DIRECTED GRAPHICAL MODELS

COMPACTLY REPRESENT JOINT DISTRIBUTION $P(x|\theta)$, HOW TO INFER GIVEN SET OF VARS GIVEN MOTHER, HOW TO LEARN DISTRIBUTION PARAMS

CHAIN RULE: $P(x_{1:V}) = P(x_1)P(x_2|x_1)P(x_3|x_2,x_1)\dots$ CONDITIONAL PROBABILITY TABLES: $O(K^V)$ SPACE, A LOT!

CONDITIONAL PROBABILITY DISTRIBUTION: $O(K^2V^2)$ SPACE, BETTER! BUT STILL DEPENDS ON ALL PREVIOUS VARS

CONDITIONAL INDEPENDENCE / MARKOV ASSUMPTION CONDITIONAL JOINT AS PRODUCT OF MARGINALS $P(x_{1:V}) = P(x_1)\prod_t^V P(x_t|x_{t-1})$

GRAPHICAL MODEL: MODELS CI IN > 1D, LACK OF EDGES → CI

- (GRAPH) NEIGHBORS → SET OF DIRECTLY CONNECTED NODES
- (GRAPH) CLIQUE → SET OF NODES ALL NEIGHBORS OF EACH OTHER

MARKOV CHAIN: INITIAL DISTRIBUTION + STATE TRANSITION MATRIX

DIRECTED GRAPHICAL MODEL: A GM WHOSE GRAPH IS A DAG. ALSO BAYESIAN NETWORKS / BELIEF NETWORKS / CAUSAL NETWORKS

- ORDERED MARKOV PROPERTY A NODE ONLY DEPENDS ON ITS IMMEDIATE PARENTS $x_s \perp x_{PRED(s)} | x_{PA(s)}$ EACH $P(x_t|x_{pa(t)})$ IS CPD
- GENERALLY: $P(1:V|\theta) = \prod_t^V P(x_t|x_{pa(t)})$, $O(VK^F)$ SPACE, F = PARENTS K = STATES

NAIVE BAYES

CAN BE SEEN AS DGM. FEATURES ARE ASSUMED CI FROM LABELS $P(y|x) = P(y)\prod_j^D P(x_j|y)$

TREE-AUGMENTED NAIVE BAYES: CAPTURES CORRELATION BETWEEN FEATURES, GRAPH IS A TREE, CAN FIND OPTIMAL STRUCTURE. CAN HANDLE MISSING FEATURES

MARKOV AND HIDDEN MARKOV MODELS

- 2ND ORDER MARKOV CHAIN = $P(1:T) = P(x_1,x_2)P(x_3|x_1,x_2)P(x_4|x_2,x_3) = P(x_1,x_2)\prod_t P(x_t|x_{t-1},x_{t-2})$
- IF EXTEND TO MUCH NUMBER OF PARAMS BLOWS UP
- HIDDEN MARKOV MODEL: HIDDEN VARIABLES z_t, OBSERVED VARIABLES x_t
  - TRANSITION MODEL $P(z_t|z_{t-1})$ - OBSERVATION MODEL $P(x_t|z_t)$ - STATE ESTIMATION $P(z_t|x_{1:t},\theta)$ ESTIMATE HIDDEN STATE

NOISY-OR MODEL / SIGMOID BELIEF NET

IF A PARENT ON → CHILD USUALLY ON, BUT OCCASIONALLY PARENT → CHILD FAILS $\theta_{ST} = 1 - q_{ST}$ PROB FAILURE

PROB CHILD = OFF $P(v_t=0|h) = \prod \theta_{st}^{I(h_s=1)}$ • DUMMY LEAK NODE: CHILD ON IF ALL PARENTS OFF θ_{0t}

• IF $w_{st} = \log(\theta_{st})$ $P(v_t=1|h) = 1 - \exp\left(w_{0t} + \sum h_s w_{st}\right)$ SIMILAR TO LOGISTIC REGRESSION

