Georgia Institute of Technology

CSIP Center for Signal & Image Processing

ECE 8803

Parameter Learning in Graphical Models

Module 10

Markov Random Fields:

1. Known Structure &
2. Fully Observed Variables

Faramarz Fekri

Center for Signal and Information Processing

Overview

- Learning Parameters in Undirected Graphical Models (MRF):
  - Fully Observed and Given Structure:
    - Example: MLE for Markov Random Fields
  - MLE for MRF with Tabular Representation
    1. Decomposable models (triangulated) and potentials are defined on maximal cliques
    2. Non-decomposable models, or potentials are defined on non-maximal cliques
       - IPF Alg.
  - Feature-based Clique Potentials (log-linear model)
    - Gradient and GIS Algorithms.

MLE for Markov Random Fields: Example

- $P(X_1, \ldots, X_k|\theta) = \frac{1}{Z(\theta)} \exp\left(\sum_{ij} \theta_{ij} X_i X_j + \sum_i \theta_i X_i\right)$
  - $= \frac{1}{Z(\theta)} \prod_{ij} \exp(\theta_{ij} X_i X_j) \prod_i \exp(\theta_i X_i)$
  - $Z(\theta) = \sum_x \prod_{ij} \exp(\theta_{ij} X_i X_j) \prod_i \exp(\theta_i X_i)$

- $l(\theta, D) = \log(\prod_{l=1}^{N} \frac{1}{Z(\theta)} \prod_{ij} \exp(\theta_{ij} x_i{}^l x_j^l) \prod_i \exp(\theta_i x_i^l))$
- $= \sum_l^N \left(\sum_{ij} \log(\exp(\theta_{ij} x_i{}^l x_j^l)) + \sum_i \log(\exp(\theta_i x_i^l)) - \log Z(\theta) \right)$
- $= \sum_l^N \left(\sum_{ij} \theta_{ij} x_i{}^l x_j^l + \sum_i \theta_i x_i^l - \log Z(\theta) \right)$

Chapter 20 of Koller book

MLE for Markov Random Fields: Example (II)

- $l(\theta, D) = \frac{1}{N}\sum_l^N \left(\sum_{ij} \theta_{ij} x_i{}^l x_j^l + \sum_i \theta_i x_i^l - \log Z(\theta) \right)$
- $\frac{\partial l(\theta,D)}{\partial \theta_{ij}} = \frac{1}{N} \sum_l^N x_i{}^l x_j^l - \frac{\partial \log Z(\theta)}{\partial \theta_{ij}}$

$$= \frac{1}{N}\sum_l^N x_i{}^l x_j^l - \frac{1}{Z(\theta)} \frac{\partial Z(\theta)}{\partial \theta_{ij}}$$

Z couples the parameters θ_{ij}

$$= \frac{1}{N}\sum_l^N x_i{}^l x_j^l - \frac{1}{Z(\theta)} \sum_x \prod_{ij} \exp(\theta_{ij} X_i X_j) \prod_i \exp(\theta_i X_i) \; X_i X_j$$

empirical covariance matrix

need to do inference

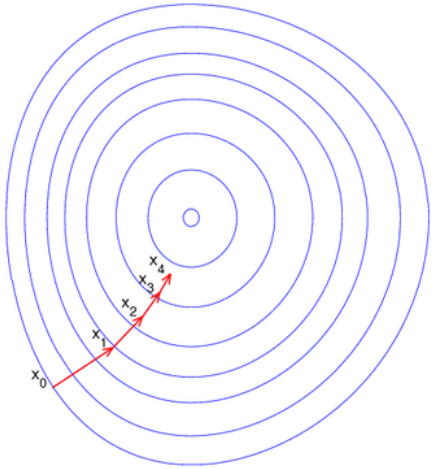
covariance matrix from model

- $\hat{P}(X_i, X_j) = \frac{1}{N}\sum_{l=1}^N \delta(X_i, x_i^l)\, \delta(X_j, x_j^l)$
- $\frac{\partial l(\theta,D)}{\partial \theta_{ij}} = E_{\hat{P}(X_i,X_j)}[X_i X_j] - E_{P(X|\theta)}[X_i X_j] = 0$

Moment Matching condition

MLE for Markov Random Fields: Example (III)

- Hessian of $\log Z(\theta)$: $\nabla^2_\theta \log Z(\theta) =$ Cov( ), positive semi-definite
  - $\max_\theta l(\theta, D)$ is a convex optimization problem.
  - Can be solve by many methods, such as gradient descent, conjugate gradient.



