

6.438 Final Exam

Collaborators: rdeits

Gaussian Random Variables

$$\underline{x} \sim N(\underline{\mu}, \Lambda)$$

$$E[\underline{x}] = \underline{\mu}$$

$$E[(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})^T] = \Lambda$$

$$p_x(\underline{x}) = |2\pi\Lambda|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}[(\underline{x} - \underline{\mu})^T \Lambda^{-1}(\underline{x} - \underline{\mu})]\right]$$

Transformations

$$\underline{z} \sim N(\underline{\mu}_z, \Lambda_z) \quad \underline{x} = A\underline{z} + \underline{b} \rightarrow \underline{x} \sim N(\underline{\mu}_x, \Lambda_x)$$

$$\underline{\mu}_x = A\underline{\mu}_z + \underline{b} \quad \Lambda_x = A\Lambda_z A^T$$

Gaussian Information Form

$$p(x) = \frac{1}{Z} \exp\left[-\frac{1}{2}x^T J x + h^T x\right]$$

$$J = \Lambda^{-1} \quad h = J\mu$$

Marginalization

$$p(x_1) \sim N(\mu_1, \Lambda_{11}) \quad p(x_1) \sim N(h_1 t, J_{11} t)$$

$$h'_1 = h_1 - J_{12} J'_{22} h_2 \quad J'_{11} = \Lambda_{11}^{-1} = J_{11} - J_{12} J_{22}^{-1} J_{21} \quad (\text{Schur complement})$$

Conditioning

$$p(x_1|x_2) \sim N^{-1}(h''_1, J''_{11})$$

$$J''_{11} = J_{11} \quad h''_1 = h_1 - J_{12} x_2$$

$$\mu''_1 = \mu_1 + \Lambda_{12} \Lambda_{22}^{-1} (x_2 - \mu_2) \quad \Lambda''_{11} = \Lambda_{11} - \Lambda_{12} \Lambda_{22}^{-1} \Lambda_{21}$$

Dependencies

G is an I-map of P if $CI(G) \subseteq CI(P)$ e.g. G is fully connected

G is a D-map of P if $CI(G) \supseteq CI(P)$ e.g. G is unconnected

G is a P-map of P if $CI(G) = CI(P)$

G is a minimal I-map if removing any edge would make it no longer an I-map

A directed model turned into a moralized undirected model is a P-map if moralization adds no edges.

Undirected graph G has a directed P-map iff G is chordal

Directed graph H has an undirected P-map iff moralization adds no edges

Variable Elimination

$$m_i(x_{s_i}) = \sum_{x_i} \prod_{\varphi_i \in \Psi} \varphi_i(x_i, x_{s_i})$$

Sum-Product

$$p(x) = \prod_i \phi_i(x_i) \prod_{(i,j) \in E} \psi_{ij}(x_i, x_j)$$

$$m_{i \rightarrow j}^{t+1}(x_j) = \sum_{x_i} \phi_i(x_i) \psi_{ij}(x_i, x_j) \prod_{k \in N(i) \setminus \{j\}} m_{k \rightarrow i}^t(x_i)$$

$$p_{x_i}(x_i) \propto \phi_i(x_i) \prod_{k \in N(i)} m_{k \rightarrow i}(x_i)$$

Forward-Backward Probabilities

Markov chain with nodes $(x_1, \dots, x_N, \hat{y}_1, \dots, \hat{y}_N)$

$$\underbrace{p(y_{i+1}|x_{i+1}) m_{i \rightarrow i+1}(x_{i+1})}_{\alpha_{i+1}(x_{i+1})} = \sum_{x_i} p(x_{i+1}|x_i) \underbrace{m_{y_i \rightarrow x_i}(x_i) m_{i-1 \rightarrow i}(x_i)}_{\alpha_i(x_i)}$$

$$\underbrace{m_{i+1 \rightarrow i}(x_i)}_{\beta_i(x_i)} = \sum_{x_{i+1}} p(x_{i+1}|x_i) p(\hat{y}_{i+1}|x_{i+1}) \underbrace{m_{i+2 \rightarrow i+1}(x_{i+1})}_{\beta_{i+1}(x_{i+1})}$$

$$p(x_i|\hat{y}_1, \dots, \hat{y}_N) = \frac{\alpha_i(x_i) \beta_i(x_i)}{\sum_{x'_i} \alpha_i(x'_i) \beta_i(x'_i)}$$

Sum-Product for Factor Tree

Factor \rightarrow node:

$$m_{a \rightarrow j}(x_j) = \sum_{x_k, k \in N(a) \setminus \{j\}} f_a(x_{N(a)}) \prod_{k \in N(a) \setminus \{j\}} m_{k \rightarrow a}(x_k)$$

Node \rightarrow factor:

$$m_{j \rightarrow a}(x_j) = \prod_{b \in N(j) \setminus \{a\}} m_{b \rightarrow j}(x_j)$$