DIRECTED GAUSSIAN GRAPHICAL MODELS

DGM WHERE ALL VARS ARE REAL VALUED, CPD $P(x_t|x_{pa(t)}) = N(x_t|\mu_t + w^T x_{pa(t)}, \sigma_t^2)$ LINEAR GAUSSIAN CPD

GAUSSIAN BAYES NET: $P(x) = N(x|\mu,\Sigma)$ μ: $x_t = \mu_t + \sum w_{ts}(x_s - \mu_s) + \sigma_t z_t$ $z_t \sim N(0,1)$, σ_t CONDITIONAL STDDEV OF x_t GIVEN PARENTS

$\rightarrow \mu = (\mu_1 \dots \mu_D)$ w_{ts} EDGE WEIGHT, μ_t LOCAL MEAN

Σ: $(x-\mu) = W(x-\mu) + S_z$, $S = \text{DIAG}(\sigma)$; $E = S_z$ = VECTOR OF NOISE TERMS $\rightarrow x - \mu = U \cdot S \cdot z$, $U = (I-W)^{-1}$ CHOLESKY DECOMPOSITION

$\rightarrow \Sigma = U \cdot S^2 \cdot U^T$

INFERENCE

WE HAVE $P(x_{1:n}|\theta)$ AND θ ARE UNKNOWN. VISIBLE VARS X_v HIDDEN VARS X_h. COMPUTING THE POSTERIOR OF THE UNKNOWNS GIVEN THE KNOWNS

$$P(X_h|X_v,\theta) = \frac{P(x_h,x_v|\theta)}{P(x_v|\theta)} = \frac{P(x_h,x_v|\theta)}{\sum P(x'_h,x_v|\theta)}$$

- WE CONDITION ON DATA BY CLAMPING ~~EVIDENCE~~ VISIBLE VARS TO OBSERVED VALUES X_v AND NORMALIZE ON DATA LIKELIHOOD / EVIDENCE
- IF ONLY SOME OF HIDDEN VARS ARE INTERESTING, OTHERS ARE MARGINALIZED OUT

$$P(X_q|X_v,\theta) = \sum_{X_N} P(X_q,X_N|X_v,\theta) \quad X_q = \text{INTERESTING} \quad X_N = \text{NUISANCE}$$

LEARNING

$$\hat{\theta} = \underset{\theta}{\text{ARGMAX}} \sum_{i=1}^{N} \log P(x_{i,v}|\theta) + \log P(\theta)$$

MAP ESTIMATE OF PARAMS GIVEN DATA

- UNIFORM PRIOR: MLE
- IF BAYESIAN: PARAMS ARE ALSO UNKNOWN VARS. NO INFERENCE/LEARNING DISTINCTION WE ADD PARAMS TO GRAPH AS NODES, CONDITION ON D, THEN INFER VALUES. HOWEVER PARAM VARS DO NOT GROW WITH DATA, WE CAN POINT ESTIMATE THEM.

- PLATE NOTATION FOR CONVENIENT DRAWING, NESTED PLATES, ROLLED/UNROLLED REPRESENTATIONS

WHEN DATA IS COMPLETE:

- LIKELIHOOD: $P(D|\theta) = \prod_{t=1}^{V} P(D_t|\theta_t)$ DATA ASSOCIATED WITH NODE t FAMILY (t + PARENTS). LIKELIHOOD DECOMPOSES ACCORDING TO GRAPH STRUCTURE
- PRIOR: $P(\theta) = \prod_{t=1}^{V} P(\theta_t)$
- POSTERIOR: $P(\theta|D) = \prod_{t=1}^{V} P(D_t|\theta_t)P(\theta_t)$ EACH CPD IS INDEPENDENT. • FACTORED PRIOR + FACTORED LIKELIHOOD ⟶ FACTORED POSTERIOR

MISSING DATA/LATENT VARIABLES: LIKELIHOOD NO LONGER FACTORIZES. NO LONGER CONVEX. ONLY LOCAL ML/MAP ESTIMATES. APPROX. INFERENCE

CI PROPERTIES

$I(G)$ = SET OF CI STATEMENTS ENCODED IN GRAPH. G IS I-MAP FOR P IFF $I(G) \subseteq I(P)$. IF GRAPH DOESN'T ASSERT FALSE STATEMENTS ABOUT P.