  - Initialize model parameters θ
  - Loop until convergence
    - Compute $\frac{\partial l(\theta,D)}{\partial \theta_{ij}} = E_{\hat{\mathrm{P}}(X_i,X_j)}[X_i X_j] - E_{P(X|\theta)}[X_i X_j]$
    - Update $\theta_{ij} \leftarrow \theta_{ij} + \eta \frac{\partial l(\theta,D)}{\partial \theta_{ij}}$

General Conclusions for MLE

- For directed graphical models, the log-likelihood decomposes into a sum of terms, node and its parents.
- For undirected graphical models, the log-likelihood does not decompose, because the normalization constant Z is a function of **all** the parameters
- In general, we will need to do inference (i.e., marginalization) to learn parameters for MRF, even in the fully observed case.

MLE for MRF with Tabular Representation (I)

$$P(x_1,\ldots,x_n)=\frac{1}{Z}\prod_{c\in C}\psi_c(\mathbf{x}_c) \qquad Z=\sum_{x_1,\ldots,x_n}\prod_{c\in C}\psi_c(\mathbf{x}_c)$$

- GOAL: MLE Estimate potential functions for each configuration: $\psi_c(\mathbf{x}_c)$
- Assume tabular representation of potentials in MRF.
- Define number of times that a configuration $\mathbf{x}$ is observed in a dataset $D=\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$ as:
  - $m(\mathbf{x}) = \sum_n \delta(\mathbf{x}, \mathbf{x}_n)$ (total count in data for a particular configuration)
  - $m(\mathbf{x}_c) = \sum_{\mathbf{x}\setminus\mathbf{x}_c} m(\mathbf{x})$ (total count for a specific configuration for a clique)
    - $m(\mathbf{x}_c)$ is computed like a marginal clique configuration count.

$$p(D|\theta)=\prod_n\prod_{\mathbf{x}} p(\mathbf{x}\,|\,\theta)^{\delta(\mathbf{x},\mathbf{x}_n)}$$

$$\log p(D|\theta)=\sum_n\sum_{\mathbf{x}}\delta(\mathbf{x},\mathbf{x}_n)\log p(\mathbf{x}\,|\,\theta)=\sum_{\mathbf{x}}\sum_n\delta(\mathbf{x},\mathbf{x}_n)\log p(\mathbf{x}\,|\,\theta)$$

$$\ell=\sum_{\mathbf{x}} m(\mathbf{x})\log\left(\frac{1}{Z}\prod_c\psi_c(\mathbf{x}_c)\right)=\sum_c\sum_{\mathbf{x}_c} m(\mathbf{x}_c)\log\psi_c(\mathbf{x}_c)-N\log Z$$

- Again, Z appears in the log-likelihood.

MLE for MRF with Tabular Representation (II)

- Computing gradients vs $\psi_c(\mathbf{x}_c)$:
  - Log-likelihood: $\ell = \sum_c \sum_{\mathbf{x}_c} m(\mathbf{x}_c) \log \psi_c(\mathbf{x}_c) - N \log Z$
  - First term: $\frac{\partial \ell_1}{\partial \psi_c(\mathbf{x}_c)} = m(\mathbf{x}_c) \Big/ \psi_c(\mathbf{x}_c)$
  - Second term:

$$\frac{\partial \log Z}{\partial \psi_c(\mathbf{x}_c)} = \frac{1}{Z} \frac{\partial}{\partial \psi_c(\mathbf{x}_c)} \left(\sum_{\tilde{\mathbf{x}}} \prod_d \psi_d(\tilde{\mathbf{x}}_d) \right)$$

