CMPUT463/563
Probabilistic Graphical Models

Supervised Learning: Markov Networks

Lili Mou

Dept. Computing Science, University of Alberta

lmou@ualberta.ca

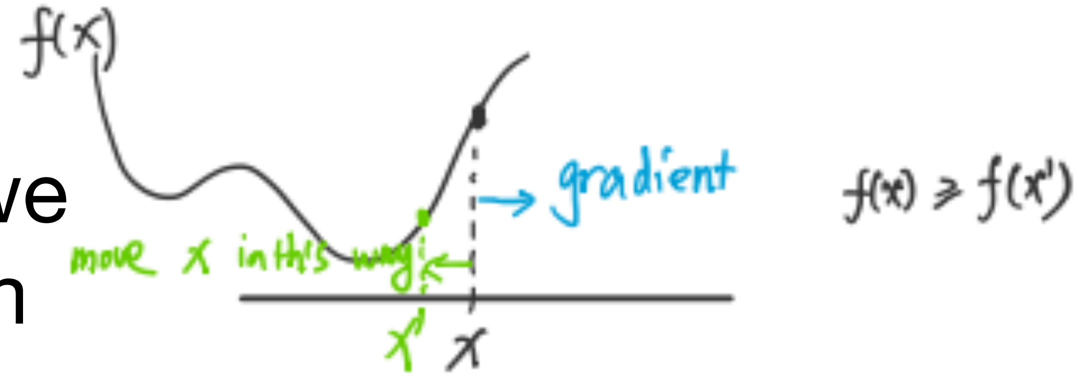
Outline

- MLE of MN does not decompose
- Exponential family is a general form of distributions
  - Gaussian, Bernoulli, and feature-based MN
  - Moment generating property
  - Gradient: Expectation in data - Expectation in model
- CRF: a conditional version of MRF
  - Requires inference for each sample
  - Still much more efficient than MRF

Note: The slides are largely derived from CMPUT466/566

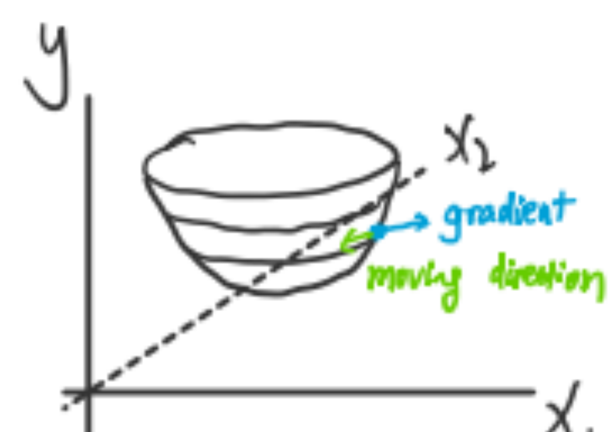
The parameters of MN do not decompose

- Log-likelihood $\log P(X) = \log \frac{\prod_k \phi_k}{\sum_x \prod_i \phi_k} = \sum_k \log \phi_k - \log \sum_x \prod_i \phi_k$
  - The first term looks good, but the second term entangles all parameters together
  - In general, MN training does not have closed-form solutions
  - We may still resort to gradient-based optimization



When learning MN parameters, we typically work with log-linear form

- ϕ must be nonnegative
- Log-linear parameters are unbounded
- Addition often more convenient than multiplication



Exponential family

Def (exponential family): A distribution belonging to the exponential family has the following form

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}^T \mathbf{T}(\boldsymbol{x}) - A(\boldsymbol{\eta})\}$$

where $A(\boldsymbol{\eta}) = \log \int h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}^T \mathbf{T}(\boldsymbol{x})\} d\boldsymbol{x}$

is a log-normalizing factor. In other words,

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = \frac{1}{Z} h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}^T \mathbf{T}(\boldsymbol{x})\}$$

where $Z = \int h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}^T \mathbf{T}(\boldsymbol{x})\} d\boldsymbol{x}$ [assuming convergent]

Here, $\boldsymbol{T}(\boldsymbol{X})$ are the sufficient statistics.

$\boldsymbol{\eta}$ are the parameters, and are called the natural parameters.

Basically, exponential family is log-linear in the sufficient statistics of $\boldsymbol{x}$, with some residual effect $h(\boldsymbol{x})$.

Apparently, MRF belongs to exponential family

Example

Gaussian

$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{1}{\sigma^2}\mu x - \left(\frac{\mu^2}{2\sigma^2} + \log\sigma\right)\right\}$$

Thus $\mathbf{T}(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$

$$\boldsymbol{\eta} = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$$

Let $\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$. Then $\mu = -\frac{\eta_1}{2\eta_2}$, $\sigma^2 = -\frac{1}{2\eta_2}$

$$A(\boldsymbol{\eta}) = \frac{\mu^2}{2\sigma^2} + \log\sigma = -\frac{\eta_1^2}{4\eta_2} + \log\sqrt{-\frac{1}{2\eta_2}}$$

$$h(x) = \frac{1}{\sqrt{2\pi}}$$

Example

Bernoulli

$$p(x;\pi) = \pi^x (1-\pi)^{1-x}$$

$$= \exp\{x \log \pi + (1-x)\log(1-\pi)\}$$

$$= \exp\{[\log \pi - \log(1-\pi)]\cdot x + \log(1-\pi)\}$$

Thus, $T(x) = x$

$$\eta = \log \frac{\pi}{1-\pi} \quad \Rightarrow \pi = \frac{e^\eta}{1+e^\eta} = \frac{1}{1+e^{-\eta}}$$

$$A(\eta) = -\log(1-\pi) = \log(1+e^\eta)$$

$$h(x) = 1$$

They are known as moment parameters, and have some correspondence with natural parameters.

