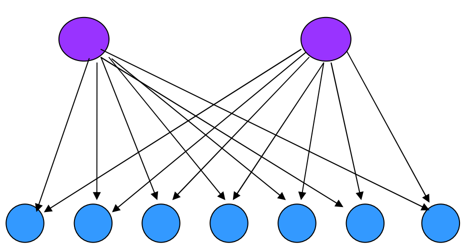
CS 3750 Machine Learning
Lecture 15

Latent variable models
Variational approximations.

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

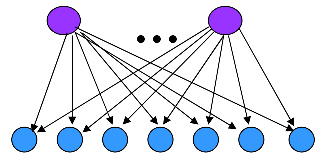
Cooperative vector quantizer

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

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

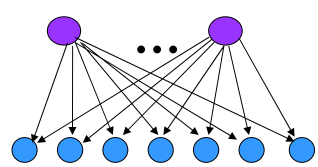
We assume $\Sigma = \sigma I$

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

Joint for one instance of x and s:

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

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:

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

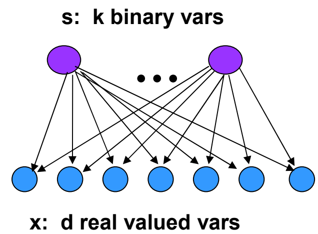
$$\sum_{n=1}^{N} -d\log\sigma - \frac{1}{2\sigma^2}(\mathbf{x}^{(n)}-\mathbf{Ws}^{(n)})^T(\mathbf{x}^{(n)}-\mathbf{Ws}^{(n)}) + \sum_{i=1}^{k} s_i^{(n)}\log\pi_i + (1-s_i^{(n)})\log(1-\pi_i) + c$$

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:

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

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

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

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

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

Log-likelihood of data

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

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

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

Log-likelihood of data

EM algorithm

Algorithm (general formulation)

Initialize parameters Θ

Repeat

Set $\Theta' = \Theta$

1. **Expectation step**

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

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

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

Log-likelihood of data

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

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

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

Log-likelihood of data

Variational approximation

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

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

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

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

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

Parameters: Θ, λ

Why? $\log P(D \mid \Theta, \xi) \geq F(Q, \Theta)$

Maximization of F pushes up the lower bound on the log-likelihood

Variational approximation

- **Comparison:**
  - **EM uses the true posterior** $P(H \mid D, \Theta', \xi)$
  - **Variational EM uses a surrogate posterior** $Q(H \mid \lambda)$

EM:

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

Variational EM:

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

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

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

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 $Q(H \mid \lambda)$
  - the 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

$$Q(\mathbf{H} \mid \boldsymbol{\lambda}) = \prod_{n=1,..N} Q(\mathbf{s}^{(n)} \mid \boldsymbol{\lambda}^{(n)})$$

$$Q(\mathbf{s}^{(n)} \mid \lambda^{(n)}) = \prod_{i=1,..d} Q(s_i^{(n)} \mid \lambda_i^{(n)})$$

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

Mean Field Approximation

Functional F for the mean field:

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

Assume just one data point $\mathbf{x}$ and corresponding $\mathbf{s}$:

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

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

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

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

Mean Field Approximation

Functional F. Part 1:

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

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

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

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

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

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):

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

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

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

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

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 and for all n

Variational EM: M step

Optimization of the functional F with respect to Θ.

Start with π**:**

For N data points

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

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

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

Closed form solution

Variational EM: M step

Optimization of the functional F with respect to Θ.

Parameters **w:**

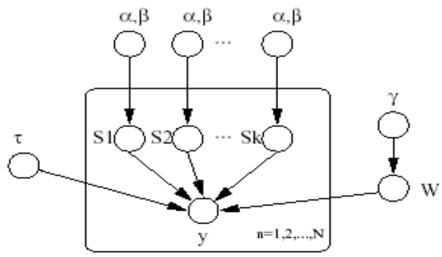
$$\frac{\partial}{\partial w_{uv}} F = \sum_{n=1}^{N} -\frac{1}{2\sigma^2}\left[\lambda_v^{(n)} x_u^{(n)} + 2\sum_{j\neq v} \lambda_v^{(n)} \lambda_j^{(n)} w_{uj} + 2\lambda_v^{(n)} w_{uv} \right] = 0$$

