10708 - PGM - Readings 24/04/2020

(L1) - introduction

Jordan (2004):

(W): I don't understand the significance of kernels in the PGM specification

(W): It is not fully specified how the assertions of conditional independence in directed and undirected graphs differ (i.e. arrows/edge)

(W): I don't fully understand how we move from the formalism for cliques C → formalism for factors
(undirected graph) (factor graph)

(W): Conversion of directed → undirected formalism; work with (2) ~~it.~~

(W): 3.1 exact algorithms

- Not entirely sure how / needs investigation of distributive law for marginalisation of $p(x_1)$ i.e. how

$$p(x_1) = \sum_{x_2}\sum_{x_3}\sum_{x_4}\dots\sum_{x_6}\frac{1}{Z}\psi(x_1,x_2)\psi(x_1,x_3)\psi(x_2,x_4)\psi(x_3,x_5)\psi_{2}(x_2,x_5,x_6)$$

is simplified

(W): Key terminology: - elimination order, triangulation algorithm, tree width

- (W): At a high-level, what is stated in this algorithm is an efficient way of reducing the computational complexity of marginalising a joint probability distn

- Rest is details, machinery for doing so. (exact elimination algorithm)
- Elimination algorithm → sum-product → junction-tree algorithm.

PI-708 26/04/2020

- supplementing (4)
- Anything new, interesting, important
- 4 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. $P(x) = Q(x) \; \forall x \in X$
- ~~KO~~ 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 $\longrightarrow$ (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

(?): Distinct from marginal correlation (in regression coefficients?) — this nuance is important

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

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:-

$$R_{ij} = \rho(X_i, X_j | X_{-ij})$$

$$R_{ij} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}}\sqrt{\Theta_{jj}}}$$ where Θ 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 (?)
- shortcut: impose Gaussian assumption on r.v.s. $\rho(X,Y|Z)$ iff $X \perp Y|Z$ i.e. (X,Y,Z) jointly Gaussian