Uncertainty in Artificial Intelligence - Cheat sheet

Dieter Castel & Pierre Carbonnelle

August 20, 2020

1 General Probability

Formula	Comment
$p(x,y,z) = p(x \wedge y \wedge z) = p(x,z\|y)p(y)$	Joint Probability Distribution (JPD)
$p(x \vee y) = p(x) + p(y) - p(x \wedge y)$	Disjunction for probabilities
$p(x\|y) = \frac{p(x,y)}{p(y)} = \frac{p(y\|x)p(x)}{p(y)}$	definition of Conditional Probablity Dist. (CPD)
$p(\neg x) = 1 - p(x)$	only valid for probability distributions i.e. normalized!
$p(x) = \sum_{x,z} p(x,y,z)$	marginalisation
$\sum_{y,z} p(x\|y,z) = 1$	marginalisation of CPD sums to one.
$size(p(\hat{x},\hat{y},\hat{z})) = \#dom(\hat{x}) * \#dom(\hat{y}) * \#dom(\hat{z})$	**without** independence assumptions
$p(x_1,\ldots,x_n) = p(x_1)p(x_2\|x_1)p(x_3\|x_2,x_1)\ldots p(x_k\|x_{k+1},\ldots,x_n)$	Chain rule

Scientific inference:

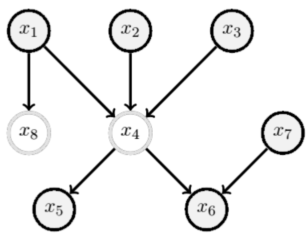
$$\text{Posterior distribution} = p(\Theta|D) = \frac{p(D|\Theta)p(\Theta)}{\int_\Theta p(D|\Theta)p(\Theta)} = \frac{\text{generative Model * Prior}}{\text{Normalization Constant}(=Z)} = \frac{\text{likelihood} * \text{prior}}{\text{evidence}}$$

Point Estimates: Model M, Data D and Parameters Θ	Name (ABRV, comment)
$\Theta_{ML} = argmax_\Theta p(D\|\Theta, M) = argmax_\Theta p(\Theta, D\|M)$	Maximum Likelihood (ML) $(discrete + fullyobsv. = justcounting)$
$\Theta_{MAP} = argmax_\Theta p(\Theta\|D, M) = argmax_\Theta p(D\|\Theta, M)p(\Theta\|M)$	Maximum A Postiori (MAP,)
$\Theta_{MoP} = \langle p(\Theta\|D, M)\rangle$	Mean of Postiori (MoP)

2 Distributional Independence

Formula	Comment
$X \perp\!\!\!\perp Y \iff \forall x \in X, y \in Y : p(x,y) = p(x)p(y)$	marginal distributional independence for variable sets X,Y
$X \perp\!\!\!\perp Y \iff \forall x \in X, y \in Y : p(x\|y) = p(x)$	marginal distributional independence with CPD def.
$X \perp\!\!\!\perp Y\|Z \iff \forall x \in X, y \in Y : p(x\|y,z) = p(x\|z)$	conditional distributional independence.

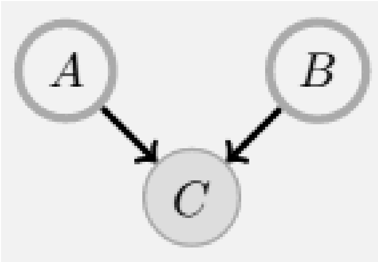
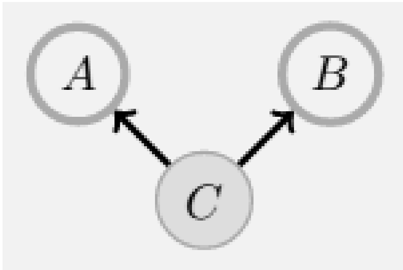
Below is the markov blanket of x_4: parents, childeren and parents of its childeren.



3 Graphical Independence/d-separation

In this section $\perp\!\!\!\perp$ and the like only mean GRAPHICAL (in)dependence, **BEWARE**:

(d-separation, $\perp\!\!\!\perp$, $\not\top\!\!\!\top$ ==) Graphical independence $\Rightarrow$ distributional independence.
(d-connected, $\not\perp\!\!\!\perp$, $\top\!\!\!\top$ ==) Graphical dependence $\not\Rightarrow$ distributional dependence.

colliders in BN	**non-colliders** in BN
	
$A \not\perp\!\!\!\perp B \mid C$ or $A \top\!\!\!\top B \mid C$	$A \perp\!\!\!\perp B \mid C$ or $A \not\top\!\!\!\top B \mid C$

3.1 Path-blocking

$$\forall p \in allpaths(x,y) : blocked(p,Z) \rightarrow isDSeparated(x,y)$$

$$\exists p \in allpaths(x,y) : infoflows(p,Z) \rightarrow isDConnected(x,y)$$

A path p is $blocked(p) \iff (1) \vee (2)$ (d-separation see p43 BRML.)

