Session 6 Reading 3/1/2022

PRML :

- 4.4 on Laplace Approx (skip 4.4.1)
- 13 on models for sequential data
  - skim : 13.2.1 , 13.2.3
  - skip : 13.2.4 , 13.2.6 , 13.3.2-4

4.4 The Laplace Approximation

Laplace approx: framework to find a Gaussian approx to a probability density over a set of continuous variables

- consider a single continuous var. z

$$p(z) = \frac{1}{Z} f(z) \quad \text{where} \quad \underbrace{Z}_{\text{value is unknown}} = \int f(z)\, dz$$ "normalization coef."

- goal : find a Gaussian approx. $q(z)$ which is centered on a mode of $p(z)$

- start : find a mode of $p(z)$; ie. find a z_0 s.t. $p'(z_0) = 0$:

$$\left.\frac{df(z)}{dz}\right|_{z=z_0} = 0$$

- the log of a Gaussian is a quadratic fxn of the variables so can use Taylor expansion:

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2} A (z - z_0)^2 \qquad A = -\left.\frac{d^2}{dz^2} \ln f(z)\right|_{z=z_0}$$

$$f(z) \simeq f(z_0) \exp\left[-\frac{A}{2}(z - z_0)^2\right]$$

- normalized distribution $q(z)$ from the standard normalization of a Gaussian:

$$q(z) = \left(\frac{A}{2\pi}\right)^{\frac{1}{2}} \exp\left[-\frac{A}{2}(z-z_0)^2\right]$$

- extend Laplace method over M-dimensional space $\mathbf{z}$

$$p(\mathbf{z}) = \frac{1}{Z} f(\mathbf{z})$$

- stationary point $\mathbf{z}_0$ where $\nabla f(\mathbf{z}) = 0$

$$\ln f(\mathbf{z}) \simeq \ln f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T \mathbf{A} (\mathbf{z}-\mathbf{z}_0)$$

where $\mathbf{A}$ is a $M \times M$ Hessian matrix: $\mathbf{A} = -\nabla\nabla \ln f(\mathbf{z})\big|_{\mathbf{z}=\mathbf{z}_0}$

"square matrix of 2nd order partial derivatives of a scalar valued function"

$$f(\mathbf{z}) \simeq f(\mathbf{z}_0) \exp\left[-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T \mathbf{A} (\mathbf{z}-\mathbf{z})\right]$$

- $q(\mathbf{z})$ is proportional to $f(\mathbf{z})$
- can find normalization coef w/ standard result for a normalized multivariate Gaussian

$$q(\mathbf{z}) = \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}} \exp\left[-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T \mathbf{A} (\mathbf{z}-\mathbf{z}_0)\right]$$

$|\mathbf{A}|$: determinate of $\mathbf{A}$

$$= N(\mathbf{z} \mid \mathbf{z}_0, \mathbf{A}^{-1})$$

13. Sequential Data

- Markov models: assume that future predictions are independent of all but the most recent observations
- we'll focus on two "state space models": hidden Markov model (HMM), linear dynamical system (LDS)

13.1 Markov models

- relax assumption of i.i.d data
- product rule to express joint distribution for a sequence of observations:

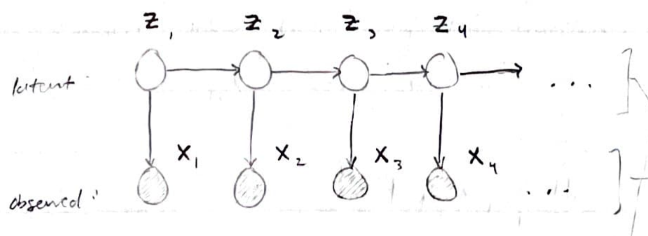
$$p(x_1, \ldots, x_N) = \prod_{n=1}^{N} p(x_n \mid x_1, \ldots, x_{n-1})$$

- assume x_n is independent of all but most recent obs x_{n-1} → "first-order Markov chain"

$$p(x_1, \ldots, x_N) = \prod_{n=2} p(x_n \mid x_{n-1})$$

- 2^{nd} order MC: $p(x_1, \ldots, x_N) = \prod_{n=3}^{N} p(x_n \mid x_{n-1}, x_{n-2})$
- autoregressive (AR) model: for continuous variables, use linear Gaussian conditional distributions where each node is a Gaussian dist. w/ mean as a linear function of the parents

