Lecture 4+5 Classical Inference & Structure Learning on Trees

Last time: undirected graphical models, conditional indep., testing Ising vs uniform.

What else do we want to with graphical models?

Separate 2 settings:

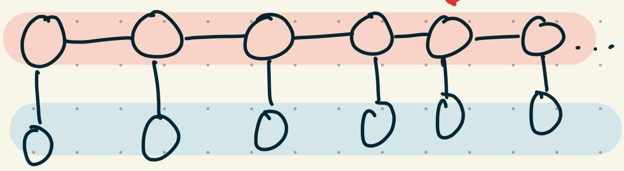
① Willing to assume the world is described by some known graphical model.

② Believe world is described by some unknown graphical model.

① Observe values of random variables corresponding to a subset of nodes. Infer something about observations / rest of graphical model.

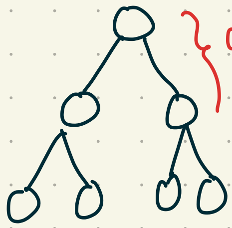
Example models :

- Hidden Markov Models

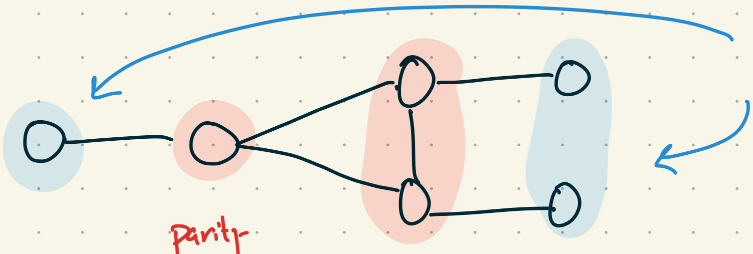


Speech recognition

- Evolution of genomes



- Error-correcting codes



Example inference tasks:

- Compute $\Pr(\text{observations}) = \sum_{\substack{x \text{ matching} \\ \text{observed values on} \\ \text{observed nodes}}} \Pr(x)$

- Compute marginal distribution $\Pr(x_A)$ for some $A \subseteq V$

- Compute conditional distribution $\Pr(x_A \mid x_B = y)$

  "Posterior inference"

- Compute most likely x – "mode".

These are classical – a whole course at MIT, "Algorithms for

Inference". So we will only scratch the surface here, then move on to other topics.

Naive algorithms: involve sum or maximization over all possible values for x - intractable b/c $(\#\text{ of vals per node})^{|V|}$ possible values

In general shouldn't hope to beat naive algs by much - NP, #P hardness

But, can do (much) better in special cases.

On trees, all of these can be solved by dynamic programming!

message passing, Belief propagation, Viterbi, sum-product, max-product, junction tree, ...

Dynamic Programming for Marginal & conditional distributions

Computing marginal & conditional is ≈ same - just question of whether we fix values of some nodes by introducing potential

$$\delta(x) = \begin{cases} 1 & \text{if } x = \text{desired value} \\ 0 & \text{o.w.} \end{cases}$$

Problem: Given a tree T and factors $\{\psi_C\}_{C \in \text{Cliques of } T}$, compute $\{\Pr(x_v = y)\}_{v \in T, y}$ - ie all 1-wise marginals.

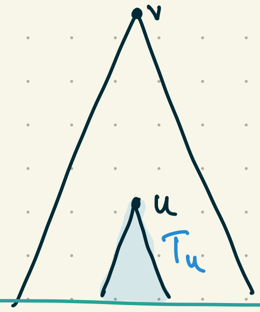
Assuming discrete distributions over universe Ω.

<u>Observation:</u> only cliques in T are edges + individual nodes.

Won't use this, but for intuition, therefore,

$$\Pr(x) \propto \prod_i \psi_i(x_i) \cdot \prod_{ij} \psi_{\underbrace{i\sim j}_{i,j \text{ adjacent in tree}}}(x_i, x_j)$$

Root the tree at v. For any $u \in T$, let T_u be the graphical model we get by restricting to subtree rooted at u.



Let $x_v \in \Omega$.

$$\Pr(X_v = x_v) = \frac{1}{Z} \sum_{x \in \Omega^{T \setminus \{v\}}} \prod_i \psi_i(x_i) \prod_{i \sim j} \psi_{ij}(x_i, x_j)$$

$$= \frac{1}{Z} \psi_v(x_v) \cdot \boxed{\prod_{i \in \text{children}(v)} \boxed{\sum_{x_i \in \Omega} \psi_{iv}(x_i, x_v) \cdot \underbrace{\psi_i(x_i) \cdot \sum_{x \in \Omega^{T_i}} \prod_{j \in T_i} \psi_j(x_j) \prod_{j \sim_{T_i} k} \psi_{jk}(x_j, x_k)}_{\text{the things we would've multiplied to compute } \Pr_{T_i}(X_i = x_i)}}}$$

If we knew , could compute in $O(|\Omega|)$ time, and in $O(|\Omega| \cdot \text{degree})$ time.

