Lecture 19 - exponential families

November 12, 2021 2:53 PM

today: - exp. family
- estimation DGM/UGM
- sampling

moment matching different than MLE in exp family

gamma dist. $\Gamma(\alpha,\beta)$ has $T(x)=\begin{bmatrix}\log x\\ x\end{bmatrix}$

so moment matching with $\tilde{T}(x)=\begin{bmatrix}x\\ x^2\end{bmatrix}$ will yield different estimate than MLE

example 2: 1d Gaussian

$X \sim N(\mu,\sigma^2)$ $\mathcal{X}=\mathbb{R}$ $\Theta=(\mu,\sigma^2)$ "moment parameterization"

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

$$= \exp\left(\underbrace{-\frac{x^2}{2}\left[\frac{1}{\sigma^2}\right]}_{\eta_2} + x\underbrace{\left[\frac{\mu}{\sigma^2}\right]}_{\eta_1} - \underbrace{\left[\frac{\mu^2}{2\sigma^2}+\frac{1}{2}\log(2\pi\sigma^2)\right]}_{A(\eta)}\right)$$

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

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

$\eta_1=\eta_2\cdot\mu$

$h(x)=1$ (but some people use $h(x)=\frac{1}{\sqrt{2\pi}}$ for Gaussian)

$\Omega=\{(\eta_1,\eta_2): \eta_2>0, \eta_1\in\mathbb{R}\}$

[we'll see later: multivariate Gaussian $T(x)=\begin{bmatrix}x\\ -\frac{xx^T}{2}\end{bmatrix}$ $\begin{cases}\Lambda=\Sigma^{-1}\\ \eta=\Lambda\mu=\Sigma^{-1}\mu\end{cases}$]

example 3: discrete UGM

Let $p\in\mathcal{L}(G)$, G is undirected with $\psi_C(x_C)>0$ $\forall C, x_C$

$$p(x)=\frac{1}{Z}\prod_{C\in\mathcal{C}}\psi_C(x_C)=\exp\left(\sum_C\log\psi_C(x_C)-\log Z\right)$$

$$=\exp\left(\sum_{C\in\mathcal{C}}\sum_{y_C\in\mathcal{X}_C}\underbrace{\mathbb{1}\{y_C=x_C\}}_{T_{C,y_C}(x)}\underbrace{\log\psi_C(y_C)}_{\eta_{C,y_C}}-\log Z\right)$$

$$T(x)=\left(\begin{array}{c}\vdots\\ \mathbb{1}\{x_C=y_C\}\\ \vdots\end{array}\right)\begin{array}{l}\leftarrow y_C\in\mathcal{X}_C\\ C\in\mathcal{C}\end{array} \qquad \mathcal{X}_C=\{(y_i)_{i\in C}: y_i\in\mathcal{X}_i\}$$

$$\eta(\theta)=\left(\begin{array}{c}\vdots\\ \log\psi_C(y_C)\\ \vdots\end{array}\right)_{y_C\in\mathcal{X}_C}$$ [not a minimal representation]

$$\eta(\theta) = \begin{pmatrix} \vdots \\ \log \psi_C(y_C) \\ \vdots \end{pmatrix}_{\substack{y_C \in \mathcal{X}_C \\ C \in \mathcal{C}}}$$ [not a minimal representation]

notes: a) Mult(π) is a special case where has complete graph (1 big clique)

b) feature perspective: instead of using all possible indicators $\mathbb{1}\{y_C = x_C\}$ you could use a subset for a task

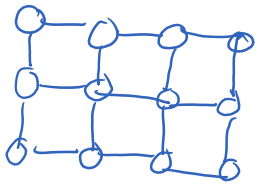
for example: suppose x is a sentence, x_i is a word

feature on x_i, x_{i+1} e.g. $\mathbb{1}\{x_i$ is a verb, x_{i+1} is a noun$\}$

↝ much smaller set of parameters

c) binary Ising model

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



suppose use nodes & pairs (edges) as cliques

⇒ dimension of $T(x)$ is $2|V| + 4|E|$ ↝ "overparameterized" exp family

$\sum_{y_C} T_{C,y_C}(x) = 1$ for any C

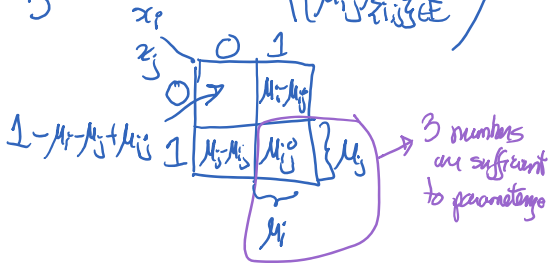
⇒ not a min. exp family

* a minimal representation

is. $T(x) = \begin{pmatrix} (x_i)_{i \in V} \\ (x_i x_j)_{x_i, x_j \in E} \end{pmatrix}$ → dim: $|V| + |E|$

$x_i x_j \to \mathbb{1}\{x_i = 1, x_j = 1\}$

$\mathbb{E}T(x) = \begin{pmatrix} (\mu_i)_{i \in V} \\ (\mu_{ij})_{x_i, x_j \in E} \end{pmatrix}$



properties of A:

- $\nabla_\eta A(\eta) = \mathbb{E}_{p(x;\eta)}[T(x)] \triangleq \mu(\eta)$ "moment vector" (for $\eta \in \text{int}(\Omega)$)
- $\left(\frac{\partial^2 A(\eta)}{\partial \eta_i \partial \eta_j}\right)_{ij} = \mathbb{E}_{p(x;\eta)}[(T(x) - \mu(\eta))(T(x) - \mu(\eta))^T] = \text{Cov}(T(x))$

(proof as exercise)

"cumulant generating fct."

