - \underbrace{\log\Big(\sum_{j=1}^{K-1} \exp(\eta_j) + 1 \Big)}_{A(\eta)} \Big)$$

recall: $\nabla_\eta A(\eta) = \mathbb{E}_{p(x;\eta)}[T(x)]$ (valid for $\eta \in \text{int}(\Omega)$)

for multinoulli, $\frac{\partial A(\eta)}{\partial \eta_j} = \frac{1}{Z(\eta)} \exp(\eta_j) = p(x = j \mid \eta) = \mathbb{E}_{p(x;\eta)}[\mathbb{1}_{j}(x)]$ as required

(*) $T(x)$ is called "sufficient" as in statistics:

$T: x \mapsto T(x)$ is "sufficient" for a parametric family p_θ

iff $\forall \theta \in \Theta \quad p_\theta(x) = \underbrace{h(x)}_{\text{fixed for family}} g(T(x); \theta)$

i.e. dependence on θ only happens through $T(x)$

i.i.d. model in exp. family:

$$\log p(x_1, \dots, x_n \mid \eta) = \sum_{i=1}^{n} [\log h(x_i) + \eta^T T(x_i) - A(\eta)]$$

$\rightsquigarrow$ exp. family on $(x_1, \dots, x_n)$ with reference density $\tilde{h}(x) = \prod_i h(x_i)$

sufficient statistics $T(x) = \sum_{i=1}^{n} T(x_i)$

log-partition $A(\eta) = nA(\eta)$

Example 2: (1D) Gaussian

$X \sim N(\mu, \sigma^2)$ $\quad \mathcal{X} = \mathbb{R}$

$$p(x; (\mu, \sigma^2)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$= \exp\left(-\frac{x^2}{2}\left[\frac{1}{\sigma^2}\right] + x\left(\frac{\mu}{\sigma^2}\right) - \left[\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right]\right)$$

$$T(x) = \begin{bmatrix} x \\ -\frac{x^2}{2} \end{bmatrix} \quad \eta(\theta) = \begin{bmatrix} \mu/\sigma^2 \\ 1/\sigma^2 \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}$$

$\eta_2 = \frac{1}{\sigma^2}$ = precision > 0

$\eta_1 = \eta_2 \cdot \mu$ $\qquad \Omega = \{(\eta_1, \eta_2) : \eta_2 > 0\}$

Example 3: UGM?

Let $p \in \mathcal{F}(G)$ G is undirected

with $\psi_c(x_c) > 0$

$$p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(x_c) = \exp\left(\sum_c \ln \psi_c(x_c) - \log Z\right)$$

$$= \exp\left(\sum_{c \in \mathcal{C}} \sum_{y_c \in \mathcal{X}_c} \underbrace{\mathbb{1}\{y_c = x_c\}}_{T_{c,y_c}(x)} \underbrace{\log \psi_c(x_c)}_{\eta_{c,x_c}} - \log Z\right)$$

for discrete value of x & y.

for every value of C & y_c

I have $T_{c,y_c}(x) = \mathbb{1}\{y_c = x_c\}$

$$T(x) = \begin{pmatrix} \mathbb{1}\{x_c = y_c\} \end{pmatrix}_{\substack{y_c \in \mathcal{X}_c \\ c \in \mathcal{C}}} \qquad \eta_{c,y_c} = \log \psi_c(y_c)$$

$$\eta(\theta) = \begin{pmatrix} \log \psi_c(y_c) \end{pmatrix}_{\cdots}$$

notes: a) Mult(x) is special case with complete graph (1 big clique)

b) feature perspective: instead of using all indicators $\mathbb{1}\{x_c = y_c\}$

you could choose a subset

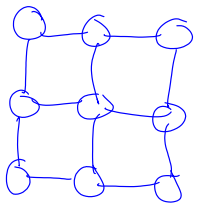
for example; suppose x_i is a word

feature on x_i & x_{i+1} $\mathbb{1}\{x_i \text{ is a verb } \& \ x_{i+1} \text{ is a noun}\}$

