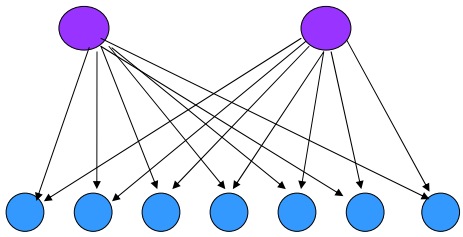
CS 3750 Machine Learning
Lecture 19

Latent variable models
Variational approximations.

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

Cooperative vector quantizer

Latent variables (s): binary vars
Dimensionality k



Observed variables x: real valued vars
Dimensionality d

Cooperative vector quantizer

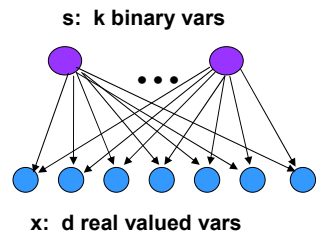
Model:

Latent var s_i:

~ Bernoulli distribution

parameter: π_i

$$P(s_i \mid \pi_i) = \pi_i^{s_i}(1-\pi_i)^{1-s_i}$$



Observable variables x:

~ Normal distribution

parameters: **W, Σ**

$$P(\mathbf{x} \mid \mathbf{s}) = N(\mathbf{W}\mathbf{s}, \Sigma)$$

We assume $\Sigma = \sigma I$

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} & .. & w_{1k} \\ w_{21} & & & \\ & .. & & \\ w_{d1} & .. & .. & w_{dk} \end{pmatrix}$$

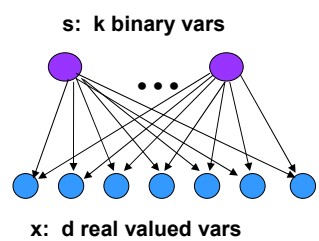
Joint for one instance of x and s:

$$P(\mathbf{x},\mathbf{s} \mid \Theta) = (2\pi)^{d/2}\sigma^{d/2}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{W}\mathbf{s})^T(\mathbf{x}-\mathbf{W}\mathbf{s})\right\}\prod_{i=1}^{k}\pi_i^{s_i}(1-\pi_i)^{(1-s_i)}$$

Cooperative vector quantizer

Our objective:

- **Learn the parameters of the model W, π, σ**
- **One can use the data likelihood or loglikelihood and optimize ..**



Learning if x and s are observable

Log likelihood:

$$\sum_{n=1}^{N} \log P(\mathbf{x}^{(n)},\mathbf{s}^{(n)} \mid \Theta) =$$

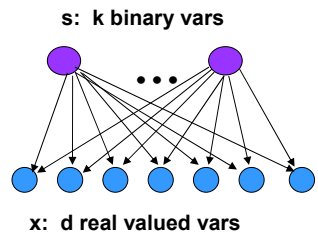
$$\sum_{n=1}^{N} -d\log\sigma - \frac{1}{2\sigma^2}(\mathbf{x}^{(n)}-\mathbf{W}\mathbf{s}^{(n)})^T(\mathbf{x}^{(n)}-\mathbf{W}\mathbf{s}^{(n)}) + \sum_{i=1}^{k} s_i^{(n)}\log\pi_i + (1-s_i^{(n)})\log(1-\pi_i)$$

Solution: nice and easy

Cooperative vector quantizer

Our objective:

- **Learn the parameters of the model W, π, σ**
- **One can use the data likelihood or loglikelihood and optimize ..**



Learning if only x are observable

Log likelihood of data:

$$\log P(D|\Theta) = \sum_{n=1}^{N} \log P(\mathbf{x}^{(n)}|\Theta) = \sum_{n=1}^{N} \log \sum_{\{\mathbf{s}^n\}} P(\mathbf{x}^{(n)}, \mathbf{s}^{(n)}|\Theta)$$

Solution: does not let us benefit from the decomposition

EM: used to work in such cases …

EM

Let H – be a set of all variables with hidden or missing values

Derivation

$$P(H, D|\Theta,\xi) = P(H|D,\Theta,\xi)P(D|\Theta,\xi)$$

$$\log P(H, D|\Theta,\xi) = \log P(H|D,\Theta,\xi) + \log P(D|\Theta,\xi)$$

$$\underline{\log P(D|\Theta,\xi)} = \log P(H, D|\Theta,\xi) - \log P(H|D,\Theta,\xi)$$

Log-likelihood of data

Average both sides with $P(H|D,\Theta',\xi)$ for Θ'

$$E_{H|D,\Theta'} \log P(D|\Theta,\xi) = E_{H|D,\Theta'} \log P(H, D|\Theta,\xi) - E_{H|D,\Theta'} \log P(H|D,\Theta,\xi)$$

$$\underbrace{\log P(D|\Theta,\xi)}_{\text{Log-likelihood of data}} = F(\Theta|\Theta') = E(\Theta|\Theta') + H(\Theta|\Theta')$$

