'08 PGMS (L1) - intro 23/04/2020

· You already have notes; use this space to record instructor assistance/exposition

- Koller + Friedman → intense
- Jordan - An Introduction - easier
(!) Supplement with materials e.g. papers and tutorials

· List grading on GitHub → this will give you a sense of how much you are missing out
· 4 HWs are not trivial

· IID data e.g. $X_1, X_2, \ldots, X_n \sim P$ (usual ML setting)
· EX: How graphs can uniquely and consistently specify a model M_G,
under what conditions we can estimate model parametrization and topology

e.g. NN as a graphical model
· EX: PGM - define probability distribution over data with complex structure
- in distinction against rule-based inference; reasoning under uncertainty,
- ~~rules~~ noise
- PGMs give systematic methods for reasoning under uncertainty

EX: Research questions
- see slides
representation - Articulation in common language (mathematics, algorithms)

inference - Prediction/estimation
- may be able to ask valid questions that are intractable to answer (NP-hard)

learning - Combine model + data in some kind of (function ⇒) score that encodes optimality over all possible models specifications

· Example: Multiple representations of same data (trees)

EX: Can we mathematically express/quantify what makes a 'good' representation → e.g. via some 'distance metric'

- There may exist hidden nodes; allow for placeholders
- PGM: Think through lens representation - inference - learning
  - observation data - hidden states - structures - probabilities

- Difference between "models" and "graphical models" representation
- (1): we can write down a joint distribution of a collection of random variables (assuming independence) (on binary r.v.s)

- via a probability table (remember Wasserman)

- $2^8 - 1$ state configurations for 8 binary r.v.s.

(1): memory issues as no. of r.v.s ↑; computer scientist would balk at using a large joint distrib. table

EX: in any case, there may be r.v.s we cannot observe if too many

Inference: e.g. $p(H|A)$

EX: - All questions (using enumeration) is NP-hard
- Probability distributions with no structure other than an enumerative table is likely not going to be helpful.

- e.g. 1000 stocks on NASDAQ for a portfolio
  - structure: → correlation via sector? and dependencies

- EX: How can we make use of domain knowledge/structure to make models more economical than enumerative probability tables.

Graphical Models

- Molecular biology
- PGM: Structure simplifies representation
  e.g. via physical location/communicative pathways (dependencies amongst variables)

PGM: instead of enumeration; think about traversal (factorisation law)
- Given a graph, traverse it, whenever you run into a node; write down ~~conditionally~~ ~~ind~~ distribution of r.v.s given their parents;

– no ~~pa~~ parents → marginal probability
· multiply together
· currently we assume this is feasible (prove this later)
· Rewriting joint as factorisation; more parsimonious representation of probability + dependencies

(w): (A) - check calculation - Benefits of PGM

1) Handle large multivariate distri using graph structure to factorise the distribution (representation cost)

- Formally; using conditional independence (next)

=

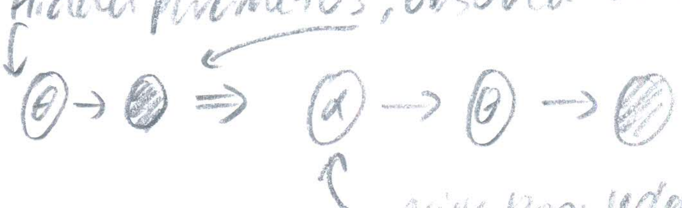
data integ
· Each term is self-contained, local-conditional distri
· In context of biology → allows for parallelism/data integration over biological labs; each lab only works with relevant LCD.
· use PGM to combine LCD at each "modality"

(w): possibilities for combining diverse, heterogeneous data sources in a modular fashion

=

statistical inference
- use priors to confine search for distribution of earth surface temperature (domain knowledge)
(e.g. not −273°C)
- via Bayes Theorem: Allows inference; placeholder for injection of prior knowledge

- PGM - Hidden parameters, observed data



prior knowledge over hidden parameters/r.v.s.

- universal way of representing structure of knowledge/mathematical algorithms ~~for~~

Ex: Also lots of downsides; PGM

· PGM is a particular mode of inference (not really ~~probab~~ 'model')

EX: · simplify exponentially-large probability distri without associated costs
  · and endow with structured semantics

Formal description: A family of distri on a set of r.v.s. compatible
with all probabilistic independence propositions
encoded with a graph that connects variables

- emphasis on allowing/enabling scientific communication

(a) - 2 GMs:

1) Directed edges: causality rel. (Bayesian Networks / Directed
Graphical Models)

2) Undirected edges: correlations (Markov Random Field / ...)

