DS-GA 1005: INFERENCE AND REPRETENTATION

Logistics:

- Instructor: Alberto Bietti
- TA: Carles Domingo-Enrich
- Office Hours: Mondays, 5:00-6:45 pm ET (Zoom)
- Website for notes/readings/HW
- Campuswire

Lecture 1: INTRODUCTION & MOTIVATION

1 Motivation: Reasoning with probabilities

- Goal: reason about variables $x = (x_1, \dots, x_d)$

Ex: (i) x_1 = cough? x_2 = flu? x_3 = fever? x_4 = winter?

$x_i \in \{0,1\}$

$x \in \{0,1\}^d$ (hypercube)

Q: "I have a cough. Do I have the flu?"

(ii) $x = (z, y)$ — z: image $\in \mathbb{R}^{128 \times 128}$, y: label $\in \{0,1\}$ — Image classification

Q: "Is this an image of a cat?"

(iii) $x = (x_{1 \dots T})$ — Language modeling

$x_i \in \{1, \dots, K\}$ are words (K = vocabulary size)

Q: "How should I complete this sentence?"

Probabilistic modeling

→ **Representation** of data using prob. distribution $p(x; \theta)$ (parameters)

→ **Inference** to answer questions:

- $p(\text{flu} \mid \text{cough})$?
- $p(y=1 \mid z)$?
- $p(x_T = \text{"mat"} \mid x_{1:T-1} = \text{"the cat sat on the"})$

only need:

- conditioning / Bayes rule

$$p(a,b) = p(a)\,p(b|a) = p(b)\,p(a|b)$$

- marginalization

$$p(a) = \sum_b p(a,b)$$

→ **Learning**: find a good θ from data $\{x^{(i)}\}_{i=1\ldots n}$

often maximum likelihood estimation

$$\hat{\theta}^{MLE} := \arg\max_\theta \prod_{i=1}^{n} p(x^{(i)}; \theta)$$

Note: In Bayesian statistics, θ are typically treated as variables ⇒ Learning is framed as inference

e.g. $p(y=1|z) = \int p(y=1|z,\theta)\,p(\theta|z)\,d\theta$

instead of $p(y=1|z; \hat{\theta})$

2. The role of structure

Curse of dimensionality: with no assumptions, things become exponentially worse with the dimension.

Example: $x \in \{0,1\}^d$

→ no assumptions: $p(x;\theta) = \theta_x$

equivalent model: $x \in \{1,\dots,K\}$, $\boxed{K = 2^d}$ (multinomial)

$\theta \in [0,1]^K$, $\sum_{i=1}^K \theta_i = 1$

→ representation requires $K-1 = 2^d - 1$ parameters

→ inference: computing a marginal over one variable costs $O(2^d)$ time

$$p(x_1 = 1) = \sum_{x_2,\dots,x_d} p(x_1 = 1, x_2, \dots, x_d)$$

→ learning (using MLE) may require $O(2^d)$ samples: →

Error of MLE

- true model: $p(x = k) = \theta_k$
- observations $x^{(i)} \in \{1,\dots,K\}$, $i = 1,\dots,n$
- MLE: $\hat{\theta}_k = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{x^{(i)} = k\}$

$$\mathbb{E}\,\|\hat{\theta}_k - \theta_k\|^2 = \mathrm{Var}(\hat{\theta}_k) = \frac{1}{n}\mathrm{Var}\left(\mathbb{1}\{x^{(1)} = k\}\right) = \frac{1}{n}\theta_k(1-\theta_k)$$

$Z = \mathbb{1}\{x^{(1)} = k\}$ Bernoulli r.v. with param θ_k

$\mathrm{Var}\, Z = \theta_k(1-\theta_k)$

$$\frac{\mathbb{E}\,\|\hat{\theta}-\theta\|^2}{\|\theta\|^2} = \frac{\sum_k \theta_k(1-\theta_k)}{n\|\theta\|^2} = \frac{1-\|\theta\|^2}{n\|\theta\|^2}$$

In many cases, $\|\theta\|^2 \sim \frac{1}{K}$, e.g. when $\theta_k = \frac{1}{K}$

Then $\frac{\mathbb{E}\,\|\hat{\theta}-\theta\|^2}{\|\theta\|^2} \sim \frac{K}{n} = \frac{2^d}{n}$