Set the value of variables to $\mathbf{x}$

$$= \frac{1}{Z} \sum_{\tilde{\mathbf{x}}} \delta(\tilde{\mathbf{x}}_c, \mathbf{x}_c) \frac{\partial}{\partial \psi_c(\mathbf{x}_c)} \left(\prod_d \psi_d(\tilde{\mathbf{x}}_d) \right)$$

$$= \sum_{\tilde{\mathbf{x}}} \delta(\tilde{\mathbf{x}}_c, \mathbf{x}_c) \frac{1}{\psi_c(\tilde{\mathbf{x}}_c)} \frac{1}{Z} \prod_d \psi_d(\tilde{\mathbf{x}}_d)$$

$$= \frac{1}{\psi_c(\mathbf{x}_c)} \sum_{\tilde{\mathbf{x}}} \delta(\tilde{\mathbf{x}}_c, \mathbf{x}_c) p(\tilde{\mathbf{x}}) = \frac{p(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)}$$

-Xing

MLE for MRF with Tabular Representation (III)

- Conditions on Clique Marginals:

$$\frac{\partial \ell}{\partial \psi_c(\mathbf{x}_c)} = \frac{m(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)} - N\frac{p(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)} = 0$$

  - Unfortunately, parameter of interest $\psi_c(\mathbf{x}_c)$ is canceled out.

- We know that: $p^*_{MLE}(\mathbf{x}_c) = \frac{m(\mathbf{x}_c)}{N} \overset{\text{def}}{=} \tilde{p}(\mathbf{x}_c)$

  - That is: at the MLE setting of the parameters for each clique, the observed marginals (empirical counts) must match the model marginals.

- Observation 1: In decomposable (trangularized) MRF, where potentials are defined on maximal cliques, MLE of clique potentials equate to the empirical marginals (or conditionals) of the corresponding clique.

- Observation 2: In non-decomposable, and or the potentials are defined on non-maximal cliques, we could not equate MLE of cliques potentials to empirical marginals (or conditionals).

Decomposable Models (triangulated) and Potentials are Defined on Maximal Cliques (I)

- Decomposable models:
  - G is decomposable $\Leftrightarrow$ G is triangulated $\Leftrightarrow$ G has a junction tree
  - Potential based representation:

$$p(\mathbf{x}) = \frac{\prod_c \psi_c(\mathbf{x}_c)}{\prod_s \varphi_s(\mathbf{x}_s)}$$

- Consider a chain $X_1 - X_2 - X_3$. The cliques are (X_1, X_2) and (X_2, X_3); the separator is X_2
  - The empirical marginals must equal the model marginals.

- Let us guess that

$$\hat{p}_{MLE}(x_1, x_2, x_3) = \frac{\tilde{p}(x_1, x_2)\tilde{p}(x_2, x_3)}{\tilde{p}(x_2)}$$

  - We can verify that such a guess satisfies the conditions:

$$\hat{p}_{MLE}(x_1, x_2) = \sum_{x_3} \hat{p}_{MLE}(x_1, x_2, x_3) = \tilde{p}(x_1 \mid x_2) \sum_{x_3} \tilde{p}(x_2, x_3) = \tilde{p}(x_1, x_2)$$

  - Likewise:

$$\hat{p}_{MLE}(x_2, x_3) = \tilde{p}(x_2, x_3)$$

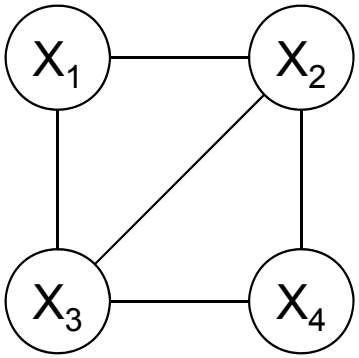
-Xing

Decomposable Models (triangulated) and Potentials are Defined on Maximal Cliques (II)

- Let us guess that $\hat{p}_{MLE}(x_1, x_2, x_3) = \frac{\tilde{p}(x_1, x_2)\tilde{p}(x_2, x_3)}{\tilde{p}(x_2)}$
- To compute the clique potentials, just equate them to the empirical marginals (or conditionals), i.e., the separator must be divided into one of its neighbors. Then $Z = 1$.