- want to build a model w/o a specific Markov order but still limited num. of free parameters
  - need to introduce some latent variables
- for each observation x_n, introduce a corresponding latent var. z_n of lower dim
  - form the MC w/ the latent variables to form a "state space model"



$$p(x_1, \ldots, x_N, z_1, \ldots, z_N) = p(z_1)\left[\prod_{n=2}^{N} p(z_n \mid z_{n-1})\right]\left[\prod_{n=1}^{N} p(x_n \mid z_n)\right]$$

13.2 Hidden Markov Models (skim: 13.2.1 & 13.2.3; skip 13.2.4, 13.2.6)

- state space model w/ discrete latent variables
  - can be interpreted as an extension of a mixture model w/ choice of mixture component dependent on component of previous observation
- latent variables are discrete multinomial variables $\mathbf{z}_n$ w/ 1-of-K coding scheme
  - describes which component produced observation $\mathbf{x}_n$
- prob. of $\mathbf{z}_n$ is dependent on $\mathbf{z}_{n-1}$: $p(\mathbf{z}_n | \mathbf{z}_{n-1})$
  - this conditional dist. corresponds to table $\mathbf{A}_{K \times K}$ containing "transition prob."

$$A_{jk} \equiv p(z_{nk} = 1 \mid z_{n-1,j} = 1) \equiv \text{prob. of going from } j \to k \text{ component}$$

$$p(\mathbf{z}_n | \mathbf{z}_{n-1}, \mathbf{A}) = \prod_{k=1}^{K} \prod_{j=1}^{K} A_{jk}^{z_{n-1,j} z_{nk}}$$

special case for first latent node $\mathbf{z}_1$:

$$p(\mathbf{z}_1 | \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{1k}}$$ where $\boldsymbol{\pi}$ is a vector of prob.: $\pi_k \equiv p(z_{1k} = 1)$ & $\sum_k \pi_k = 1$

- see Fig 13.6 (pg. 611) for a "state transition diagram"
- Fig 13.7 (pg. 612) shows the state transition diagram unfolded over time

- finish prob. models by defining conditional dist. of the observed variables $\mathbf{x}_n$:
  $p(\mathbf{x}_n | \mathbf{z}_n, \boldsymbol{\phi})$ where $\boldsymbol{\phi}$ are the parameters of the dist. ("emission prob.")
  - $p(\mathbf{x}_n | \mathbf{z}_n, \boldsymbol{\phi})$ consists of a vector of K numbers corresponding to the K possible states of a binary vector $\mathbf{z}_n$
- emission probabilities:

$$p(\mathbf{x}_n | \mathbf{z}_n, \boldsymbol{\phi}) = \prod_{k=1}^{K} p(\mathbf{x}_n | \boldsymbol{\phi}_k)^{z_{nk}}$$ (recall z_{nk} is either 1 or 0)

- for a homogeneous model, all latent vars. use same $\mathbf{A}$ & all emission dist. use the same $\boldsymbol{\phi}$

- joint prob. dist. over latent & observed variables:

$$p(X, Z \mid \theta) = p(z_1 \mid \pi)\left[\prod_{n=2}^{N} p(z_n \mid z_{n-1}, A)\right]\left[\prod_{m=1}^{N} p(x_m \mid z_m, \phi)\right]$$

$$X = \{x_1, \dots, x_N\}, \quad Z = \{z_1, \dots, z_n\}, \quad \theta = \{\pi, A, \phi\}$$

13.3 Linear Dynamical System (LDS)

- continuous latent variables
  - the general form of the inference algorithms are the same as for HMM
- we will consider a linear-Gaussian state space model
  - latent vars. $\{z_n\}$ & observed variables $\{x_n\}$ are multivariate Gauss.
    - means of these dist. are linear fxns of their parents in the graph
- can interpret LDS as a generalization of continuous latent variable models such as prob. PCA
  - each pair of nodes $\{z_n, x_n\}$ represents a linear-Gaussian latent var. model for the observation, but now the latent vars are not independent
- because LDS is a linear-Gaussian model, all joint, marginal, & conditional dist. will be Gaussian.

