CSE 291B 2/5

Probabilistic modeling

- Notation
  - Evidence X
  - Hidden variables H
  - Model parameters β
- Bayes rule

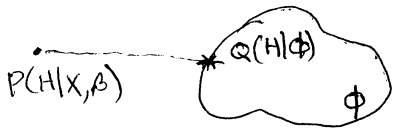
  Posterior $P(H|X,\beta) = \frac{P(X|H,\beta)P(H|\beta)}{P(X|\beta)}$
- challenges

  Inference: how to compute $P(X|\beta)$ and statistics of $P(H|X,\beta)$?

  Learning: how to compute $\arg\max_\beta P(X|\beta)$

  Often intractable

Variational methods



- Inference

  Approximate intractable $P(H|X,\beta)$ by tractable $Q(H|\phi)$.

  Vary ϕ to minimize $KL(Q(H|\phi), P(H|X,\beta))$.
- Useful bound:

$$\log P(X|\beta) = \sum_H Q(H|\phi) \log \frac{P(X,H|\beta)}{P(H|\beta)}$$

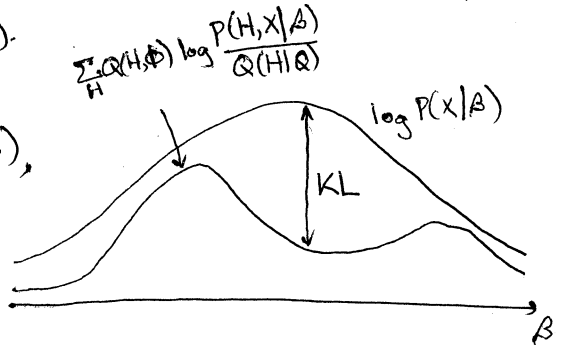
$$= \sum_H Q(H|\phi) \log \frac{P(X,H|\beta)}{Q(H|\phi)} + KL(Q,P)$$

$$\geq \sum_H Q(H|\phi) \log \frac{P(X,H|\beta)}{Q(H|\phi)}$$

- Connection to learning

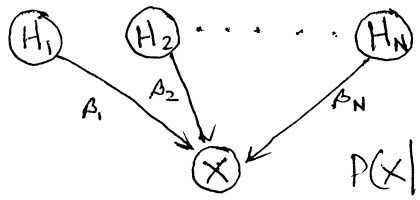
  The better the approximation $Q(H|\phi)$, the smaller the error $KL(Q,P)$, the tighter the bound on $\log P(X|\beta)$.
- Variational learning

  When intractable to compute $\log P(X|\beta)$, maximize instead its lower bound.



Example: combinatorial Gaussian mixture model

- Generative model
  - Flip N fair coins with unknown cash values β_n.
  - Ignore coins with tails, sum coin values for heads.
  - Observe summed value corrupted by Gaussian noise.
- Graphical model



$H_n \in \{0,1\}$
$P(H_n = 1) = \frac{1}{2}$ for all n

$$P(X|H) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X - \sum_{n=1}^{N} \beta_n H_n)^2}$$

Joint distribution:

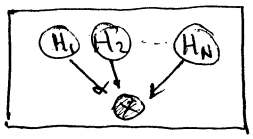
$$P(X,H) = P(X|H) \prod_i P(H_i)$$

Marginal distribution

$$P(X) = \sum_{H \in \{0,1\}^N} P(X,H)$$ intractable to compute exactly

mixture components $\sim 2^N$

- Variational approximation

Replace  by 

$$Q(H|\phi) = \prod_{n=1}^{N} Q(H_n|\phi_n) = \prod_n \phi_n^{H_n} (1-\phi_n)^{1-H_n}$$

- Approximation error

$$KL(Q,P) = \sum_H Q(H|\phi) \log \frac{Q(H|\phi)}{P(H|X,\beta)} = \sum_H Q(H|\phi) \log \frac{Q(H|\phi)}{P(H,X|\beta)} \cdot P(X|\beta)$$

$$= \sum_H Q(H|\phi) \log \frac{Q(H|\phi)}{P(H,X|\beta)} + \text{constant independent of } \phi$$

