BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Probabilistic Graphical Model

Session # 15: Learning - Markov Network

Seetha Parameswaran

seetha.p@pilani.bits-pilani.ac.in

Table of Contents

Conditional Random Field

- Random Variables $X_1, \ldots, X_n$
- Gibbs Distribution $\Phi = \{\phi_1(D_1), \ldots, \phi_k(D_k)\}$
- Un-normalized distribution

$$\tilde{P}_\Phi(X, Y) = \prod_{i=1}^{k} \phi_i(D_i)$$

- Partition function

$$Z_\Phi(X) = \sum_{Y} \tilde{P}_\Phi(X, Y)$$

- CRF

$$P_\Phi(Y \mid X) = \frac{1}{Z_\Phi(X)} \tilde{P}_\Phi(X, Y)$$

Log-Linear Representation

- Incorporate local structure in Markov Network

un-normalized

$$\tilde{P}_{\Phi}(X, Y) = \prod_{i=1}^{k} \phi_i(D_i)$$

$$= \exp\left[-\sum_j w_j f_j(D_j)\right]$$

can have multiple indicator functions on the same scope

w_j – coefficient of each feature f_j, which is used to represent the network

- Any factor can be represented by a log-linear model by including all of the appropriate features.

Example

- Binary random variables X_1 and X_2 with $\Phi(X_1, X_2) = \begin{bmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{bmatrix}$
- To represent this as a log-linear model, use indicator functions

$$f_{00} = \mathbf{1}\{X_1 = 0, X_2 = 0\}$$
$$f_{01} = \mathbf{1}\{X_1 = 0, X_2 = 1\}$$
$$f_{10} = \mathbf{1}\{X_1 = 1, X_2 = 0\}$$
$$f_{11} = \mathbf{1}\{X_1 = 1, X_2 = 1\}$$

$\phi(X_1 = 0, X_2 = 0) = a_{00}$

$f_{00}(X_1, X_2) = 1$ iff $X_1 = 0$, $X_2 = 0$ and 0 otherwise

Example

$e^{+\log w_{00}} = \underline{w_{00}}$

$w_{00} = -\log a_{00}$

- Factors can be represented as

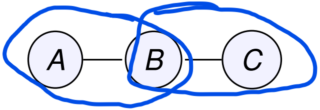
$$\Phi(X_1, X_2) = \exp\left[-\sum_{k,l} w_{kl} \underline{f^{kl}(X_1, X_2)}\right]$$

$$w_{kl} = -\log a_{kl} = \begin{bmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{bmatrix}$$

$$w_{kl} \quad \begin{bmatrix} -\log a_{00} & -\log a_{01} \\ -\log a_{10} & -\log a_{11} \end{bmatrix}$$

Table of Contents

Maximum Likelihood for Markov Networks



- Joint Distribution:

$$P_\Phi(A, B, C) = \frac{1}{Z}\underline{\phi_1(A, B)}\,\underline{\phi_2(B, C)}$$

$A[1], B[1], C[1]$, . . . $A[M], B[M], C[M]$

$A[2], B[2], C[2]$

Maximum Likelihood for Markov Networks

$P(D) = P\phi(A[1], B[1], C[1])\, P\phi(A[2], B[2], C[2]) \dots$

- Log likehood:

$$\ell(\theta : \mathcal{D}) = \sum_{m} [\ln \phi_1(a[m], b[m]) + \ln \phi_2(b[m], c[m]) - \ln Z(\theta)]$$

m instances

- Using sufficient statistics:

$$\ell(\theta : \mathcal{D}) = \sum_{a,b} \underline{M[a, b]} \ln \phi_1(a, b) + \sum_{b,c} M[b, c] \ln \phi_2(b, c) - M \ln Z(\theta)$$

- Partition function:

$$Z(\theta) = \sum_{a,b,c} \phi_1(a, b)\phi_2(b, c)$$

Maximum Likelihood for Markov Networks

$M \ln Z(\theta)$

$$P_\Phi(A,B,C) = \sum_{a,b} M[a,b] \ln \phi_1(a,b) + \sum_{b,c} M[b,c] \ln \phi_2(b,c) - M \ln \sum_{a,b,c} \phi_1(a,b)\phi_2(b,c)$$

- Partition function couples the parameters
  - No decomposition of likelihood
  - No closed form solution for optimization

Maximum Likelihood for Log-Linear Models

- Use Log-linear representation

$$P(X_1, \ldots, X_n : \theta) = \frac{1}{Z(\theta)} \exp\left[\cancel{-} \sum_{i=1}^{k} \theta_i f_i(D_i)\right]$$

- Log-likelihood

$$\ell(\theta : \mathcal{D}) = \sum_i \theta_i \left(\sum_m f_i(x[m]) \right) - M \ln Z(\theta)$$

concave

$$\ln Z(\theta) = \ln \sum_x \exp\left[- \sum_i \theta_i f_i(x) \right]$$

Log Partition Function

Convex → min

Concave → max

Theorem

$$\rightarrow \frac{\partial}{\partial\theta}\ln Z(\theta) = \underline{\mathrm{E}_\theta[f_i]}$$

$$\rightarrow \frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln Z(\theta) = \underline{Cov_\theta[f_i, f_j]}$$

ln Z(θ) is convex

− ln Z(θ) is concave

Hessian of ln Z(θ) is a covariance matrix which is always positive semidefinite

- ~~Log partition function is a Hessian;~~ hence a convex function.
- Negation of log partition function is a concave function.