Remark: For continuous variables, e.g. $x_i \in [0,1]$, this can be even harder: density estimation

---

$\Rightarrow$ We need more assumptions on problem structure (prior knowledge, inductive bias)

- Factorization, conditional independencies

(chain rule) recall that we always have

$$p(x_1,\dots,x_d) = p(x_1)\cdot p(x_2|x_1)\,p(x_3|x_1,x_2)\cdots p(x_d|x_1,\dots,x_{d-1})$$

→ still cursed representation for binary variables

$(1+2+4+\cdots+2^{d-1} = 2^d-1$ parameters)

→ structural assumptions using (conditional) independence

Example: $p(x_3|x_1,x_2) = p(x_3|x_2)$ $\quad x_3 \perp x_1 \mid x_2$

$p(x_t|x_{1:t-1}) = p(x_t|x_{t-1})$ (Markov)

Then we have the Factorization:

$$p(x_1,\dots,x_d) = p(x_1)\,p(x_2|x_1)\,p(x_3|x_2)\cdots p(x_d|x_{d-1})$$

- Much more compact representation !!

  $1 + 2 + 2 + \cdots + 2 = 1 + 2(d-1)$ parameters

- Efficient inference using Dynamic Programming
- Efficient learning (convex, parametric)

Ex: $x = (\text{Season}, \text{Flu}, \text{Hayfever}, \text{Muscle-pain}, \text{Congestion})$

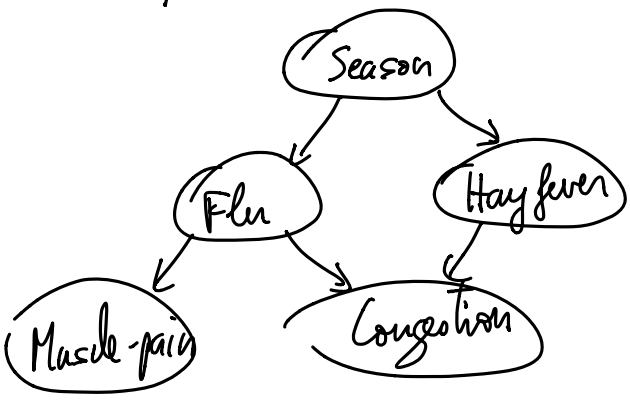
- Independencies e.g. $C \perp S \mid F, H$

  $P(\text{Congestion} \mid \text{Season}, \text{Flu}, \text{Hayfever}) = P(\text{Congestion} \mid \text{Flu}, \text{Hayfever})$

- Factorization:

  $p(x) = p(S)\,p(F|S)\,p(H|S)\,p(C|F,H)\,p(M|F)$

- Conveniently encoded with a directed (acyclic) graph (DAG)



3. Contents of this course

- **Probabilistic Graphical Models**

  directed / undirected) Graph $G = (V, E)$ — V: vertices/nodes; E: edges

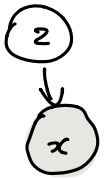
  - Equivalence for a distribution $p(x)$:

    $p(x)$ Factorizes according to G $\iff$ $p(x)$ satisfies a set of conditional independencies encoded by G

  - directed vs undirected : encode $\neq$ independencies
  - Powerful abstraction : generic graph algorithms for inference on large families of distributions

- **Examples of families of distributions**

  → Latent-variable models

  (unobserved)
  (observed)

  $$p(x) = \int p(x|z)\, p(z)\, dz$$

  often, we want to compute posterior $p(z|x)$

  e.g. : • Mixture models

  $z \in \{1, \dots, K\}$ mixture component

  $x|z=k \sim \mathcal{N}(\mu_k, \Sigma_k)$ Gaussian

- **Bayesian Inference:**
  parameters θ are a latent variable

- **Random graph models** (Stochastic Block Model)
  community detection
  node $z_i \in \{\pm 1\}$ community
  edge $p(e_{ij}=1 \mid z_i, z_j) = \begin{cases} p & \text{if } z_i = z_j \\ 1-p & \text{o/w} \end{cases}$