- Negative entropy term

$$\sum_H Q(H|\Phi) \log Q(H|\Phi) = \sum_H \left[\prod_n Q(H_n|\Phi_n) \log \prod_n Q(H_n|\Phi_n) \right]$$

$$= \sum_{H_1} \sum_{H_2} \cdots \sum_{H_N} Q(H_1|\Phi_1) \cdots Q(H_N|\Phi_N) \sum_n \log Q(H_n|\Phi_n)$$

$$= \sum_n \sum_{H_1} \cdots \sum_{H_N} Q(H_1|\Phi_1) \cdots Q(H_N|\Phi_N) \log Q(H_n|\Phi_n)$$

$$= \sum_n \sum_{H_n} Q(H_n|\Phi_n) \log Q(H_n|\Phi_n)$$

$$= \sum_n \left[\Phi_n \log \Phi_n + (1-\Phi_n) \log(1-\Phi_n) \right]$$

- other term in $KL(Q,P)$:

$$\sum_H Q(H|\Phi) \log P(H,X|\beta) = \sum_H Q(H|\Phi) \log \left\{ P(X|H,\beta) \prod_n P(H_n) \right\}$$

$$= \sum_H Q(H|\Phi) \left\{ -\tfrac{1}{2}\log(2\pi) - \tfrac{1}{2}(X - \textstyle\sum_n \beta_n H_n)^2 + \log\left(\tfrac{1}{2}\right)^N \right\}$$

$$= -\tfrac{1}{2}\log(2\pi) + N\log\left(\tfrac{1}{2}\right) - \tfrac{1}{2} E_{Q(H|\Phi)}\left[(X - \textstyle\sum_n \beta_n H_n)^2 \right]$$

Aside: $E_Q[H_n] = \Phi_n(1) + (1-\Phi_n)(0) = \Phi_n$

$$E_Q[H_n H_m] = \begin{cases} E_Q[H_n]\, E_Q[H_m] = \Phi_n \Phi_m & \text{if } m \neq n \\ E_Q[H_n^2] = E[H_n] = \Phi_n & \text{if } m = n \end{cases}$$

↑ because $H_n \in \{0, 1\}$

- Up to numerical constants:

$$\sum_H Q(H|\Phi) \log P(H,X|\beta) = -\tfrac{1}{2} E_Q\left[X^2 - 2X \textstyle\sum_n \beta_n H_n + \sum_{nm} \beta_n \beta_m H_n H_m \right] + \text{numerical constants}$$

$$= -\tfrac{1}{2} X^2 + X \textstyle\sum_n \beta_n \Phi_n - \tfrac{1}{2} \sum_{nm} \beta_n \beta_m \left[\Phi_n \delta_{nm} + \Phi_n \Phi_m (1 - \delta_{nm}) \right]$$

↑ discrete delta function

- Optimizing the approximation:

$$KL(Q,P) = \sum_n [\phi_n \log\phi_n + (1-\phi_n)\log(1-\phi_n)] - X\sum_n \beta_n\phi_n + \frac{1}{2}\sum_n \beta_n^2\phi_n + \frac{1}{2}\sum_{n\neq m} \beta_n\beta_m\phi_n\phi_m + \dots$$

terms independent of ϕ_n

Setting $\frac{\partial}{\partial\phi_n}[KL(Q,P)] = 0$:

$$\log\phi_n + \frac{\phi_n}{\phi_n} - \log(1-\phi_n) - \frac{1-\phi_n}{1-\phi_n} = X\beta_n - \frac{1}{2}\beta_n^2 - \beta_n\sum_{m\neq n}\beta_m\phi_m$$