Can use a dynamic program, computing for each choice of $x_i \in \Omega$ and each subtree T_i, when computation for T_i happens before its parent.

Time: $O(n \cdot \text{degree} \cdot |\Omega|^2)$

(compute Z by adding appropriate table entries.)

- Makes it look like would need $O(n^2)$ to compute all marginals, but there is a clever way to do all at once in same $O(n \cdot \text{degree} \cdot |\Omega|)$ time. (Can Google "sum product" or "Belief Propagation")

What happens on non-trees?

- Can view as a "message" passed by X_i to its parent, constructed from similar "messages" it recieved from its children.

- Could use same formula for constructing messages and passing them around, but now on non-trees. "loopy BP".
  <u>Heuristic</u>, sometimes seems ok in practice, maybe expected to work if graph has no short cycles ("locally tree-like") and weak long-range correlations.

- Can try other algs – MCMC, variational inference, ...
  always heuristic, maybe w/ guarantees in special cases.
  take "Algorithms for Inference" @ MIT.

Moving on to 2 :

Learning Graphical Models

Assume getting samples $X_1 \dots X_n$ iid from some unknown graphical model.

What can we learn about it?

Fully-connected graph $\Leftrightarrow$ represent any distn. So need some assumptions.

2 learning tasks:

1. TV learning

2. Structure learning - find the underlying graph

   - how to distinguish no edge, edge w/ $\psi_{ij} \approx 1$?

   - need assumptions on ψ_{ij}'s.

Today : trees.

Chow-Liu (infinite sample version) :

Instead of iid samples, let's pretend we get access to marginal distribution of every pair of variables X_i, X_j.

- Compute $I(X_i;X_j) = \mathbb{E}_{x_i,x_j} \log \frac{\Pr(x_i,x_j)}{\Pr(x_i)\Pr(x_j)} = KL(\{x_i,x_j\} \| \{x_i\} \otimes \{x_j\})$

- Let G be a graph where weight of edge i,j is $I(X_i;X_j)$

- Output maximum spanning tree of G

Reminder: $= \arg\max_{\text{Tree } T \text{ on vertices of } G} \text{weight}(T)$

<u>Theorem</u>: Suppose T is a tree-structured graphical model. Then Chow-Liu, run on marginals of T, returns T. (Exception: if there is another tree T' which can represent same distn, can get T' — MST will not be unique.)

Proof: Follows from 2 key claims:

① If S is another tree-structured graphical model on the same set of variables, s.t. for every edge $i,j \in S$, $\underbrace{\{x_i, x_j\}_S}_{\text{joint dist'n of } x_i, x_j \text{ under } S} = \{x_i, x_j\}_T$ and $\{x_i\}_S = \{x_i\}_T$ for all i, then the distribution of S = distribution of T iff S is a maximum spanning tree in G.

② For every spanning tree S of G $\exists$ a distribution which is Markov w.r.t. S satisfying hypotheses of ①.

So, let S be MST in T, define a dist'n Markov w.r.t. S as in ②.

Then that dist'n must $= T$.

Proof of ①: We have

shorthand for $\Pr_T(x)$

$$KL(T \| S) = \mathop{\mathbb{E}}_{x \sim T} \log \frac{T(x)}{S(x)}$$

$$= \underbrace{\mathop{\mathbb{E}}_{x \sim T} \log T(x)}_{\text{indep of } S} - \mathop{\mathbb{E}}_{x \sim T} \log S(x)$$

$$= \mathop{\mathbb{E}}_{x \sim T} \log T(x) - \mathop{\mathbb{E}}_{x \sim T} \log \prod_i \Pr_S(x_i \mid x_{\text{parent}_S(i)})$$

$$= \mathop{\mathbb{E}}_{x \sim T} \log T(x) - \mathop{\mathbb{E}}_{x \sim T} \log \prod_i \Pr_T(x_i \mid x_{\text{parent}_S(i)})$$

$$= \mathop{\mathbb{E}}_{x \sim T} \log T(x) - \left(\mathop{\mathbb{E}}_{x \sim T} \sum_i \log \frac{\Pr_T(x_i, x_{\text{parent}_S(i)})}{\Pr_T(x_i) \Pr_T(x_{\text{parent}_S(i)})} - \sum_i H(x_i) \right)$$

$$= \mathop{\mathbb{E}}_{x \sim T} \log T(x) + \sum_i H(x_i) - \sum_i I(x_i ; x_{\text{parent}_S(i)}).$$