- A FULLY CONNECTED GRAPH IS IMAP FOR ALL DISTRIBUTIONS

D-SEPARATION

- A PATH IS D-SEPARATED BY A SET OF NODES E (FOR EVIDENCE) IFF EITHER
  - P CONTAINS A CHAIN WHERE MIDDLE $\in E$ $S \rightarrow M \rightarrow T$
  - P CONTAINS FENT/FORK WITH $M \in E$ IS VERTEX
  - P CONTAINS A COLLIDER/V-STRUCT WHERE M IS VERTEX, $M \notin E$, NOR ANY DESCENDANT OF M
- A SET A IS D-SEPARATED FROM SET B BY SET E IFF EACH (UNDIR) PATH FROM $\forall a \in A$ TO $\forall b \in B$ IS D-SEP BY E

$$X_A \perp_G X_B \mid X_E \longleftrightarrow A \text{ IS D-SEP FROM } B \text{ GIVEN } E$$

C.I. ⟷ D-SEP

DIRECTED GLOBAL MARKOV PROPERTY

- CONDITIONING ON BOTTOM OF V/COLLIDER MAKES PARENTS DEPENDANT → EXPLAINING AWAY/INTER-CAUSAL REASONING/BERKSON'S PDOX

IE IF YOU OBSERVE THE SUM OF TWO VALUES, KNOWING ONE GIVES YOU THE OTHER.

ALGO: BAYES BALLS

DIRECTED LOCAL MARKOV PROPERTY: $t \perp ND(t) \mid PA(t)$ ND: NON-DESCENDANTS ORDERED MARKOV PROPERTY: $t \perp PRED(t) \mid PA(t)$

- THESE PROPS ARE EQUIVALENT.

MARKOV BLANKET

SET OF NODES RENDERING NODE t CI FROM ALL OTHER NODES IN GRAPH. MB IS PARENTS, CHILDREN, AND CO-PARENTS (ALSO NODES PARENTS OF ITS CHILDREN)

FULL CONDITIONAL

FULL PRODUCT OF CPD WITH X_t IN THEIR SCOPE $P(x_t|X_{-t}) \propto P(x_t|x_{pa(t)}) \prod_{s \in children(t)} P(x_s|x_{pa(s)})$

INFLUENCE/DECISION DIAGRAM

IT'S A DAG + UTILITY NODES, DECISION NODES. OIL-WILD-CATTER PROBLEM.

- EXPECTED UTILITIES — POSTERIORS - EXHAUST THEM FOR ALL SCENARIOS → COMPUTE OPTIMAL POLICIES
- MAX EXPECTED UTILITY: $MEU = \max_{\delta} \sum_{s} P(s)\, EU(d^{\delta}(s)|s)$
- VALUE OF PERFECT INFORMATION: Δ UTILITY DUE TO GAINING MORE INFO. CAN HAVE COSTS, USED TO SET POLICIES/THRESHOLDS
  - VPI: $MEU(I+T \rightarrow D) - MEU(I)$

POMDP - PARTIALLY OBSERVED MARKOV DECISION PROCESS

IT'S A HMM WITH AUGMENTED ACTION AND REWARD NODES → USED FOR PERCEPTION - ACTION CYCLES IN INTELLIGENT AGENTS

MDP - MARKOV DECISION PROCESS. LIKE POMDP BUT FULLY OBSERVED. EASIER TO SOLVE

↳ STATE NOT FULLY OBSERVED
CHOOSE ACTIONS BASED ON BELIEF STATE
$P(z_t|x_t, a_{1:t})$