Collecting terms

$$\log\left(\frac{\phi_n}{1-\phi_n}\right) = X\beta_n - \frac{1}{2}\beta_n^2 - \beta_n\sum_{m\neq n}\beta_m\phi_m$$

inverse sigmoid function

$$\phi_n = \sigma\left(\beta_n\left[X - \sum_{m\neq n}\beta_m\phi_m - \frac{\beta_n}{2}\right]\right) \quad \text{where } \sigma(z) = \frac{1}{1+e^{-z}}$$

- Intuition:

If $\underbrace{X - \sum_{m\neq n}\beta_m E[H_m]}_{\text{residual value after excluding the other coins}} > \frac{\beta_n}{2}$, then n^{th} coin is needed to account for observed X, with $\phi_n > \frac{1}{2}$

- Variational E-step:

To find local minimum of $KL(Q,P)$ with respect to $\{\phi_n\}$:

Iterative optimization:

1) Initialize ϕ_n at random. [Or use ϕ_n from previous E-step (i.e., previous round)]
2) For $n = 1 \dots N$

$$\phi_n \leftarrow \sigma\left(\beta_n\left(X - \sum_{n\neq m}\beta_m\phi_m - \frac{\beta_n}{2}\right)\right)$$

3) Repeat #2 until convergence

- Variational learning

  Given data $\{X_t\}_{t=1}^{T}$, how to recover the cash value β_n of each coin?

  Maximum likelihood estimation:

  Choose β to maximize $\mathcal{L} = \sum_t \log P(X_t|\beta)$

  But intractable to compute

- Log-likelihood bound

$$\mathcal{L}(\beta) \geq \sum_t \sum_H Q(H|\phi^{(t)}) \log \frac{P(X_t, H|\beta)}{Q(H|\phi^{(t)})}$$

$$= \sum_t \left\{ X_t \sum_n \beta_n \phi_n^{(t)} - \frac{1}{2} \sum_n \beta_n^2 \phi_n^{(t)} - \frac{1}{2} \sum_{n \neq m} \beta_n \beta_m \phi_n^{(t)} \phi_m^{(t)} \right\} + \text{terms independent of } \beta$$

where $\phi_n^{(t)}$ are chosen to minimize $KL(Q,P)$

$$\delta_{nm} = \begin{cases} 1 & \text{if } n=m \\ 0 & \text{otherwise} \end{cases}$$

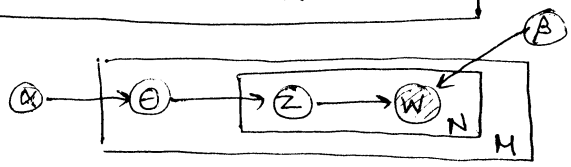
- Variational M-step

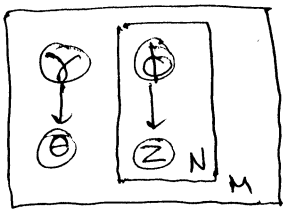
  Setting $\frac{\partial}{\partial \beta_n}[\text{RHS}] = 0$ yields linear set of equations.

$$\sum_m \left\{ \sum_t \left[\phi_m^{(t)} \delta_{nm} + (1-\delta_{nm}) \phi_n^{(t)} \phi_m^{(t)} \right] \right\} \beta_m = \sum_t X_t \phi_n^{(t)}$$

  Update β by solving these equations.

Ex: Latent Dirichlet allocation

Replace 

by 

$P(\theta, \vec{z} \mid \vec{w}, \alpha, \beta)$
intractable

$$Q(\theta, \vec{z} \mid \gamma, \phi) = Q(\theta \mid \gamma) \prod_n Q(z_n \mid \phi_n)$$

$$Q(\theta \mid \gamma) = \frac{\Gamma(\sum_k \gamma_k)}{\prod_k \Gamma(\gamma_k)} \prod_k \theta_k^{\gamma_k - 1}$$ Dirichlet

$$Q(z_n \mid \phi_n) = \prod_{k=1}^{K} \phi_{kn}^{I(z_n, k)}$$ multinomial