- transition dist: $p(z_n \mid z_{n-1}) = \mathcal{N}(z_n \mid A z_{n-1}, \Gamma)$

- emission dist: $p(x_n \mid z_n) = \mathcal{N}(x_n \mid C z_n, \Sigma)$

- initial latent var. dist: $p(z_1) = \mathcal{N}(z_1 \mid \mu_0, V_0)$

- alternative formulations of above:

$$z_n = A z_{n-1} + w_n \qquad w \sim \mathcal{N}(w \mid 0, \Gamma)$$
$$x_n = C z_n + v_n \qquad v \sim \mathcal{N}(v \mid 0, \Sigma)$$
$$z_1 = \mu_0 + u \qquad u \sim \mathcal{N}(u \mid 0, V_0)$$

(noise terms)

13.3.1. Inference in LDS

- goal: find marginal dist. for the latent variables conditional on the data.
  - and make predictions on the next latent state & observation
- accomplish w/ the "sum-product algorithm"
  - w.r.t LDS, results in the "Kalman filter" and "Kalman smoother" equations
  - b/c only dealing w/ Gaussians, could use standard results for conditionals & marginals, but "sum-product alg." is more efficient

- begin by considering the forward equations
  - z_n: root node $h(z_1)$: leaf node
  - propagate msg from leaf to root
- propagate messages that are normalized marginal dist.:

$$p(z_n | x_1, \ldots, x_n) \rightarrow \hat{\alpha}(z_n) = N(z_n | \mu_n, V_n)$$

- recursion equation:

$$c_n \hat{\alpha}(z_n) = p(x_n | z_n) \int \hat{\alpha}(z_{n-1}) p(z_n | z_{n-1}) \, dz_{n-1}$$

Session 6 Lecture Notes 7/2/2022

Topics:

1. Laplace Approximation

2. State space models (HMM & linear dynamical systems)

1. Laplace Approx.

- normally w/ MAP estimates, do not get a value of uncertainty
  - for Laplace approx., can use the variance as an estimate of uncertainty
- Laplace approx. is best when the actual prob. dist. is Gaussian-ish
  - symmetric w/ long tails

2. State space models

A. General structure

- when making a Markov chain (MC):
  - assuming previous state is all that's needed to know the current state
  - requires very descriptive x
- state space models assume a latent state space z
- the measurements x are noisy observations of z
  - emission model: $p(x_n | z_n)$
- 2 types of tractable state space models:
  1. HMM
     - discrete latent space
     - $p(z_n | z_{n-1})$ is a transition matrix
  2. Kalman filter
     - continuous observations x and latent space z
     - linear Gaussians $p(z_n | z_{n-1})$ & $p(x_n | z_n)$

B. Filtering (forward pass)

→ "What is my estimate for $\mathbf{z}_n$ given all observations $\mathbf{x}_{1:n}$?"

assume: $p(\mathbf{z}_{n-1} \mid \mathbf{x}_{1:n-1})$

prediction: of latent space

$$p(\mathbf{z}_n \mid \mathbf{x}_{1:n-1}) = \sum_{\mathbf{z}_{n-1}} p(\mathbf{z}_n, \mathbf{z}_{n-1} \mid \mathbf{x}_{1:n-1})$$

$$= \sum_{\mathbf{z}_{n-1}} p(\mathbf{z}_n \mid \mathbf{z}_{n-1}, \cancel{\mathbf{x}_{n-1}})\, p(\mathbf{z}_{n-1} \mid \mathbf{x}_{1:n-1})$$

x_{n-1} is not needed if we are given z_{n-1}

$$= \sum_{\mathbf{z}_{n-1}} p(\mathbf{z}_n \mid \mathbf{z}_{n-1})\, p(\mathbf{z}_{n-1} \mid \mathbf{x}_{1:n-1})$$

prediction of observation:

$$p(\mathbf{z}_n \mid \mathbf{x}_{1:n}) = p(\mathbf{z}_n \mid x_n, \mathbf{x}_{1:n-1})$$

Baye's Rule

$$\propto p(x_n \mid \mathbf{z}_n, \mathbf{x}_{1:n-1})\, \underbrace{p(\mathbf{z}_n \mid \mathbf{x}_{1:n-1})}$$

C. Smoothing (forward & backward pass)

→ "What is my estimate of some $\mathbf{z}_n$ given all observations $\mathbf{x}_{1:N}$?"

· guessing at a previous latent space given future data

$$p(\mathbf{z}_n \mid \mathbf{x}_{1:N}) = p(\mathbf{z}_n \mid \mathbf{x}_{1:n}, \mathbf{x}_{n+1:N})$$

Baye's Rule

$$\propto p(\mathbf{x}_{n+1:N} \mid \mathbf{z}_n, \mathbf{x}_{1:n})\, p(\mathbf{z}_n \mid \mathbf{x}_{1:n})$$

if know z_n, do not need $x_{1:n}$

$$\propto \underbrace{p(\mathbf{x}_n \mid \mathbf{z}_n)}_{\text{"emission model"} \approx \text{likelihood of } x_n}\, \underbrace{p(\mathbf{z}_n \mid \mathbf{x}_{1:n})}_{\text{"forward pass" (above)} \approx \text{"prior"}}$$