EM algorithm

Algorithm (general formulation)

Initialize parameters Θ

Repeat

Set $\Theta' = \Theta$

1. **Expectation step**

$$E(\Theta \mid \Theta') = \langle \log P(H, D \mid \Theta, \xi) \rangle_{P(H \mid D, \Theta')}$$

2. **Maximization step**

$$\Theta = \arg\max_{\Theta} E(\Theta \mid \Theta')$$

until no or small improvement in Θ $(\Theta = \Theta')$

Problem: posterior $P(H \mid D, \Theta', \xi)$ is defined over 2^k probabilities

EM algorithm

Posterior $P(H \mid D, \Theta', \xi)$ for our model

$$P(H \mid D, \Theta') = \prod_{n=1}^{N} P(s^{(n)} \mid x^{(n)}, \Theta')$$

– Each data point n=1, …N requires us to calculate 2^k probabilities

– If k is larger then this is a bottleneck!!!

Variational approximation

Let H – be a set of all variables with hidden or missing values

Derivation

$$\underline{\log P(D \mid \Theta, \xi)} = \log P(H, D \mid \Theta, \xi) - \log P(H \mid D, \Theta, \xi)$$

Log-likelihood of data

Average both sides with $\boxed{Q(H \mid \lambda)}$

$$E_{H|\lambda} \log P(D \mid \Theta, \xi) = E_{H|\lambda} \log P(H, D \mid \Theta, \xi) - E_{H|\lambda} \log P(H \mid \Theta, \xi)$$
$$+ E_{H|\lambda} \log Q(H \mid \lambda) - E_{H|\lambda} \log Q(H \mid \lambda)$$

$$\log P(D \mid \Theta, \xi) = F(P, Q) + KL(Q, P)$$

Log-likelihood of data

Variational approximation

Let H – be a set of all variables with hidden or missing values

$$\log P(D \mid \Theta, \xi) = E_{H|\lambda} \log P(H, D \mid \Theta, \xi) - E_{H|\lambda} \log P(H \mid \Theta, \xi)$$
$$+ E_{H|\lambda} \log Q(H \mid \lambda) - E_{H|\lambda} \log Q(H \mid \lambda)$$

$$\log P(D \mid \Theta, \xi) = F(Q, \Theta) + KL(Q, P)$$

$$F(Q, \Theta) = \sum_{\{H\}} Q(H \mid \lambda) \log P(H, D \mid \Theta, \xi) - \sum_{\{H\}} Q(H \mid \lambda) \log Q(H \mid \lambda)$$

$$KL(Q.P) = \sum_{\{H\}} Q(H \mid \lambda) \left[\log Q(H \mid \lambda) - \log P(H \mid D, \Theta)\right]$$

Approximation: maximize $F(Q, \Theta)$

Parameters: Θ, λ

Variational EM

Let H – be a set of all variables with hidden or missing values

- **E step:**
  - Optimize

    $F(Q,\Theta)$ with respect to λ while keeping Θ fixed
- M step
  - Optimize

    $F(Q,\Theta)$ with respect to Θ while keeping λ**s**

Note: if $Q(H)$ is the posterior then the variational EM reduces to the standard EM

Variational EM

- So what is the deal?
  - Why should we use the variational EM?
- Hope:
  - If we choose $Q(H \mid \lambda)$ well the optimization of both λ and Θ will become easy
- A well behaved choice for $Q(H \mid \lambda)$
  - Use mean field approximation

Mean Field Approximation

Assumption:

- $Q(H|\lambda)$ is the mean field approximation.
- Variables in the *Q(H)* distribution are independent variables H_i.
- Q is completely factorized:

$$Q(H \mid \lambda) = \prod_i Q_i(H_i \mid \lambda_i)$$

- For our CVQ model
  - Hidden variables are binary sources

$$Q(\mathbf{s} \mid \lambda) = \prod_i Q_i(s_i \mid \lambda_i)$$

$$Q(s_i \mid \lambda_i) = \lambda_i^{s_i}(1-\lambda_i)^{1-s_i}$$

Mean Field Approximation

Functional F for the mean field:

$$F(Q,\Theta) = \sum_{\{H\}} Q(H \mid \lambda) \log P(H, D \mid \Theta, \xi) - \sum_{\{H\}} Q(H \mid \lambda) \log Q(H \mid \lambda)$$

$$F(Q,\Theta) = \left\langle \log P(\mathbf{x},\mathbf{s} \mid \Theta) \right\rangle_{Q(s|\lambda)} - \left\langle \log Q(\mathbf{s} \mid \lambda) \right\rangle_{Q(s|\lambda)}$$

$$= \left\langle -d \log \sigma - \frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{W}\mathbf{s})^T(\mathbf{x}-\mathbf{W}\mathbf{s}) \right\rangle_{Q(s|\lambda)} \qquad (1)$$

$$+ \left\langle \sum_{i=1}^{k} s_i \log \pi_i + (1-s_i)\log(1-\pi_i) \right\rangle_{Q(s|\lambda)} \qquad (2)$$

$$- \left\langle \sum_{i=1}^{k} s_i \log \lambda_i + (1-s_i)\log(1-\lambda_i) \right\rangle_{Q(s|\lambda)} \qquad (3)$$