If distn on S = distn on T, then this $= 0$. If not 0, then $\sum_i I(X_i; X_{\text{parent}_S(i)})$ must not be maximal $\square$

Proof of (2): define distribution via $Pr_S(x) = Pr_T(x_{\text{root}}) \cdot \prod_i Pr_T(x_i | x_{\text{parents}_S(i)})$.

Marginals match by induction on depth.

What about finite samples?

- estimate $I(x_i;x_j)$ using empirical distr's
- how accurately do we need them?

  if interactions are really weak, might need a lot of samples to distinguish

  $A - B - C$ vs $A - C - B$

2 options:

- add assumptions on interaction strength
- learn in TV — if $A - B - C$ vs $A - C - B$ hard to distinguish, describe close-by distr's anyway.

also [Bhattacharyya - Gayen - Price - Vinodchandran]

Theorem [Daskalakis-Pan]: Assume alphabet $|\Omega| = 2$ (binary Ising model).

With $O\left(\frac{n \log n}{\varepsilon^2}\right)$ samples, Chow-Liu + empirical estimates for $I(\cdot;\cdot)$ learns a distribution which is ε-close to true one in TV dist.

[Aside: what "distribution" does Chow-Liu output? As described above it only gives the tree. (Can get full distribution by estimating $\Pr(x_i \mid x_{\text{parent}(i)})$ from samples.)]

This theorem is out of scope for this class. But we will discuss why you can't beat $O\left(\frac{n \log n}{\varepsilon^2}\right)$ samples, returning to Le Cam's method from lecture 1.

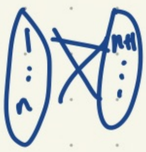
Theorem [Koehler]: To test between:

null: D is uniform on $\{\pm 1\}^n$

alternative: D is a tree-structured Ising model w/ $TV(D, \text{unif}) \geq 0.01$

requires $\Omega(n \log n)$ samples.

Idea:



always match $1 \dots n$ to $n+1 \dots 2n$

first, pick a random matching M on $\{1, 2, \dots, 2n\}$, from a set S of "allowed" matchings.

then, $X_1 \dots X_N \sim$ Ising model w/ $\Pr(x) \propto \exp\left(\frac{\beta}{\sqrt{n}} \sum_{i \le n} x_i x_{M(i)}\right)$

Claim 1: If Hamming dist. between $M, M' \geq \Omega(n)$, then

$$TV(P_M, P_{M'}) \geq \Omega(1).$$

Proof: Consider the distribution of $\sum_i x_i x_{M(i)}$.

Under P_M, it is a sum of $\frac{n}{2}$ independent ± 1 bits,

each w/ bias
$$\mathbb{E}_{x \sim P_M} x_1 x_{M(1)} = \frac{\exp\left(\frac{\beta}{\sqrt{n}}\right) - \exp\left(-\frac{\beta}{\sqrt{n}}\right)}{\exp\left(\frac{\beta}{\sqrt{n}}\right) + \exp\left(-\frac{\beta}{\sqrt{n}}\right)} = \frac{\beta}{\sqrt{n}} \pm O\left(\frac{1}{n}\right)$$

Hence, $\frac{1}{\sqrt{n}} \sum x_i x_{M(i)} \longrightarrow N(\beta, O(1))$.

In particular, $\Pr\left(\frac{1}{\sqrt{n}} \sum x_i x_{M(i)} \geq \beta\right) \to \frac{1}{2}$.

Under $P_{M'}$, bias is 0 for at least $\Omega(n)$ terms in the sum,

so $\mathbb{E}_{P_{M'}} \frac{1}{\sqrt{n}} \sum_i x_i x_{M(i)} \leq (1 - \Omega(1))\beta$.

No longer a sum of independent terms, but variance is still $O(1)$.

So, $\Pr_{P_{M'}}\left(\frac{1}{\sqrt{n}} \sum x_i x_{M(i)} \geq \beta\right) \leq \frac{O(1)}{\beta^2} \ll \frac{1}{2}$ if $\beta \gg 1$. $\square$

TV error ε w/ N samples, can identify underlying matching using N samples.

Now we need a new tool to show that identifying the underlying matching is not possible.

