Lecture 18 - scribbles

Friday, November 15, 2019 13:57

today: • fusion properties of exp. family
• estimation of graph model parameters
• sampling (approximate inference)

Exp. family (ct.ed)

example 2: 1D Gaussian

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

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

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

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

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

$\eta_1 = \eta_2 \cdot \mu$

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

[we'll see later: multivariate Gaussian $\Lambda = \Sigma^{-1}$
$\eta = \Lambda\mu = \Sigma^{-1}\mu$]

$$T(x) = \begin{bmatrix} x \\ -\frac{xx^T}{2} \end{bmatrix}$$

Example 3: discrete UGM?

let $p \in \mathcal{L}(G)$, G is undirected

with $\psi_C(x_C) > 0 \quad \forall C, x_C$

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

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

$T_{C,y_C}(x)$ n_{C,y_C}

$$T(x) = \begin{pmatrix} \vdots \\ \mathbb{1}\{x_C = y_C\} \\ \vdots \end{pmatrix} \leftarrow \begin{matrix} y_C \in \mathcal{X}_C \\ C \in \mathcal{C} \end{matrix} \qquad \mathcal{X}_C = \{ (y_i)_{i \in C} : \text{s.t. } y \in \mathcal{X} \} = \bigtimes_{i \in C} \mathcal{X}_i$$

$$\eta(\theta) = \begin{pmatrix} \vdots \\ \log \mu_C(y_C) \\ \vdots \end{pmatrix}$$

notes: a) $\text{Mult}(\pi)$ is special case where complete graph (1 big clique)

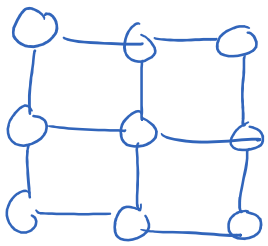
b) feature perspective: instead of using all indicators $\mathbb{1}\{x_C = y_C\}$, you could choose relevant subset for a task

for example x is a sentence and x_i is a word

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

c) binary Ising model

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



suppose use nodes & pairs (edges) as cliques

$\Rightarrow$ dimension of $T(x)$ is $2|V| + 4|E|$ $\rightsquigarrow$ "overparameterized" exp. family

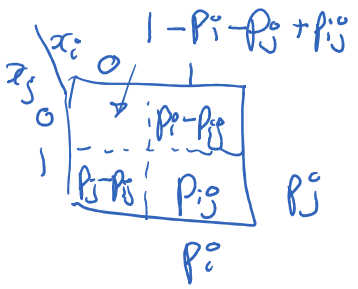
$\sum_{y_C} T_{C,y_C}(x) = 1$ for any C

$\Rightarrow$ not a min. exp. family

* a minimal representation:

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

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



properties of A:

Properties of A:

- $A(\cdot)$ is C^∞ for $\eta \in \text{int}(\Omega)$

  $A(\cdot)$ is convex (Ω is convex set)

- $\nabla_\eta A(\eta) = \mathbb{E}_{p(x|\eta)}[T(x)] \triangleq \mu(\eta)$ "moment vector" (for $\eta \in \text{int}(\Omega)$)

- $\left(\frac{\partial^2}{\partial \eta_i \partial \eta_j} A(\eta)\right)_{ij} = \mathbb{E}_{p(x|\eta)}[[T(x)-\mu(\eta)][T(x)-\mu(\eta)]^T] = \text{Cov}(T(x))$

  (Hessian) (proof as exercise)

- $Z(\eta)$ is the mgf for $p(x;\eta)$ $\quad t \mapsto Z(\eta, t)$ for $t \in \mathbb{R}^p$

  $\mathbb{E}_x \exp(t^T T(x))$

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

↝ why derivatives of A gives cumulants

1st → mean
2nd → covariance
3rd → 3rd cumulant

14h56

Estimation of parameters for DGM

DGM: parametric family $P_{\mathcal{H}} = \{p_\theta(x) = \prod_i p(x_i | x_{\pi_i}, \theta_i)\}$