  Observe edges E, infer communities $p(z \mid E)$

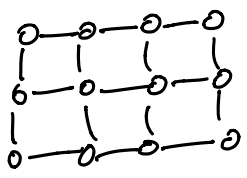
→ **Gibbs or "Energy-based" models**

$$p(x) = \frac{1}{Z} e^{-E(x)}$$

- $E(x)$: Energy function
- $Z = \int_x e^{-E(x)} dx$: Partition function

- Foundational model in statistical physics (Boltzmann/Gibbs distribution)
- $E(x)$ often involves interaction terms among few variables: $E(x) = \sum_i \lambda_i x_i \qquad x_i \in \{\pm 1\}$
  $E(x) = \sum_{i,j} \lambda_{ij} x_i x_j$

e.g. Ising model



→ **Implicit models / transport models**

Instead of modeling $p(x)$ or $p(x \mid z)$ directly, view x as the transformation of a r.v. z with simple distribution

e.g. $z \sim \mathcal{N}(0, I)$ $\quad x = \phi(z)$

$\phi : \mathcal{Z} \to \mathcal{X}$

e.g. a Neural Network

■ Inference algorithms

→ Exact inference

For certains graph structures, it is possible to perform exact inference efficiently.

Main example: trees (including chains)

- e.g. $p(x) = p(x_1)\, p(x_2 | x_1) \cdots p(x_d | x_{d-1})$

marginal on x_s ?

$$p(x_s) = \sum_{x_1, \dots, x_{s-1}} p(x_1, \dots, x_s)$$

$$= \sum_{x_{s-1}} p(x_s | x_{s-1}) \sum_{x_{s-2}} p(x_{s-1} | x_{s-2}) \cdots \sum_{x_1} p(x_2 | x_1)\, p(x_1)$$

Define $q_1(x_1) = p(x_1)$

$$q_s(x_s) = \sum_{x_{s-1}} p(x_s | x_{s-1})\, q_{s-1}(x_{s-1})$$

We have $p(x_s) = q_s(x_s)$ $\quad$ $O(sK^2)$ computation for $x_i \in \{1, \dots, K\}$

- More generally: sum-product / junction tree / belief propagation | algorithm

("message-passing" algorithms)

→ Approximate inference

Q: What if G is not a tree?
What if posterior is intractable, e.g. due to difficult integral?

(i) Can still apply algorithms beyond trees
"Loopy belief propagation"

(ii) Variational inference
- Reframe inference as optimization
- Leverage convexity properties in modeling distributions
- Notably useful w/ exponential family distributions

(iii) Sampling algorithms
- Approximate distributions using samples
  (we can then compute expectations approximately
  e.g. $\mathbb{E}_{p(z|x)}[f(z)] \approx \frac{1}{m}\sum_{i=1}^{m} f(z^{(i)})$)
- Flexible sampling algos: MCMC (Markov Chain Monte Carlo)
  Define a Markov Chain whose stationary distribution is the target distribution, e.g. $p(z|x)$

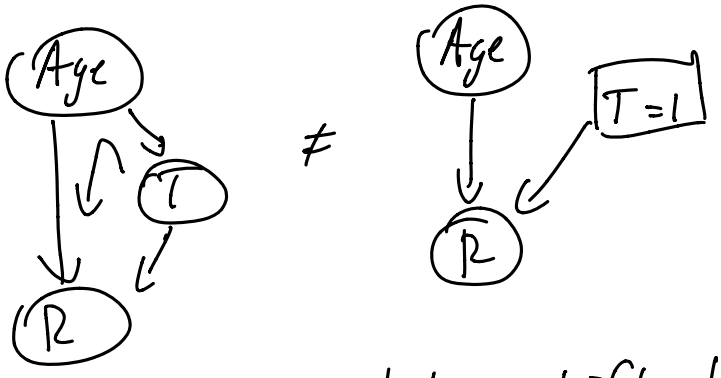
Causal Inference

So far, we only used a single distribution $p(x)$
Often, we might care about changes to the distribution
"Counterfactual" questions

- Ex: → policy decisions
  ("what happens if we increase the minimum wage?")
  → medical treatments
  ("what is the effect of vaccines")
  → A/B testing

- Causal models ≠ probabilistic models
  X: altitude Y: temperature
  causal model: we expect only X → Y works
  prob. model: X → Y and Y → X both work
  $p(x|y)p(y) = p(y|x)p(x) = p(x,y)$
  arrows have "causal" meaning

- Interventions: change only some "part" of the distributions
  e.g. $P(\text{recovery} \mid do(\text{treatment} = 1))$
  $\neq$
  $P(\text{recovery} \mid \text{treatment} = 1)$



- Related to: distribution shift, transfer learning