$$\hat{\psi}_{12}^{MLE}(x_1, x_2) = \tilde{p}(x_1, x_2) \qquad \hat{\psi}_{23}^{MLE}(x_2, x_3) = \frac{\tilde{p}(x_2, x_3)}{\tilde{p}(x_2)} = \tilde{p}(x_3 \mid x_2)$$

- One more example:



$$\hat{p}_{MLE}(x_1, x_2, x_3, x_4) = \frac{\tilde{p}(x_1, x_2, x_3)\tilde{p}(x_2, x_3, x_4)}{\tilde{p}(x_2, x_3)}$$

$$\hat{\psi}_{123}^{MLE}(x_1, x_2, x_3) = \frac{\tilde{p}(x_1, x_2, x_3)}{\tilde{p}(x_2, x_3)} = \tilde{p}(x_1 \mid x_2, x_3)$$

$$\hat{\psi}_{234}^{MLE}(x_2, x_3, x_4) = \tilde{p}(x_2, x_3, x_4)$$

-Xing

Non-decomposable and/or with non-maximal clique potentials: IPF Alg.

- Iterative Proportional Fitting (IPF):
  - Derivative of the likelihood: $$\frac{\partial \ell}{\partial \psi_c(\mathbf{x}_c)} = \frac{m(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)} - N\frac{p(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)}$$
  - Derive equivalent relationship: $$\frac{\tilde{p}(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)} = \frac{p(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)}$$
  - Note that ψ_c appears implicitly in the model marginal $p(\mathbf{x}_c)$.
- This is therefore a **fixed-point equation** for ψ_c.
  - Solving ψ_c in closed-form is hard, because it appears on both sides of this implicit nonlinear equation.
- The idea of IPF is to hold ψ_c fixed on the right hand side (both in the numerator and denominator) and solve for it on the left hand side. We cycle through all cliques, then iterate: (convergence to global maximum is shown)

$$\psi_c^{(t+1)}(\mathbf{x}_c) = \psi_c^{(t)}(\mathbf{x}_c)\frac{\tilde{p}(\mathbf{x}_c)}{p^{(t)}(\mathbf{x}_c)}$$

← Need to do inference here

-Xing

MRF: MLE in Log-Linear Model (I)

- For large cliques, general potentials (which are parameterized by general **tabular** potential functions) have exponential complexity for inference and also have exponential numbers of parameters (that we must learn from limited data).
- Feature-based models use a less general parameterization of the clique potentials.

- Let $F=\{f_i(\mathbf{D}_i)\}_{i=1,\ldots,k}$, be a given set of features, where $f_i(\mathbf{D}_i)$ is a feature function defined over the variables in clique $\mathbf{D}_i$.
- Note that θ_i *are the* parameters we need to estimate.

$$P(X_1,\ldots,X_n:\boldsymbol{\theta})=\frac{1}{Z(\boldsymbol{\theta})}\exp\left\{\sum_{i=1}^{k}\theta_i f_i(\mathbf{D}_i)\right\}$$

- This is indeed the exponential family, with features as sufficient statistics.

MRF: MLE in Log-Linear Model (II)

- Consider a clique $\mathbf{D}_i$ of random variables in a MRF, e.g. three consecutive characters $c_1c_2c_3$ in a English text.
- How would we build a model of $p(c_1c_2c_3)$?
  - If we use a single clique function for $c_1c_2c_3$, the full joint clique potential would be huge: (26^3-1) parameters.
  - However, we often know that particular joint settings of variables in a clique are quite likely or quite unlikely. e.g. ion, ing, jxk, zwy.
- A "feature" is a function which is in "don't care" over all joint settings except a few particular ones on which it is high or low.
  - For example, we might have $f_{ion}(c_1c_2c_3)$ which is 1 if the string is "ion" and 0 otherwise, and similar features for "jxk", etc.
- We can also define features for continuous random variable inputs.
- Each feature has a weight θ_i which represents the numerical strength of feature and whether it increases or decreases probability of the clique.