(1) COLLIDER	NON-COLLIDER (2)
	$u > v > w$
$u > v < w$	$u < v < w$
	$u < v > w$
$\exists v \in p \setminus \{x,y\}$:	$\exists v \in p \setminus \{x,y\}$:
$collider(v) \wedge v \notin Z \wedge descendants(v) \notin Z$	$noncollider(v) \wedge v \in Z$
colliders and descendants in Z lead to upward info flow.	non-colliders equivalent w.r.t. independence

3.2 AMDS on complete graph at once

Quickest way is with graph edits **AMDS**. For variable sets $X, Y, Z : X \perp\!\!\!\perp Y \mid Z$

1. **Ancestral** graph (keep X,Y,Z and ancestors(X,Y,Z)) **A**
2. **Moralize** (add edges between all parents of the same node) ($\forall v \in A marry(parents(v))$) **M**
3. **Disorient** (remove arrows) **D**
4. **Separate** (remove all edges from nodes in Z) **S**

In the final S graph all unconnected nodes are D-separated.

4 Independence Identities

symmetry	**decomposition**	**weak union**	**contraction**
$A \perp\!\!\!\perp B \mid C$	$A \perp\!\!\!\perp B, C$	$A \perp\!\!\!\perp B, C$	$A \perp\!\!\!\perp B \wedge A \perp\!\!\!\perp C \mid B$
$\iff$	$\Downarrow$	$\Downarrow$	$\Downarrow$
$B \perp\!\!\!\perp A \mid C$	$A \perp\!\!\!\perp B \wedge A \perp\!\!\!\perp C$	$A \perp\!\!\!\perp B \mid C \wedge A \perp\!\!\!\perp C \mid B$	$A \perp\!\!\!\perp B, D$

Graphical networks are **Markov Equivalent** $\iff$ same independencies $\iff$ same skeleton $\wedge$ same immoralities.
L_p set of independencies in JPD P. L_G set of independencies in graph.

I-map (all independencies hold)	D-Map (all dependencies hold)
$L_g \subseteq L_p$	$L_p \subseteq L_g$
EXCLUSION: Look for an independence in $L_g \notin L_p$	**EXCLUSION:** Look for independence in $L_p \notin L_g$
$\Rightarrow$ *NOT* I-map	$\Rightarrow$ *NOT* D-map

5 General Inference

6 Hidden Markov Models (HMM)

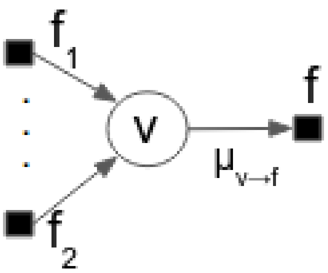
Formula	Comment
$p(v_{1:t}, h_{1:t}) = p(h_1) * \Pi_{i=2}^{t} p(v_t\|h_t)p(h_t\|h_{t-1})$	JPD for an HMM
$p(v_t\|h_t)(= p(v_1\|h_1) \iff$ stationary HMM)	emission matrix
$\forall t p(v_t\|h_t) = p(v_1\|h_1)$	emission matrix for stationary HMM
$p(h_t\|h_{t-1})$	transmission matrix
	Inference in HMMS
$p(h_t\|v_{1:t})$	Filtering (infer up to t)
$p(h_t\|v_{1:T})(T > t)$	Smoothing (use future too)
$argmax_{h_1:T} p(h_{1:T}\|v_{1:T})$	Viterbi (most likely state)

7 Sum-Product on Factor Graphs

Singly connected variant BRML p.81. Loopy? Remove problematic node R, SP over new graph, finally: sum/max over the states of removed node R.

Variable-to-Factor
The set F are all factors in the image

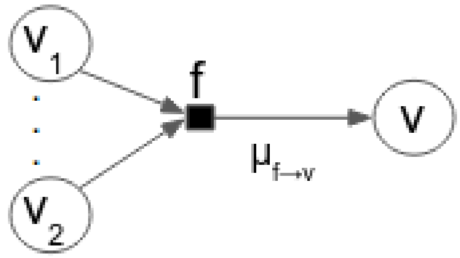
$\mu_{v\to f}(v) = \Pi_{f_i \in F \setminus f} \mu_{f_i \to v}(v)$
product of all factors, $func(v)$



V2Factor = ONLY **FACTORS**
$\mu_{v\to f}(v)$ is a function of v
Leaf Factors = $f_i(v)$

Factor to Variable
The set V are all variables in the image.

