Lecture 2 - motivation / prob. review

September 7, 2021 4:27 PM

bit.ly / IFT 6269-F21

today: • motivation
• review of prob theory

why graphical models?

$1:T \triangleq \{1, \dots, T\}$

POS tagging observation $(x_1, x_2, \dots, x_T) \triangleq x_{1:T}$ $x_{\{1,3,5\}}$

$x_t \in \{1, \dots, K\}$ (x_1, x_3, x_5)

K: size of vocabulary $[(x_3, x_1, x_5)]$

want to model $p(x_{1:T})$ issue: exponential size → in length of input
state space

$\Rightarrow K^T - 1$ parameters needed to fully describe dist.

trick: make a factorization assumption about p

$$p(x_1, \dots, x_T) = f_1(x_1) f_2(x_2|x_1) \cdots f_T(x_T|x_{T-1})$$

factor → 2 variables
$\Rightarrow \approx K^2$ parameters to specify

clique in graph. model

theme: representation

T factors $\Rightarrow$ $T \cdot K^2$ parameters $\ll K^T$

computation?

say want compute $p(x_1)$ "marginal"

$$p(x_1) = \sum_{x_2, x_3, \dots, x_T} p(x_{1:T}) \to \sum_{x_2 \in \{1,\dots,K\}} \sum_{x_3 \in \{1,\dots,K\}} \cdots \sum_{x_T \in \{1,\dots,K\}}$$

exponential sum!?

distributivity

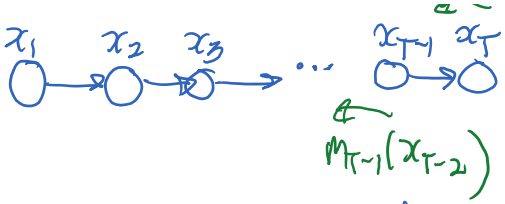
$$= \sum_{x_{2:T}} f_1(x_1) f_2(x_2|x_1) \cdots f_T(x_T|x_{T-1}) \qquad a \cdot (b+c) = a \cdot b + a \cdot c$$

$$= f_1(x_1) \left(\sum_{x_2} f(x_2|x_1) \left(\sum_{x_3} f(x_3|x_2) \left(\dots \left(\sum_{x_T} f_T(x_T|x_{T-1}) \right) \dots \right) \right) \right)$$

$m_T(x_{T-1})$ ← $O(K^2)$ to compute

$$\sum_{x_{T-1}} f_{T-1}(x_{T-1}|x_{T-2}) \cdot m_T(x_{T-1})$$

$m_T(x_{T-1})$



"message passing alg." to compute efficiently marginal $p(x_t)$

→ $T \cdot k^2$ time $\ll O(k^T)$

Key themes:

I) representation: how to represent structured prob. dist?

[prob.]

- graph → factorization
- parameterization — full table param. vs. "exponential family" param ↝ more structured

II) estimation: given data, how to learn / estimate the parameters of dist?

[statistics]

↝ learning e.g.
- maximum likelihood
- max. entropy
- moment matching

III) "probabilistic" inference:

[CS]

answer questions about data

e.g. compute $p(y \mid x)$ or $p(x)$

(y: query, x: observation; conditional; marginal)

↝ computation e.g.
- message passing
- approximate inference
  - sampling
  - variational methods

15h43

next: probability review

why? → principled framework to model "uncertainty"

sources of uncertainty

1) intrinsic uncertainty → quantum mechanics

2) partial information / observation:
- card games
- rolling a die → don't know the exact initial conditions

3) incomplete modeling of a complex phenomenon

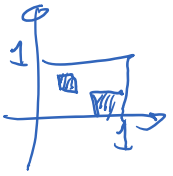
example: "most birds can fly" → simple rule can be

a complex phenomenon example: "most birds can fly" → simple rule can be advantageous but then yields uncertainty

(computational issue are also important)

• "AI"

probabilities → like areas



Notation: X_1, X_2, X_3 X Y Z

random variables (usually real-valued)

x_1, x_2, x_3 x y z

their "realizations"

X a random variable → represents an uncertain quantity e.g. X "result of a die roll"

$X = x$ represents the "event" that X takes the value x

Ω_X → the sample space of "elementary events" possible values of my R.V. $\Omega_X = \{1,2,3,4,5,6\}$

two standard types of R.V. ← discrete where Ω is countable

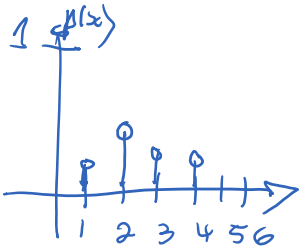
Continuous " " is uncountable

I) [assume Ω is countable → discrete case]

(discrete) R.V. X is characterized by a probability mass function (pmf)

"lower case" → $p(x)$ for $x \in \Omega$

pmf p is s.t. $\begin{cases} p(x) \geq 0 & \forall x \in \Omega \\ \sum_{x \in \Omega} p(x) = 1 \end{cases}$



probability distribution "big P"

is a mapping $P: \mathcal{E} \to [0,1]$

$\mathcal{E} = 2^{\Omega}$ ≜ set of all subsets of Ω

↑ "set of events"

that satisfy the Kolmogorov axioms

$P(E) \geq 0 \quad \forall E \in \mathcal{E}$

$P(\Omega) = 1$

$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$ when E_i's are disjoint

("σ-field" in measure theory)

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) \quad \text{when } E_i\text{'s are disjoint}$$

("σ-field" in measure theory needed when Ω is uncountable)

think area

notation $P\{X=x\} = p(x)$
$P(\{X=x\})$

$$P\{X=x \text{ or } X=x'\} = P\{X=x\} + P\{X=x'\} \overset{\text{pmf}}{=} p(x)+p(x') \quad \{\text{if } x \neq x'\}$$

$E = \{x, x'\}$

for discrete $P(E) = \sum_{x \in E} p(x)$

continuous R.V.

continuous R.V. is characterized by a probability density fct. p (pdf)

$p: \Omega \to \mathbb{R}_+$
$p(x) \geq 0 \quad \forall x \in \Omega$

p is integrable and $\int_\Omega p(x)dx = 1$

prob. of events

$\Omega = \mathbb{R} \quad P([a,b]) = \int_a^b p(x)dx$

$\mathcal{E} \to$ Borel σ-field

Recap: discrete R.V. X; pmf $p(x) \leftrightarrow P\{X=x\} = p(x)$

cts. R.V. X; pdf $p(x)$ $\quad P\{X=x\} = 0$

"$P\{X \in x \pm \frac{dx}{2}\} \approx p(x)\,dx$"

$$P\{X \in x \pm \tfrac{a}{2}\} = \int_{x-\frac{a}{2}}^{x+\frac{a}{2}} p(x')dx'$$

[Sidenote: can change pdf p on a countable # of pts. without changing "anything"] — has measure 0 according to Lebesgue measure