Mean Field Approximation

Functional F. Part 1:

$$\left\langle -d\log\sigma - \frac{1}{2\sigma^2}(\mathbf{x}-\sum_{i=1}^{k} s_i\mathbf{w}_i)^T(\mathbf{x}-\sum_{i=1}^{k} s_i\mathbf{w}_i)\right\rangle_{Q(s|\lambda)} =$$

$$= \left\langle -d\log\sigma - \frac{1}{2\sigma^2}(\mathbf{x}-\sum_{i=1}^{k} s_i\mathbf{w}_i)^T(\mathbf{x}-\sum_{i=1}^{k} s_i\mathbf{w}_i)\right\rangle_{Q(s|\lambda)}$$

$$= \left\langle -d\log\sigma - \frac{1}{2\sigma^2}\left[\mathbf{x}^T\mathbf{x} - 2\sum_{i=1}^{k}(s_i\mathbf{w}_i)\mathbf{x} + \sum_{i=1}^{k}\sum_{j=1}^{k} s_i s_j \mathbf{w}_i^T\mathbf{w}_j\right]\right\rangle_{Q(s|\lambda)}$$

$$= -d\log\sigma - \frac{1}{2\sigma^2}\left[\mathbf{x}^T\mathbf{x} - 2\sum_{i=1}^{k}\langle s_i\rangle_{Q(s_i|\lambda_i)}\mathbf{w}_i)\mathbf{x} + \sum_{i=1}^{k}\sum_{j=1}^{k}\langle s_i s_j\rangle_{Q(s|\lambda)}\mathbf{w}_i^T\mathbf{w}_j\right]$$

$$\langle s_i\rangle_{Q(s_i|\lambda_i)} = \lambda_i \qquad \langle s_i s_j\rangle_{Q(s|\lambda)} = \lambda_i\lambda_j + \delta_{ij}(\lambda_i - \lambda_i^2)$$

Mean Field Approximation

Functional F. Part 2:

$$\left\langle \sum_{i=1}^{k} s_i\log\pi_i + (1-s_i)\log(1-\pi_i)\right\rangle_{Q(s|\lambda)} = \sum_{i=1}^{k}\langle s_i\rangle_{Q(s_i|\lambda_i)}\log\pi_i + (1-\langle s_i\rangle_{Q(s_i|\lambda_i)})\log(1-\pi_i)$$

$$= \sum_{i=1}^{k}\lambda_i\log\pi_i + (1-\lambda_i)\log(1-\pi_i)$$

Functional F. Part 3:

$$\left\langle \sum_{i=1}^{k} s_i\log\lambda_i + (1-s_i)\log(1-\lambda_i)\right\rangle_{Q(s|\lambda)} = \sum_{i=1}^{k}\lambda_i\log\lambda_i + (1-\lambda_i)\log(1-\lambda_i)$$

Mean Field Approximation

Functional F:

$$F(Q,\Theta) = \langle \log P(\mathbf{x},\mathbf{s} \mid \Theta) \rangle_{Q(s|\lambda)} - \langle \log Q(\mathbf{s} \mid \lambda) \rangle_{Q(s|\lambda)}$$

$$= -d \log \sigma - \frac{1}{2\sigma^2}\left[\mathbf{x}^T\mathbf{x} - 2\sum_{i=1}^{k} \lambda_i \mathbf{w}_i)\mathbf{x} + \sum_{i=1}^{k}\sum_{j=1}^{k} \left[\lambda_i\lambda_j + \delta_{ij}(\lambda_i - \lambda_i^{\,2})\right]\mathbf{w}_i^{\,T}\mathbf{w}_j \right]$$

$$+ \sum_{i=1}^{k} \lambda_i \log \pi_i + (1-\lambda_i)\log(1-\pi_i)$$

$$+ \sum_{i=1}^{k} \lambda_i \log \lambda_i + (1-\lambda_i)\log(1-\lambda_i)$$

Parameters: W, π, σ

Mean field parameters: λ

Mean Field Approximation

Functional F (for all data points):

$$F(Q,\Theta) = \sum_{n=1}^{N} \langle \log P(\mathbf{x},\mathbf{s} \mid \Theta) \rangle_{Q(s|\lambda)} - \langle \log Q(\mathbf{s} \mid \lambda) \rangle_{Q(s|\lambda)}$$