$\mu_{f\to v}(v) = \sum_{v_i \in V \setminus v} f(v_1, \dots, v_i, v)\Pi_{y\in\{ne(f)\setminus x\}}$
sum-product over non-v variables, $func(v)$



F2Variable = **Sum/Max/Argmax AND FACTORS**
$\mu_{f\to v}(v)$ is a function of v
Leaf Variables = 1

1. Make factor graph
2. Pick Root node.
3. Set leafs (Leaf factor = factor, Leaf Variable = 1)
4. Propagate messages until required value is computed (up till root for marg inference, backtracking for argmax)

8 Bucket Elimination

To calculate $p(x_k)$ from $p(x_1, \dots, x_k, \dots, x_n)$:

1. Pick and ordering ending with x_k
2. Set all buckets to 1 distribute all factors in order to the buckets
3. Eliminate top to bottom: $\sum_v p(v|x_i, \dots, x_j) = 1$ or redistribute summed bucket to first remaining bucket.
4. Final bucket is required marginal and still a function of that variable.

Don't sum over evidential variable.

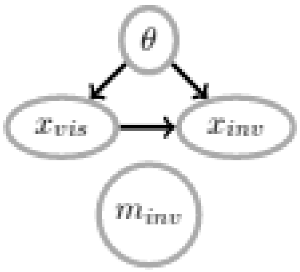
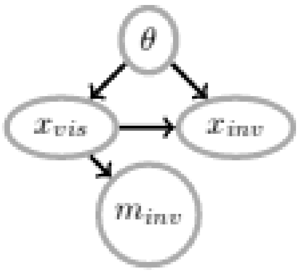
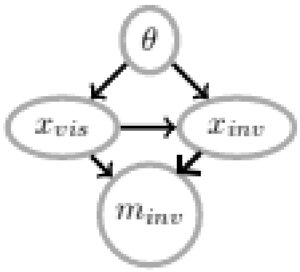
9 Distributions

Beta distribution = 'conjugate' of binomial dist. and conjugate of Beta dist. p.183 BRML; More about Beta Dist see p. 173 BRML

formula	comment
prior : $B(\alpha, \beta), x \in 0, 1 \Rightarrow$ posterior $= B(\alpha + \#_1^x, \beta + \#_0^x)$	Posterior with with Beta-disribution for binary variable x

10 Partially Observable Data - Learning with hidden variables

As opposed to fully observable data. Strategy depends on **missingness assumptions**. Direction between the visible variables $v_{vis}^i \in x_{vis} = x_{obs}$ and $h_{inv}^i \in x_{inv}$ below is irrelevant.

Missing Completely At Random (MCAR)	**Missing At Random** MAR	**Missing Not At Random** (MNAR)
		
$x_i \perp\!\!\!\perp m_i$ marginal idp between var. and its missingness var.	$x_i \perp\!\!\!\perp m_i \mid x_{vis}$ Conditional idp between var. and its missingness var. $\sum_{h_i} \ldots = \sum_{x_{inv}} \ldots$ couples vars.	$x_i \top\!\!\!\top m_i$ dependence between var. and its missingness. **NON-identifiable**
$\Downarrow$ Delete samples with missing data & regular $\Theta_* ML/...$	$\Downarrow$ **LEARNING SOLUTION** $\Downarrow$ m_i factored out + sol. 4 coupling = Expectation Maximisation (EM)	$\Downarrow$ *(out of scope of UAI)* Requires EXTRA assumptions

11 Expectation Maximisation

Learning method under MAR assumption **guaranteed** to converge.

1. guess Θ^0 (n=0 start with uniform distributions or randomly)
2. until convergence(Θ^{n-1}, Θ^n) do:
   E: compute $\forall i, k : q_k^i = p(h^i = values(k) | v^i, D, \Theta^{n-1})$ Θ^{n-1}-weighted estimate of (completed) data for all missing var.
   M: compute $\Theta_* ML$ given E by weighted counting. Probabilistic counting using q_k^i's from **E**
   $n+ = 1$

12 sampling

13 Importance "sampling" = approximate averaging

Calculates $\langle f(x) \rangle_{p(x)}$ given a known importance distribution $q(x)$ (S_q are samples of $q(x)$) & evaluable function $p*(x)$ such that $p(x) = p*(x)/Z$ and :

$$\langle f(x) \rangle_{p(x)} = \sum_{x^l \in S_q} f(x^l) w(x^l) \text{ where } w^l = w(x^l) = \frac{p^*(x^l)/q(x^l)}{\sum_{x^l} p^*(x^l)/q(x^l)} \text{ and } \sum_{x_l} w^l = 1$$

14 soft and unreliable evidence

15 global/local semantics

A LaTeXcommands

Go over all tex files see what sticks.

B Credits

Original google docs by P. Carbonelle to be found here:
https://docs.google.com/presentation/d/1sP-PJmo-pW4epfLQs3zoveZGceqw59pPeNAY9OM6dl8/edit#slide=id.p

Some images taken from the Bayesian Reasoning and Machine Learning book. Found on this URL: http://www.cs.ucl.ac.uk/staff/d.barber/brml/