Log Likelihood Function

$$\ell(\theta : \mathcal{D}) = \sum_i \theta_i \left(\sum_m f_i(x[m]) \right) - M \ln Z(\theta)$$
$$= \text{Linear function} + \text{Concave function}$$

- Log likelihood function is a concave function.
- No local optima
- Easy to optimize using hill climbing or Gradient Ascent method (L-BFGS) to obtain global optima.

Maximum Likelihood Estimation

- Log likelihood function

$$\ell(\theta : \mathcal{D}) = \sum_i \theta_i \left(\sum_m f_i(x[m]) \right) - M \ln Z(\theta)$$

$$\frac{1}{M}\ell(\theta : \mathcal{D}) = \sum_i \theta_i \left(\frac{1}{M} \sum_m f_i(x[m]) \right) - \ln Z(\theta)$$

$$\frac{\sum_m f_i(x[m])}{M}$$

- First partial derivative

$$\frac{\partial}{\partial \theta_i} \frac{1}{M}\ell(\theta : \mathcal{D}) = \mathrm{E}_D[f_i(X)] - \mathrm{E}_\theta[f_i]$$

Maximum Likelihood Estimation

Theorem

$\hat{\theta}$ is the Maximum Likelihood Estimate if and only if expectation in the data $\mathcal{D}$ equals the expectation relative to the model for each and every feature.

$$\mathrm{E}_D[f_i(X)] = \mathrm{E}_{\hat{\theta}}[f_i]$$

because $\frac{\partial}{\partial \theta_i} \frac{1}{M} \ell(\theta : D) = 0$ at optimal value

Table of Contents

Maximum Likelihood Estimation for CRF

- CRF for target Y given evidence X

$$P_\theta(Y \mid X) = \frac{1}{Z(\theta)} \hat{P}_\theta(X, Y)$$

$$Z(\theta) = \sum_{Y} \hat{P}_\theta(X, Y)$$

$$\mathcal{D} = \{x[m], y[m]\}_{m=1}^{M}$$ M data instances

Maximum Likelihood Estimation for CRF

- Log conditional likelihood

$$\ell_{Y|X}(\theta : \mathcal{D}) = \sum_{m=1}^{M} \ln P_\theta(y[m] \mid x[m], \theta)$$

$$\ell_{Y|X}(\theta : \mathcal{D}) = \sum_{i} \theta_i f_i(x[m], y[m]) - \ln Z_{x[m]}(\theta)$$

- First partial derivative

$$\frac{\partial}{\partial \theta_i} \frac{1}{M} \ell_{Y|X}(\theta : \mathcal{D}) = \frac{1}{M} \sum_{m=1}^{M} f_i(x[m], y[m]) - \mathrm{E}\left[f_i(x[m], Y]\right.$$

Maximum Likelihood Estimation for CRF

- Requires inference for each data instance $x[m]$ at each gradient step.
- Requires M inference steps.
- More expensive.
- Likelihood function is concave; optimized using gradient ascent.

Table of Contents

MAP Estimation for MRF and CRF

- MLE may over-fit the parameters to the training data.
- Hence use parameter prior to smooth out the estimates of the parameters.
- In MRF and CRF, the likelihood function cannot be maintained in closed form.
- For regularization, use MAP estimation.

Gaussian Parameter Prior

- Define a Gaussian distribution over each parameter θ_i with zero mean and a variance σ^2.

$$P(\theta : \sigma^2) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-\theta^2}{2\sigma^2}\right]$$

Laplacian Parameter Prior

- Define a Laplacian distribution over each parameter θ_i using β as the hyperparameter.

$$P(\theta : \beta) = \prod_{i=1}^{k} \frac{1}{2\beta} \exp\left[\frac{- \mid \theta \mid}{\beta}\right]$$

MAP Estimation

- MAP Estimation

$$\arg\max_{\theta} P(\mathcal{D}, \theta) = \arg\max_{\theta} P(\mathcal{D} \mid \theta) P(\theta)$$

- Find the θ that maximizes the joint distribution $P(\mathcal{D}, \theta)$

$$\arg\max_{\theta} P(\mathcal{D}, \theta) = \arg\max_{\theta} [\ell(\theta : \mathcal{D}) + \log P(\theta)]$$

MAP Estimation with Gaussian Prior

- $\log P(\theta)$ is quadratic
- L2 regularization
- Many parameters are close to zero but not exactly zero.
- Dense – many $\theta \not\approx 0$

MAP Estimation with Laplacian Prior

- $\log P(\theta)$ is linear
- L1 regularization
- Push many parameters towards zero.
- Sparse – many $\theta \approx 0$

Thank You for the support and cooperation for the entire course. :)