16h02

cumulant generating fct.

16h02

Estimation of parameters DGM/UGM

DGM:

(fully observed)

parametric family $P_{\mathcal{H}} = \{ p_\theta(x) = \prod_i p(x_i \mid x_{\pi_i}, \theta_i) : \theta = (\theta_1, \dots, \theta_{|V|}) \in \mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2 \times \dots \times \mathcal{H}_{|V|} \}$

independent parameterization i.e. no tying parameters

$\Rightarrow$ MLE decouples in $|V|$ independent MLE problems

$\{x^{(i)}\}_{i=1}^n$ $\quad p(\text{data} \mid \theta) = \prod_{i=1}^n p(x^{(i)} \mid \theta) = \prod_{i=1}^n \prod_{j=1}^{|V|} p(x_j^{(i)} \mid x_{\pi_j}^{(i)}, \theta_j)$

$\log [\ ] = \sum_{j=1}^{|V|} \underbrace{\left(\sum_{i=1}^n \log p(x_j^{(i)} \mid x_{\pi_j}^{(i)}, \theta_j) \right)}_{f_j(\theta_j)}$

example: for discrete R.V. (multinoulli conditionals) $\Rightarrow \theta_j^{MLE}$ = proportion of observations

$\hat{p}(x_j = k \mid x_{\pi_j} = \text{stuff}) = \dfrac{\#(x_j = k, x_{\pi_j} = \text{stuff})}{\#(x_{\pi_j} = \text{stuff})}$

(fully observed DGM is relatively easy)

⊛ if have latent variable (i.e. unobserved variables) $\Rightarrow$ use EM. (like HMM)

UGM:

example for expo family

$$p(x \mid \eta) = \exp\left(\sum_c \eta_c T_c(x_c) - A(\eta) \right)$$

→ unlike in a DGM, $\log p(x;\eta)$ does not separate as $\sum_c f_c(\eta_c)$

gradient ascent on log-likelihood

$$\frac{1}{n} \sum_{i=1}^n \log p(x^{(i)} \mid \eta) = \sum_c \eta_c^T \left(\frac{1}{n} \sum_{i=1}^n T_c(x_c^{(i)}) \right) - A(\eta)$$

$\hat{\mu}_c$

$$\nabla_{\eta_c} [\ ] = \hat{\mu}_c - \mu_c(\eta)$$

μ_c

$\mu_c(\eta) = \mathbb{E}_{p(x;\eta)}[T_c(x_c)]$

to compute this, need inference

e.g. Ising model $T_{ij}(x_i, x_j) = x_i \cdot x_j \quad (x_i \in \{\pm 1\})$

$\mathbb{E}[T_{ij}] = \mu_{ij} = p(x_i = 1, x_j = 1 \mid \eta)$

have need approx. inference — sampling (e.g. Gibbs sampling)

here need approx. inference
- sampling (eg [Gibbs sampling])
- variational method (eg [mean field])

Approximate inference – Sampling

example: NP hard to do exact inference in Ising model → need approximation

why sampling? $X = (X_1, \dots, X_p)$

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

b) approximate marginal $p(x_i)$

→ special case of expectations

consider: $f: \mathbb{R}^p \to \mathbb{R}$

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

special case: if $f(x) \triangleq \mathbb{1}\{X_A = x_A\}$ $\mathbb{E}_p[f(x)] = p(X_A = x_A)$

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

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

MC estimation alg:
- n samples $x^{(i)} \overset{iid}{\sim} p$
- estimate $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(x^{(i)}) = \mathbb{E}_{\hat{p}_n}[f(x)]$

emp. dist. $\hat{p}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x^{(i)} = x\}$

properties:

1) unbiased estimator: $\mathbb{E}[\hat{\mu}] = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbb{E}f(x^{(i)})}_{\mu} = \frac{n\mu}{n} = \mu$

($\mathbb{E}$: expectation over $(x^{(i)})_{i=1}^n$)

this is still true if $x^{(i)}$ are dependent

2) expected error (l_2-error) $\mathbb{E}[\|\hat{\mu} - \mu\|_2^2] = \mathbb{E}\left[\left\langle \frac{1}{n}\sum_i f(x^{(i)}) - \mu, \frac{1}{n}\sum_j f(x^{(j)}) - \mu \right\rangle\right]$

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

by independence ⇒ off-diagonal terms are zero

i.e. $\langle \mathbb{E}f(x^{(i)}) - \mu, \mathbb{E}f(x^{(j)}) - \mu \rangle$ ($\mu - \mu \to 0$)

$= \frac{1}{n^2}\sum_{i=j} \mathbb{E}[\langle f(x^{(i)}) - \mu, f(x^{(i)}) - \mu \rangle]$

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

$= \text{tr}(\text{cov}(f(x), f(x)))$

$= \frac{n\sigma^2}{n^2}$

$= \frac{1}{n^2}\mathrm{tr}(\mathrm{cov}(f(x), f(x)))$

$= \frac{p\sigma^2}{n^2}$

$$\mathbb{E}[\|\hat{\mu} - \mu\|_2^2] = \frac{\sigma^2}{n}$$

note: that no dimension in rate (apart σ^2 which could depend implicitly)

e.g. $f(x) = X$, $X_j \sim N(0, \sigma^2)$

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