c) binary Ising model

$x_i \in \{0,1\}$ $|C| = 2$

suppose use nodes & edges as cliques



$\Rightarrow$ dimension of $T(x)$ is $2|V| + 4|E|$ $\rightarrow$ "overparameterized"

a minimal representation:

$$T(x) = \begin{pmatrix} (x_i)_{i \in V} \\ (x_i x_j)_{\{i,j\} \in E} \end{pmatrix} \Rightarrow \dim = |V| + |E|$$

ie. $$p(x;\eta) = \frac{1}{Z} \exp\left(\sum_{\{i,j\} \in E} \eta_{ij} x_i x_j + \sum_{i \in V} \eta_i x_i \right)$$

def: $p(x;\eta)$ is <u>minimal</u> exp family iff there does not exist $\{a_j\}$

s.t. $\sum_{j=1}^{p} a_j T_j(x) + a_{p+1} = 0 \quad \forall x$ s.t. $h(x) \neq 0$

<u>Properties of A:</u>

- $A(\cdot)$ is C^∞ for $\eta \in \text{int}(\Omega)$
- $A(\cdot)$ is <u>convex</u> (Ω is a convex set)
- $\nabla_\eta A(\eta) = \mathbb{E}_{p(x;\eta)}[T(x)] \triangleq \mu(\eta)$ "moment vector"
  for $\eta \in \text{int}(\Omega)$
- $\nabla^2 A(\eta) = \mathbb{E}_{p(x;\eta)}\left[(T(x) - \mu(\eta))(T(x) - \mu(\eta))^T \right] = \text{cov}(T(x))$
  (Hessian)
  (proof as exercise)
- $Z(\eta)$ is the mgf for $p(x;\eta)$ $\quad t \mapsto Z(\eta + t)$ for $t \in \mathbb{R}^p$
  ↳ $\mathbb{E} \exp(t^T T(x))$

$A(\eta)$ = log mgf → cumulant generating function

→ why derivatives of A give cumulants

1st derivative give mean
2nd " " covariance
3rd " " 3rd-cumulant etc...

Approximate inference → sampling

why sampling?

$X = (X_1, \dots, X_p)$

a) simulation : $X^{(i)} \sim p$

b) approximate $p(x_j)$

→ more generally,

consider $f : \mathbb{R}^p \to \mathbb{R}^d$, approximate $\mu = \mathbb{E}_p[f(X)]$

e.g. if $f(X) = \mathbb{1}\{X_A = x_A\}$ $\mathbb{E}_p[f(X)] = p(X_A = x_A)$

Monte-Carlo integration / estimation → appears in physics, applied math, ML, etc...

to approximate $\mu = \mathbb{E}_p[f(X)]$

alg: • n samples $X^{(i)} \overset{iid}{\sim} p$

• estimate : $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} f(X^{(i)})$

this is true even if $X^{(i)}$ are dependent

properties: 1) unbiased $\mathbb{E}[\hat{\mu}_n] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} f(X^{(i)}) = \mu$

2) expected error: $\mathbb{E}[\|\hat{\mu}-\mu\|_2^2] = \mathbb{E}\left[\frac{1}{n^2}\sum_{i,j}\langle f(x^{(i)})-\mu, f(x^{(j)})-\mu\rangle\right]$

$\mathrm{tr}(\mathrm{cov}(\hat{\mu},\hat{\mu}))$

$= \frac{1}{n^2}\sum_{i=j} \mathbb{E}[\langle f(x^{(i)})-\mu, f(x^{(i)})-\mu\rangle]$ ← by independence; ← off-diagonal term are zero

$\mathbb{E}[\|f(x)-\mu\|^2]$

$\triangleq \sigma^2 = \mathrm{tr}(\mathrm{cov}(f(x), f(x)))$

$$\boxed{\mathbb{E}[\|\hat{\mu}-\mu\|^2] = \frac{\sigma^2}{n}}$$

[aside: asymptotic analysis LLN $\hat{\mu} \xrightarrow{as} \mu$

CLT $\sqrt{n}(\hat{\mu}-\mu) \xrightarrow{d} N(0,\sigma^2)$