MRF: MLE in Log-Linear Model (III)

- Let D be a data set of M instances D={ξ[1],…, ξ[M]}.
- Log-Likelihood:

$$l(\boldsymbol{\theta}:D)=\sum_i \theta_i\left(\sum_m f_i(\xi[m])\right)-M\ln Z(\boldsymbol{\theta})$$

- Sufficient statistics: sums of the feature values in the instances in D
- Normalizing by M:

$$\frac{1}{M}l(\boldsymbol{\theta}:D)=\sum_i \theta_i \mathbf{E}_D[f_i[\mathbf{d}_i]]-\ln Z(\boldsymbol{\theta})$$

- where $\mathbf{E}_D[f_i(\mathbf{d}_i)]$ is the empirical expectation of f_i, that is, its average in the data set.

MRF: MLE in Log-Linear Model (IV)

$$l(\boldsymbol{\theta}:D)=\sum_i \theta_i\left(\sum_m f_i(\xi[m])\right)-M\ln Z(\boldsymbol{\theta})$$

- Note likelihood function is a sum of two functions:
  - The first function is linear in parameters (i.e., increasing the parameters directly increases this term)
  - The second term, the log of partition function, is convex in the parameters θ_i because:
    - Hessian is equal to $\nabla^2_\theta \log Z(\theta) =$ Cov( ), positive semi-definite
- Likelihood function is convex in θ_i :

$$\frac{1}{M}l(\boldsymbol{\theta}:D)=\sum_i \theta_i \mathbf{E}_D[f_i[\mathbf{d}_i]]-\ln Z(\boldsymbol{\theta})$$

  - Setting gradient to zero give maximum points, computing gradient:

$$\frac{\partial}{\partial\theta_i}\frac{1}{M}l(\boldsymbol{\theta}:D)=\mathbf{E}_D[f_i[\mathbf{d}_i]]-\frac{1}{Z(\boldsymbol{\theta})}\sum_\xi \frac{\partial}{\partial\theta_i}\exp\left\{\sum_i \theta_i f_i(\xi)\right\}$$

$$=\mathbf{E}_D[f_i[\mathbf{d}_i]]-\sum_\xi f_i(\xi)\frac{1}{Z(\boldsymbol{\theta})}\exp\left\{\sum_i \theta_i f_i(\xi)\right\}=\mathbf{E}_D[f_i[\mathbf{d}_i]]-\mathbf{E}_{\boldsymbol{\theta}}[f_i]$$

MRF: MLE in Log-Linear Model (V)

$$\frac{\partial}{\partial \theta_i} l(\boldsymbol{\theta} : D) = M\mathbf{E}_D[f_i[\mathbf{d}_i]] - M\mathbf{E}_{\boldsymbol{\theta}}[f_i]$$

Average number of times feature f_i is true in data D

Average number of times feature f_i is true according to model

- The gradient of the log likelihood is rewritten as the expected feature vector according to the empirical distribution minus the model's expectation of the feature vector.
  - At the optimum, the gradient will be zero, so the empirical distribution of the features will match the model's predictions (Moment matching Problem).
    - Numerical optimization: gradient ascent method or 2nd order-based (Newton's) method, i.e. requires inference at each step (hence slow)

- Update $\theta_{ij} \leftarrow \theta_{ij} + \eta \frac{\partial l(\theta, D)}{\partial \theta_{ij}}$: Gradient method

MRF: MLE in Log-Linear Model (VI)

- Alternative to gradient method:
  - Generalized Iterative Scaling (GIS):

$$\tilde{\ell}(\theta;D) = \ell(\theta;D)/N = \frac{1}{N}\sum_n \log p(x_n \mid \theta)$$

X_n is entire set of variables

$$= \sum_x \tilde{p}(x) \log p(x \mid \theta) \quad = \sum_x \tilde{p}(x) \sum_i \theta_i f_i(x) - \log Z(\theta)$$

Where $\tilde{p}(x)$ is the observed (empirical) marginal

  - Instead of optimizing this objective directly, use its lower bound:
    - Logarithm has a linear upper bound $\log Z(\theta) \le \mu Z(\theta) - \log \mu - 1$
    - This bound holds for all μ, in particular, for $\mu = Z^{-1}(\theta^{(t)})$

$$\tilde{\ell}(\theta;D) \ge \sum_x \tilde{p}(x) \sum_i \theta_i f_i(x) - \frac{Z(\theta)}{Z(\theta^{(t)})} - \log Z(\theta^{(t)}) + 1$$