Kalman Filtering

$$x_{t+1} = Ax_t + v_t, v \sim N(0, Q), x_0 \sim N(0, \Lambda_0)$$

$$y_t = Cx_t + w_t, w_t \sim N(0, R)$$

$$x_0 \sim N(0, \Lambda_0)$$

$$x_{t+1}|x_t \sim N(Ax_t, Q)$$

$$y_t|x_t \sim N(Cx_t, R)$$

Filtering

$$\alpha(x_{i+1}) = \int \alpha(x_i) p(x_{i+1}|x_i) p(y_{i+1}|x_{i+1}) dx_i$$

Prediction

$$\mu_{i+1|i} = A\mu_{i|i}$$

$$\Sigma_{i+1|i} = A\Sigma_{i|i}A^T + Q$$

$$\mu_{0|-1} = 0$$

$$\Sigma_{0|-1} = \Lambda_0$$

Update

$$\mu_{i+1|i+1} = \mu_{i+1|i} + G_{i+1}(y_{i+1} - C\mu_{i+1|i})$$

$$\Sigma_{i+1|i+1} = \Sigma_{i+1|i} - G_{i+1}C\Sigma_{i+1|i}$$

$$G_{i+1} = \Sigma_{i+1|i} C^T (C\Sigma_{i+1|i} C^T + R)^{-1}$$

Smoothing

$$\gamma(x_i) = \int \gamma(x_{i+1}) \left[\frac{\alpha(x_i) p(x_{i+1}|x_i)}{\int \alpha(x'_i) p(x_{i+1}|x'_i) dx'_i} \right] dx_{i+1}$$

$$\gamma(x_i) = \frac{\alpha(x_i) \beta(x_i)}{p(y_0^t)}$$

$$\mu_{i|t} = \mu_{i|i} + F_i(\mu_{i+1|t} - \mu_{i+1|i})$$

$$\Sigma_{i|t} = F_i(\Sigma_{i+1|t} - \Sigma_{i+1|i}) F_i^T + \Sigma_{i|i}$$

$$F_i = \Sigma_{i|i} A^T \Sigma_{i+1|i}^{-1}$$

Junction Trees

If a graph is chordal, then it has a junction tree.

Loopy BP

If we have a graph $\mathcal{G} = (V, E)$ with

$$p_x(x) \propto \prod_{i \in V} \exp(\phi_i(x_i)) \prod_{(i,j) \in E} \exp(\psi_{ij}(x_i, x_j))$$

$$m_{i \rightarrow j}^{t+1}(x_j) \propto$$

$$\sum_{x_i \in \mathcal{X}} \exp(\phi_i(x_i)) \exp(\psi_{ij}(x_i, x_j)) \prod_{k \in N(i) \setminus \{j\}} m_{k \rightarrow i}^t(x_i)$$

Node and edge marginals:

$$b_i^t(x_i) \propto \exp(\phi_i(x_i)) \prod_{k \in N(i)} m_{k \rightarrow i}^t(x_i)$$

$$b_{ij}^t(x_i, x_j) \propto \exp(\phi_i(x_i) + \phi_j(x_j) +$$

$$\psi_{ij}(x_i, x_j)) \prod_{k \in N(i)} m_{k \rightarrow i}^t(x_i) \prod_{\ell \in N(j)} m_{\ell \rightarrow j}^t(x_j)$$

Variational Methods

K-L Divergence

$$KL(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Bethe approximation

Constraints

$$\mu(x) = \prod_{i \in V} \mu_i(x_i) \prod_{(i,j) \in E} \frac{\mu_{ij}(x_i, x_j)}{\mu_i(x_i) \mu_j(x_j)}$$

$$\mu_i(x_i) \geq 0$$

$$\sum_{x_i \in \mathcal{X}} \mu_i(x_i) = 1$$

$$\mu_{ij}(x_i, x_j) \geq 0$$

$$\sum_{x_j \in \mathcal{X}} \mu_{ij}(x_i, x_j) = \mu_i(x_i)$$

$$\sum_{x_i \in \mathcal{X}} \mu_{ij}(x_i, x_j) = \mu_j(x_j)$$

Mean Field

$$\mu(x) = \prod_{i \in V} \mu_i(x_i)$$

$$\mu_i^{t+1}(x_i) \propto \exp \left[\phi_i(x_i) + \sum_{j \in N(i)} \sum_{x_j \in \mathcal{X}} \mu_j^t(x_j) \psi(x_i, x_j) \right]$$

Variational Objective

$$\text{Maximize: } \mathcal{F}(\mu) = \sum_{x \in \mathcal{X}^N} \mu(x) \theta(x) - \sum_{x \in \mathcal{X}^N} \mu(x) \log \mu(x)$$

$$\text{where } P(x) = \frac{1}{Z(\theta)} e^{\theta(x)}$$

Sampling

Markov Chain Monte Carlo

Metropolis-Hastings

Require a reversible Markov chain:

$$P(x)P(x \rightarrow x') = P(x' \rightarrow x)P(x')$$

and regular: $\exists n$ s.t. $P(X(n) = x' | X(0) = x) > 0$

Proposal distribution: $K(x \rightarrow x')$

Prob. of accepting a move from $x \rightarrow x'$:

$$A(x \rightarrow x') = \min \left[1, \frac{K(x' \rightarrow x)P(x')}{K(x \rightarrow x')P(x)} \right]$$

$$P(x \rightarrow x') = K(x \rightarrow x')A(x \rightarrow x')$$

$$P(x \rightarrow x) = 1 - \sum_{x' \neq x} K(x \rightarrow x')A(x \rightarrow x')$$