$\theta = (\theta_1, \dots \theta_{|V|})$

$\theta \in \mathcal{H} = \mathcal{H}_1 \times \dots \mathcal{H}_{|V|}$

independent parameterization

i.e. no tying of parameters

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

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

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

$$\sum_{j=1} \Big\{ \underbrace{\sum_{i=1} \cdots}_{f_j(\theta_j)} \Big\}$$

example: for discrete R.V. $\Rightarrow \theta_j^{ML}$ = proportion of observations

$$\frac{\#(x_j = k,\ x_{\pi_j} = \text{something})}{\#\ x_{\pi_j} = \text{something}}$$

gives $\hat{p}(x_j = k \mid x_{\pi_j})$

(*) if have latent variable (i.e. unobserved variable)

$\Rightarrow$ use E.M. (like we did for HMM)

<u>UGM:</u>

example for exp family

$$p(x \mid \eta) = \exp\Big(\sum_C \eta_C^T T_C(x_C) - A(\eta)\Big)$$

$\rightarrow$ unlike in a DGM, $\log p(x \mid \eta)$ does <u>not</u> separate as just $\sum_C f_C(\eta_C)$

gradient ascent on log-likelihood

$$\frac{1}{n}\sum_{i=1}^{n} \log p(x^{(i)} \mid \eta) = \sum_C \eta_C^T \Big(\underbrace{\frac{1}{n}\sum_{i=1}^{n} T_C(x_C^{(i)})}_{\hat{\mu}_C}\Big) - \frac{n A(\eta)}{n}$$

$$\nabla_\eta [\quad] = \hat{\mu}_C - \mu_C(\eta)$$

$\hookrightarrow \mathbb{E}_{p(x \mid \eta)}[T_C(x_C)]$

to compute this, need <u>inference</u>

e.g. Ising model $T_{ij}(x_i, x_j) = x_i x_j$

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

here need to use approximate inference ← sampling / variational method

<u>Approximate inference</u>

sampling

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

Why sampling?

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

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

b) approximate $p(x_i)$ "marginal inference"

→ special case of expectations

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

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

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

properties: 1) unbiased $\mathbb{E}_p[\hat{\mu}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_p[f(X^{(i)})] = \frac{n}{n}\mu = \mu$ (this is true even if $X^{(i)}$ are dependent)

expectation over $(X^{(i)})_{i=1}^n$

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

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

by independence → off diagonal terms are zero

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

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

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

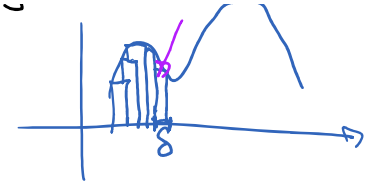
$$\boxed{\mathbb{E}[\|\hat{\mu} - \mu\|^2] = \frac{\sigma^2}{n}}$$

note that no dimension in rate (apart σ^2)

aside on numerical computing: 1d is "easy"

- numerical integration in 1D

- numerical integration in 1D



for function f L-Lipschitz

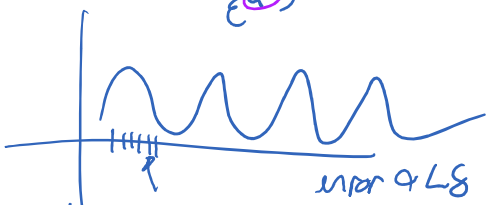
$|f(x) - f(y)| \leq L\|x-y\|$

error $\leq \varepsilon$ $\Rightarrow$ complexity $O(\frac{1}{\varepsilon})$ of approximating integral within ε

$\rightsquigarrow$ use $\delta \approx \varepsilon$

but grid in dimension d $O(\frac{1}{\varepsilon^d})$ curse of dimensionality!

- global optimization in 1d



complexity $O(\frac{1}{\varepsilon})$ to find global optimum