· Bayesian networks: $\boxed{53:20}$
- conditional independence of yellow x of red, conditional on green.
- social network interpretation
parents, children, co-parents

(Q) (A1): Be clear on terminology

(A1): $P(X|Y, \ldots) = P(X|Y)$

(W) (A3): Be clear on distinction of c.i in BN/MRFs

· MRFs
- conditional independence

· Given graph; use topology to extract conditional independence relations
· some formalism is required mathematically
between conditional independence relations - topological representation
· EX: 2 ways of specifying distri:-
i) Identify independence exhaustively via graph traversal algorithm;
write down distri that satisfies via testing proc.

ii) use factorisation; superimpose graphs on top of r.v.s.; use
graph factorisation rules and multiply.

EX: Are i) and ii) the same? (there are proofs in Koller + Friedman)

- (W) A4: Equivalence theorem → get to the point

- OM: EX: Formalises ML / stats in terms of graphs

EX: Allows contextualisation of many algorithms; PGM allows explicit consideration of topology

---

- DNA of PGMs:-
  1920s - Wright
  1980s - Spiegelhalter, Lauritzen, Judea Pearl
  (stats) (CS)

- many slides are in Appendix; a lot of slides, don't cover all; understand, distill key principles

PI-708

26/04/2020

- supplementing (U)
- Anything new, interesting, important
- U Notes
- Recall independence $\Rightarrow$ uncorrelated; but in general uncorrelated $\not\Rightarrow$ independence
- example in notes
- limitations of Pearson correlation $\rightarrow$ cannot capture non-linear dependences
- other measures of association leverage some kind of distance metric between distributions.
- independence: $f_{X,Y}(x,y) = f_X(x) f_Y(y)$
- KL-divergence, HSIC characterise distances between densities

Mutual info:

$$KL(P,Q) = \int_{x \in X} P(x) \log \frac{P(x)}{Q(x)} \, dx$$

$P(\cdot)$ $Q(\cdot)$ are density functions

- when $P = Q$; (in distribution?), $KL(P,Q) = KL(P,P) = 0$ i.e. $P(x) = Q(x)$ $\forall x \in X$
- ~~KL~~ when $P \neq Q$; $KL(P,Q) > 0$
- The desired measure of mutual information:

$$I(X,Y) = KL(f_{X,Y}, f_X f_Y)$$ i.e. KL between joint and product of marginal densities.

- successfully captures non-linear dependencies.
- computational issues (integral intractability)

HSIC $\rightarrow$ @(Gretton)

- Also for nonlinear dependences
- maximum mean discrepancy (MMD) between joint $f_{X,Y}$ and prod. marg $f_X f_Y$
- $MMD(P,Q) = \|\mu_K(P) - \mu_K(Q)\|_{H_K}$

$\mu_K(P) = \mathbb{E}_{z \sim P}[\phi(z)]$ - kernel embed. of P

$\phi(z)$: feature map of kernel K.

- HSIC$(X,Y)=0$ iff $X \perp Y$.

- partial correlation

this nuance is important

(*): Distinct from marginal correlation (in regression coefficients?)

- correlation between 2 variables given another

- X, Y, Z ; condition on Z.

- correlation between X and Y after conditioning on Z; or after eliminating linear effect of Z

- $$\rho(X,Y|Z) = \rho(e_X, e_Y) = \frac{cov(e_X, e_Y)}{\sqrt{var(e_X)}\sqrt{var(e_Y)}}$$

i) Regress X on Z; get residuals e_X

ii) regress Z on X; —— " —— e_Y

} correlation between residuals e_X, e_Y

(*)

$$X \perp\!\!\!\perp Y | Z \Rightarrow \rho(X,Y|Z)=0 \; ; \; \rho(X,Y|Z) \not\Rightarrow X \perp\!\!\!\perp Y|Z$$

- can use to create more meaningful gm than marginal dependency graph

- Analogous L.A form:-

$$\underline{R}_{ij} = \rho(X_i, X_j | X_{-ij})$$

$$\underline{R}_{ij} = - \frac{\underline{\Theta}_{ij}}{\sqrt{\underline{\Theta}_{ii}}\sqrt{\underline{\Theta}_{ij}}}$$ where $\underline{\Theta}$ is inverse covariance matrix

- conditional independence (*)

- $X \perp Y | Z$ - X is conditionally independent of Y; given Z

$$X \perp Y | Z \Leftrightarrow P(X,Y|Z) = P(X|Z)P(Y|Z)$$ (similar analogies) ~~to~~ - just independence qualified with conditioning

- Difficult to extract conditional independence if we use strong dependency measures / partial correlation

(*)

i.e. (X,Y,Z) jointly Gaussian

- shortcut: impose Gaussian assumption on r.v.s. $\rho(X,Y|Z)$ iff $X \perp Y|Z$