Generalized Iterative Scaling (GIS)

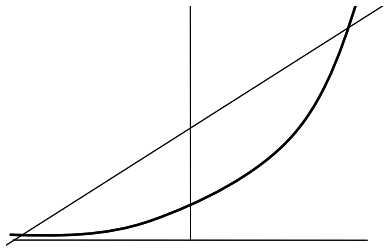
- Define $\Delta\theta_i^{(t)} \overset{\text{def}}{=} \theta_i - \theta_i^{(t)}$

$$\tilde{\ell}(\theta;D) \geq \sum_x \tilde{p}(x)\sum_i \theta_i f_i(x) - \frac{1}{Z(\theta^{(t)})}\sum_x \exp\left\{\sum_i \theta_i f_i(x)\right\} - \log Z(\theta^{(t)}) + 1$$

$$= \sum_i \theta_i \sum_x \tilde{p}(x) f_i(x) - \frac{1}{Z(\theta^{(t)})}\sum_x \exp\left\{\sum_i \theta_i^{(t)} f_i(x)\right\}\exp\left\{\sum_i \Delta\theta_i^{(t)} f_i(x)\right\} - \log Z(\theta^{(t)}) + 1$$

$$= \sum_i \theta_i \sum_x \tilde{p}(x) f_i(x) - \sum_x p(x \mid \theta^{(t)})\exp\left\{\sum_i \Delta\theta_i^{(t)} f_i(x)\right\} - \log Z(\theta^{(t)}) + 1$$

- Relax again
  - Assume $f_i(x) \geq 0, \quad \sum_i f_i(x) = 1$
  - Convexity of exponential: $\exp\left(\sum_i \pi_i x_i\right) \leq \sum_i \pi_i \exp(x_i)$



- We have:

$$\tilde{\ell}(\theta;D) \geq \sum_i \theta_i \sum_x \tilde{p}(x) f_i(x) - \sum_x p(x \mid \theta^{(t)}) \sum_i f_i(x)\exp\left(\Delta\theta_i^{(t)}\right) - \log Z(\theta^{(t)}) + 1 \overset{\text{def}}{=} \Lambda(\theta)$$

Generalized Iterative Scaling (GIS)

- Take derivative: $\frac{\partial \Lambda}{\partial \theta_i} = \sum_x \tilde{p}(x) f_i(x) - \exp\left(\Delta\theta_i^{(t)}\right) \sum_x p(x \mid \theta^{(t)}) f_i(x)$
- Set to zero

$$e^{\Delta\theta_i^{(t)}} = \frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p(x \mid \theta^{(t)}) f_i(x)} = \frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p^{(t)}(x) f_i(x)} Z(\theta^{(t)})$$

  - where $p^{(t)}(x)$ is the unnormalized version of $p(x|\theta^{(t)})$

Multiply to $f_i(x)$ and apply sum and then take exp from both sides

- Update

$$\theta_i^{(t+1)} = \theta_i^{(t)} + \Delta\theta_i^{(t)} \Rightarrow p^{(t+1)}(x) = p^{(t)}(x) \prod_i e^{\Delta\theta_i^{(t)} f_i(x)}$$

$$p^{(t+1)}(x) = \frac{p^{(t)}(x)}{Z(\theta^{(t)})} \prod_i \left(\frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p^{(t)}(x) f_i(x)} Z(\theta^{(t)}) \right)^{f_i(x)}$$

$$\Rightarrow = \frac{p^{(t)}(x)}{Z(\theta^{(t)})} \prod_i \left(\frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p^{(t)}(x) f_i(x)} \right)^{f_i(x)} \left(Z(\theta^{(t)}) \right)^{\sum_i f_i(x)}$$

$$= p^{(t)}(x) \prod_i \left(\frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p^{(t)}(x) f_i(x)} \right)^{f_i(x)} \Rightarrow \theta_i^{(t+1)} = \theta_i^{(t)} + \log\left(\frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p^{(t)}(x) f_i(x)} \right)$$

-Xing