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

For each variable v:

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

Bayesian CVQ Model

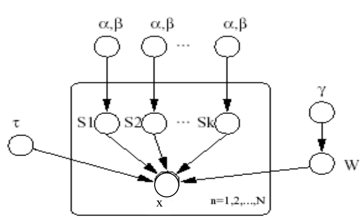


$$y = \sum_{k=1}^{K} s_k w_k + \varepsilon$$

Bayesian model:
Distributions over parameters

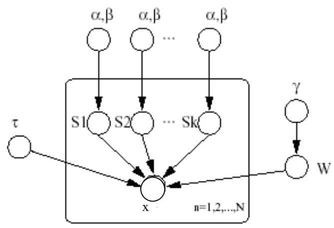
$$P(y \mid S, \theta) \sim N\left(\sum_{k=1}^{K} s_k w_k, \tau^{-1} I \right)$$

Model Specification



$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2 \ldots, \boldsymbol{x}_n\}$	observed data
$\boldsymbol{S} = \{s_1, \ldots, s_k\}$	latent sources
$\pi = \{\pi_1, \pi_2 \ldots, \pi_k\}$	probability of $s_k = 1$
$\boldsymbol{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_k\}$	DxK weight matrix
$\gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_k\}$	Variance of $\boldsymbol{W}$
τ	Precision of noise

Priors



$$P(\boldsymbol{\pi}) = \prod_{k=1}^{K} Beta\,(\pi_k \mid \alpha, \beta)$$

$$P(\mathbf{W}) = \prod_{k=1}^{K} N\left(w_k \mid 0, \gamma_k\right)$$

$$P(\boldsymbol{\gamma}) = \prod_{k=1}^{K} Gamma\,\left(\gamma_k \mid a_\gamma, b_\gamma\right)$$

$$P(\tau) = Gamma\,\left(\tau \mid c_\tau, d_\tau\right)$$

Variational approximation

- Approximation: loglikelihood of data

$$\log P(X) = \log \int_\theta P(X,\theta) d\theta$$

$$= \log \int_\theta \sum_H P(X,H,\theta) d\theta$$

$$= \log \int_\theta \sum_H P(X,H \mid \theta) P(\theta) d\theta$$

$$\geq \int_\theta \sum_H Q(H,\theta) \log \frac{P(X,H \mid \theta)P(\theta)}{Q(H,\theta)} d\theta = F(Q)$$

Where Q is a distribution with different parameterization

Variational approximation

- Approximation: loglikelihood of observable data

$$\log P(X) = F(Q) + KL(Q(H,\theta), P(H,\theta))$$

- Optimization of F(Q) is pushing up the lower bound on the loglikelihood of observable data
- How to choose Q ?

$$Q(H,\theta) = Q_\theta(\theta) Q_H(H)$$

- Then:

$$F(Q) = \int_\theta Q_\theta(\theta)\left[\sum_H Q_H(H)\log\frac{P(X,H\,|\,\theta)}{Q_H(H)}\right] d\theta$$

$$+ \int_\theta Q_\theta(\theta)\log\frac{Q_\theta(\theta)}{P(\theta)}\, d\theta \quad \longleftarrow \text{ KL distance}$$

Variational Bayes approximation

- Evaluation of $Q(H,\theta)$ is intractable
- Meanfield approximation

$$Q(H,\theta) = \prod_{k=1}^{K} Q(H_k) \prod_{i=1}^{P} (\theta_i)$$

- Allows analytical evaluation of $F(Q)$

VB learning

Learn Model with an EM like algorithm

(1) VBE – Optimize Q(H)

Estimate state of latent variables

$$Q^*_H(H) \propto \exp\langle \log P(D, H \mid \theta) \rangle_{Q_\theta(\theta)}$$

(2) VBM – Optimize Q(Θ)

Estimate parameters

$$Q^*_\theta(\theta) \propto P(\theta) \exp\langle \log P(D, H \mid \theta) \rangle_{Q_H(H)}$$