Fano's Inequality: Let M, X be joint random variables, M discrete taking values in finite set $\mathcal{M}$. Let $f(X) \in \mathcal{M}$ take values in $\mathcal{M}$. Then $\Pr(f(X) \neq M) \geq \frac{H(M) - I(M;X) - 1}{H(M)}$.

Intuition: if X doesn't contain much information about M, can't identify M using X.

How can we bound $I(M; X_1 \ldots X_N)$?

Lemma $I(A;B) \leq \max_{a,a'} KL(\underbrace{\{B \mid A=a\}}_{\text{distribution of } B \text{ conditioned on } A=a} \| \{B \mid A=a'\})$

Deferring proof of lemma for now, how do we use it?

$$KL\left(\{X_1 \ldots X_N \mid M\} \,\|\, \{X_1 \ldots X_N \mid M'\}\right) = N \cdot KL\left(\{X \mid M\} \,\|\, \{X \mid M'\}\right)$$

by tensorization.

Now, $KL(\{X \mid M\} \,\|\, \{X \mid M'\}) = \underset{X \sim P_M}{\mathbb{E}} \log \dfrac{\exp\left(\frac{\beta}{\sqrt{n}} \sum_i x_i x_{M(i)}\right)}{\exp\left(\frac{\beta}{\sqrt{n}} \sum_i x_i x_{M'(i)}\right)}$

$$= \underset{X \sim P_M}{\mathbb{E}} \frac{1}{\sqrt{n}} \sum_i x_i \left(x_{M(i)} - x_{M'(i)}\right) \leq O(1)$$

(if $\beta = O(1)$)

Applying Fano, if we tried to use a function $f(X_1 \ldots X_N)$ to guess M, we would have $\Pr(f(X_1 \ldots X_N) \neq M) \geq 1 - \dfrac{O(N)}{\#\text{ possible matchings}}$

Fact: there is a set of $n^{\Omega(n)}$ matchings all w/ Hamming dist $\geq \Omega(n)$

So, if $N = o(n \log n)$, can't identify M from $X_1 \ldots X_N$. $\square$

<u>Loose ends</u>: ① Proof of Fano's inequality

By data processing, $I(M;X) \geq I(M; f(X))$

$$= H(M) - H(M \mid f(X)) = (*)$$

let E be a 0/1 r.v., $E = \begin{cases} 0 & \text{if } f(X) = M \\ 1 & \text{o.w.} \end{cases}$. Then $H(M \mid f(X)) = H(M, E \mid f(X))$,

so $(*) = H(M) - H(M, E \mid f(X))$

$$= H(M) - H(E \mid f(X)) - H(M \mid E, f(X))$$

$$\geq H(M) - 1 - H(M \mid E, f(X))$$

$$= H(M) - 1 - \Pr(E=1) \cdot H(M \mid E=1, f(X)) - \Pr(E=0) \cdot \underbrace{H(M \mid E=0, f(X))}_{=0}$$

$$= H(M) - 1 - \Pr(E=1) \cdot H(M \mid E=1, f(X)).$$

Rearranging, we get

$$\Pr(f(X) \neq M) = \Pr(E=1) \geq \frac{H(M) - 1 - I(M;X)}{H(M \mid E=1, f(X))} \geq \frac{H(M) - 1 - I(M;X)}{H(M)},$$

since conditioning reduces information.

② Proof of Lemma:

$$I(A;B) = KL\left(\{A,B\} \,\|\, \{A\} \otimes \{B\} \right)$$

$$= \mathop{\mathbb{E}}_{A,B} \log \frac{\Pr(A,B)}{\Pr(A)\Pr(B)}$$

$$= \mathop{\mathbb{E}}_{a\sim A} \mathop{\mathbb{E}}_{B} \log \frac{\quad}{\Pr(B)}$$

$$= \mathop{\mathbb{E}}_{a\sim A} \mathop{\mathbb{E}}_{B} \log \frac{\Pr(B|a)}{\mathop{\mathbb{E}}_{a'\sim A} \Pr(B|a')}$$

$$= \mathop{\mathbb{E}}_{a\sim A} KL\left(\{B|a\} \,\|\, \mathop{\mathbb{E}}_{a'\sim A} \{B|a'\} \right)$$

$$\le \mathop{\mathbb{E}}_{a,a'} KL\left(\{B|a\} \,\|\, \{B|a'\} \right)$$

convexity of KL divergence!

$$\le \max_{a,a'} KL\left(\{B|a\} \,\|\, \{B|a'\} \right)$$