Gibbs Sampling

Subclass of M-H with $A(x \rightarrow x') = 1$

0) select any x

1) pick k at random

2) Sample $x_k' \sim P(x_k | x_{-k}) = P(x_k | x_{N(k)})$

Importance Sampling

Instead of P, sample from q using weighting $w^k = \frac{P(x^k, y)}{q(x^k)}$

$$\frac{\sum_k w^k f(x^k)}{\sum_k w^k} = E_{x \sim P_{x|y}} f(x)$$

Particle Filtering

samples = $k \in 1, \dots, K$

$$x_0^k \sim p_{x_0}(\cdot)$$

$$w_0^k = \frac{1}{K} p_{y_0|x_0}(y_0|x_0)$$

$$x_{n+1}^k \sim p_{x_{n+1}|x}(x_{n+1}^k | x_n^k)$$

$$w_{n+1}^k = w_n^k \times p_{y_{n+1}|x_{n+1}}(y_{n+1} | x_{n+1}^k)$$

Bayesian Estimation

Treat θ as a random variable

Dirichlet prior: $P(\theta) = \frac{1}{Z} \prod_x \theta_x^{\alpha_x - 1}$

$$Z = \frac{\prod_x \Gamma(\alpha_x)}{\Gamma(\sum_x \alpha_x)}$$

Inferring Structure

Bayesian Information Criterion

$$\ell(\hat{\theta}^{ML}; D) - \frac{\text{num. params}}{2} \log n$$

$$\text{approximates } \log P(D; G) = \log \int P(\theta) L(\theta; D) d\theta$$

$$\text{score}(G) = \sum_{i=1}^N \text{score}(i | \text{pa}_i; D) = \log \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ijk})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})} \text{ for}$$

BN

Learning Models

Maximum Likelihood estimation

$$\hat{\theta}^{ML} = \arg \max_{\theta} L(\theta; x)$$

Mutual information:

$$I(u; v) \triangleq \sum_{u,v} p_{u,v}(u, v) \log \frac{p_{u,v}(u, v)}{p_u(u)p_v(v)}$$

$$H(u) \triangleq - \sum_u p_u(u) \log p_u(u) \geq 0$$

$$\hat{\ell}(G, D) = \sum_{i=1}^N \hat{I}(x_i; x_{\pi_i}) - \sum_{i=1}^N \hat{H}(x_i)$$

Expectation Maximization Algorithm

$y = (y_1, \dots, y_N)$ observed

$x = (x_1, \dots, x_{N'})$ latent

assume we know $p_{y,x}(\cdot, \cdot; \theta)$ w/ param θ , want $\hat{\theta}^{ML}$

$\ell(\theta; y)$ incomplete log likelihood

$\ell(\theta; y, x)$ complete log likelihood

Choose distribution q over x : $q(\cdot | y)$

$$\ell(\theta; y) \geq \sum_x q(x | y) \log \frac{p_{y,x}(y, x; \theta)}{q(x | y)} \triangleq \tilde{\ell}(q, \theta), \text{ maximize } \tilde{\ell}(q, \theta)$$

E-step: $q^{(i+1)} = \arg \max_q \tilde{\ell}(q, \theta^{(i)})$

M-step: $\theta^{(i+1)} = \arg \max_{\theta} \tilde{\ell}(q^{(i+1)}, \theta)$

Solving those steps gives:

E-step: $q^{(i+1)} = p_{x|y}(\cdot | y; \theta^{(i)})$

M-step: $\theta^{(i+1)} = \arg \max_{\theta} \mathbb{E} [\log p_{y,x}(y, x; \theta) | \mathbf{y} = y; \theta^{(i)}]$

Estimating Undirected Models

$$\ell(\theta; D) = \frac{1}{n} \sum_{t=1}^n \log P(x^t; \theta)$$

$$\hat{p}_c(x_c) = \frac{1}{n} \sum_{t=1}^n \mathbf{1}(x_c, x_c^t)$$

$$p_c(x_c; \theta) = \frac{\partial}{\partial \theta_c(x_c)} \log Z(\theta)$$

Iterative Proportionality Fitting

For each clique c , eval $P_c^{(i)}(x_c)$

$$P^{(i+1)}(x_1, \dots, x_N) = P^{(i)}(x_1, \dots, x_N) \frac{\hat{p}_c(x_c)}{p_c^{(i)}(x_c)}$$

Closed-form solution for a chordal graph:

$$p(x) = \frac{\prod_c \psi_c(x_c)}{\prod_s \phi_s(x_s)} = \frac{\prod_c \hat{p}_c(x_c)}{\prod_s \hat{p}_s(x_s)} \text{ with cliques } c \text{ and separators } s$$