$$= -d \log \sigma - \frac{1}{2\sigma^2}\left[\mathbf{x}^{(n)T}\mathbf{x}^{(n)} - 2\sum_{i=1}^{k} \lambda_i^{(n)} \mathbf{w}_i)\mathbf{x}^{(n)} + \sum_{i=1}^{k}\sum_{j=1}^{k} \left[\lambda_i^{(n)}\lambda_j^{(n)} + \delta_{ij}(\lambda_i^{(n)} - \lambda_i^{(n)2})\right]\mathbf{w}_i^{\,T}\mathbf{w}_j \right]$$

$$+ \sum_{i=1}^{k} \lambda_i^{(n)} \log \pi_i + (1-\lambda_i^{(n)})\log(1-\pi_i)$$

$$+ \sum_{i=1}^{k} \lambda_i^{(n)} \log \lambda_i^{(n)} + (1-\lambda_i^{(n)})\log(1-\lambda_i^{(n)})$$

Parameters: W, π, σ

Mean field parameters: $\lambda = \lambda^{(1)}, \lambda^{(2)}, \ldots \lambda^{(N)}$

Variational EM: E step

Optimization of the functional F with respect to λ:

$$\frac{\partial}{\partial \lambda_u} F = \frac{1}{\sigma^2}(\mathbf{x} - \sum_{j \neq u} \lambda_j \mathbf{w}_j)^T \mathbf{w}_u - \frac{1}{2\sigma^2} \mathbf{w}_u{}^T \mathbf{w}_u + \log \frac{\pi_u}{1-\pi_u} - \log \frac{\lambda_u}{1-\lambda_u}$$

set $\quad \frac{\partial}{\partial \lambda_u} F = 0$

$$\lambda_u = g\left(\frac{1}{\sigma^2}(\mathbf{x} - \sum_{j \neq u} \lambda_j \mathbf{w}_j)^T \mathbf{w}_u - \frac{1}{2\sigma^2} \mathbf{w}_u{}^T \mathbf{w}_u + \log \frac{\pi_u}{1-\pi_u} \right)$$

$$g(x) = \frac{1}{1+e^{-x}}$$

Defines a fixed point equation

Iterate a set fixed point equations for all indexes u=1..k

Variational EM: M step

Optimization of the functional F with respect to Θ.

Start with **π:**

$$\frac{\partial}{\partial \pi_u} F = \lambda_u \log \frac{1}{\pi_u} - (1-\lambda_u) \log \frac{1}{(1-\pi_u)}$$ For 1 data point

$$\frac{\partial}{\partial \pi_u} F = \sum_{n=1}^{N} \lambda_u{}^n \log \frac{1}{\pi_u} - (1-\lambda_u{}^n) \log \frac{1}{(1-\pi_u)}$$ For N data points

set $\quad \frac{\partial}{\partial \pi_u} F = 0$

$$\pi_u = \frac{\sum_{n=1}^{N} \lambda_u{}^{(n)}}{N}$$ Closed form solution

Variational EM: M step

Optimization of the functional F with respect to Θ.

Parameters **w:**

$$\frac{\partial}{\partial w_{uv}} F = \sum_{n=1}^{N} -\frac{1}{2\sigma^2}\left[\lambda_v^{(n)} x_u^{(n)} + 2\sum_{j\neq v} \lambda_v^{(n)} \lambda_j^{(n)} w_{uj} + 2\lambda_v w_{uv}\right] = 0$$

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} & .. & w_{1k} \\ w_{21} & & & \\ & .. & & \\ w_{d1} & .. & .. & w_{dk} \end{pmatrix} \qquad \mathbf{W} = (\mathbf{w_1}\ \mathbf{w_2}\ \ldots\ \mathbf{w_k})$$

For each variable v:

The equations define a set of k linear equations that can be solved

Mean Field approximation

Let H – be a set of all variables with hidden or missing values

- E step:
  - Optimize $KL(Q.P) = \sum_{\{H\}} Q(H \mid \lambda)\left[\log Q(H \mid \lambda) - \log P(H \mid D, \Theta)\right]$

$$\log P(D \mid \Theta, \xi) = F(P, Q) + KL(Q, P)$$

$$F(P,Q) = \sum_{\{H\}} Q(H \mid \lambda) \log P(H, D \mid \Theta, \xi) - \sum_{\{H\}} Q(H \mid \lambda) \log Q(H \mid \lambda)$$

$$KL(Q.P) = \sum_{\{H\}} Q(H \mid \lambda)\left[\log Q(H \mid \lambda) - \log P(H \mid D, \Theta)\right]$$

Approximation: maximize $F(P,Q)$

Parameters: Θ, λ