$$\psi : \mu \mapsto \eta \qquad \psi^{-1} : \eta \mapsto \mu$$

A few properties

Def $$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}^T \mathbf{T}(\boldsymbol{x}) - A(\boldsymbol{\eta})\}$$

where $$A(\boldsymbol{\eta}) = \log \int h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}^T \mathbf{T}(\boldsymbol{x})\} d\boldsymbol{x}$$

- $A(\boldsymbol{\eta})$ is convex in natural parameters $\boldsymbol{\eta}$.
- $\frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \mathbb{E}[\mathbf{T}(\boldsymbol{x})]$ $\quad \frac{\partial^2 A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} = \mathrm{Var}[\mathbf{T}(\boldsymbol{x})]$

Moment generating property

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \eta_i} = \frac{1}{\log \int h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}^T \mathbf{T}(\boldsymbol{x})\} d} \int h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}^T \mathbf{T}(\boldsymbol{x})\} T_i(\boldsymbol{x}) d\boldsymbol{x}$$

$$= \int \frac{h(x) \exp\{\boldsymbol{\eta}^T \mathbf{T}(x)\}}{\log \int h(\boldsymbol{x}) \exp\{\boldsymbol{\eta}^T \mathbf{T}(x)\} d\boldsymbol{x}} T_i(\boldsymbol{x}) d\boldsymbol{x}$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\boldsymbol{x}; \boldsymbol{\eta})} [T_i(\boldsymbol{x})]$$

Gradient of the likelihood

$$\text{log-likelihood } \ell(\boldsymbol{\eta}) = \log \prod_{m=1}^{M} \left[h(\boldsymbol{x}^{(m)}) \exp\{\boldsymbol{\eta}^T \mathbf{T}(\boldsymbol{x}^{(m)}) - A(\boldsymbol{\eta})\} \right]$$

$$= \sum_{m=1}^{M} \left[\log h(\boldsymbol{x}^{(m)}) \right] + \boldsymbol{\eta}^T \sum_{m=1}^{M} \mathbf{T}(\boldsymbol{x}^{(m)}) - M A(\boldsymbol{\eta})$$

$$\frac{\partial}{\partial \boldsymbol{\eta}} \frac{1}{M} \ell(\boldsymbol{\eta}) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{T}(\boldsymbol{x}^{(m)}) - \frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}$$

$$= \mathbb{E}_{x \sim \mathcal{D}} [\mathbf{T}(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\eta}} [\mathbf{T}(\boldsymbol{x})]$$

Expectation in data **Expectation in model**

Set the gradient to 0:

Moment parameter $\hat{\boldsymbol{\mu}} = \mathbb{E}_{\boldsymbol{x} \sim \boldsymbol{\eta}} [\mathbf{T}(\boldsymbol{x})] = \mathbb{E}_{x \sim \mathcal{D}} [\mathbf{T}(\boldsymbol{x})]$

CRF: A conditional version of MRF

- MRF

$$P(x) = \frac{1}{Z}\exp\Big\{\sum_i \theta_i f_i(x)\Big\}$$

$$Z = \sum_{x'}\exp\Big\{\sum_i \theta_i f_i(x')\Big\}$$

$$\frac{\partial}{\partial\theta_i}\log P(x) = \mathbb{E}_{x\sim\mathcal{D}}[f_i(x)] - \mathbb{E}_{x\sim P(x)}[f_i(x)]$$

$\Downarrow$ Rename X to Y

Condition everything on X

- CRF

$$P(y\,|\,x) = \frac{1}{Z_x}\exp\Big\{\sum_i \theta_i f_i(y,x)\Big\}$$

$$Z_x = \sum_{y}\exp\Big\{\sum_i \theta_i f_i(y,x)\Big\}$$

$$\frac{\partial}{\partial\theta_i}\log P(\overset{y\,|\,x}{\cancel{x}}) = \mathbb{E}_{x\sim\mathcal{D}}[f_i(\overset{y,\,x}{\cancel{x}})] - \mathbb{E}_{\underset{y\,\sim\,P(y\,|\,x)}{\cancel{x\sim P(x)}}}[f_i(\overset{y,\,x}{\cancel{x}})]$$

MRF/CRF learning requires inference

- MRF $$\frac{\partial}{\partial \theta_i} \log P(\mathscr{D}) = \mathbb{E}_{x \sim \mathscr{D}}[f_i(x)] - \mathbb{E}_{\boxed{x \sim P(x)}}[f_i(x)]$$
- CRF $$\frac{\partial}{\partial \theta_i} \log P(y \mid x) = \mathbb{E}_{\boxed{x,y}}[f_i(y, x)] - \mathbb{E}_{\boxed{x \sim P(x)}}[f_i(x)]$$
- MN does not have the label bias problem. However, it comes with a cost.
  - Singleton marginal may not be enough
  - Factor MP, junction tree, or approximate inference
  - MRF: one inference for **all** samples in one gradient update
  - CRF: one inference for **one** samples
  - Which is more efficient?

MRF vs CRR

- Object detection problem
  - X: pixels, Y: labels
  - MRF: $(X, Y) \sim P(X, Y)$
  - CRF: $Y \sim P(Y | X)$
  - In practice, CRF is much more efficient than MRF
  - In this case, both requiring sampling techniques