Max-Margin Markov Networks

Ben Taskar, Carlos Guestrin, Daphne Koller

Presented by
Lu Wang
Cornell University
Department of Computer Science

Review

- Problem
  - Learning tasks have complex output spaces
- Structural SVM
  - Notation: $\vec{\omega}$, $\psi(X,Y)$
  - Prediction: $f(X) = \arg\max_{Y}\{\vec{\omega}\cdot\psi(X,Y)\}$
  - Soft-Margin Struct SVM(Margin Rescaling)

$$\min_{\vec{\omega},\vec{\xi}} \frac{1}{2}\vec{\omega}^T\vec{\omega} + C\sum_{i=1}^{n}\xi_i$$
$$s.t. \forall y \in Y \setminus y_1 : \vec{\omega}^T\psi(x_1,y_1) \ge \vec{\omega}^T\psi(x_1,y) + \Delta(y_1,y) - \xi_1$$
$$\ldots$$
$$\forall y \in Y \setminus y_1 : \vec{\omega}^T\psi(x_1,y_1) \ge \vec{\omega}^T\psi(x_1,y) + \Delta(y_1,y) - \xi_1$$

Goals of Paper

- This paper proposes Maximum Margin Markov(M3)Networks
  - That incorporates the advantages of SVM
    - Using kernels to deal with high-dimensional features efficiently
    - Having strong generalization guarantees
  - That incorporates the advantage of probabilistic graphical model
    - Having ability to capture correlations in structured data

Outline

- Structure in classification problem
  - How to construct the model to integrate the kernel models with graphical models?
- Margin-based structured classification
- Exploiting structure in M3 networks
  - How to reduce the number of constraints from exponential to polynomial?
- SMO learning of M3 networks
  - How to deal with the massive matrix when solving the QP?

Structure in classification problem

- Markov network(pairwise Markov Network)
  - Defined as a graph: $G = (Y, E)$
  - Potential: $\psi_{ij}(\mathrm{x}, y_i, y_j)$, corresponding to edge(i,j)
  - The network encodes a joint conditional probability distribution as

$$P(\mathrm{y}|\mathrm{x}) \propto \Pi_{(i,j)\in E}\psi_{ij}(\mathrm{x}, y_i, y_j)$$

Probabilistic graphical model

  - A set of features

$$f_k(\mathrm{x},\mathrm{y}) = \sum_{(i,j)\in E} f_k(\mathrm{x}, y_i, y_j)$$

Train W using struct SVM

  - The network potentials are

$$\psi_{ij}(\mathrm{x}, y_i, y_j) = \exp[\sum_{k=1}^{n} w_k f_k(\mathrm{x}, y_i, y_j)] = \exp[\mathrm{w}^T\mathrm{f}(\mathrm{x}, y_i, y_j)]$$

Margin-based structured classification

- Primal formulation

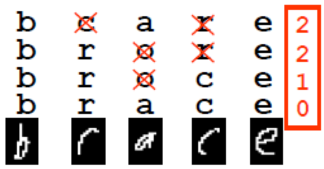
$$\min \frac{1}{2}\|w\|^2 + C\sum_x \xi_x$$
$$s.t.\ \mathrm{w}^T\Delta\mathrm{f}_x(\mathrm{y}) \ge \Delta\mathrm{t}_x(\mathrm{y}) - \xi_x, \forall \mathrm{x},\mathrm{y}$$

1. Integrate per-label loss, such as the proportion of incorrect labels predicted
2. Integrate slack variable

  - $\Delta\mathrm{f}_x(\mathrm{y}) = \mathrm{f}(\mathrm{x},\mathrm{t}(\mathrm{x})) - \mathrm{f}(\mathrm{x},\mathrm{y}) = \sum_{(i,j)} \Delta\mathrm{f}_x(y_i, y_j)$
  - $\Delta\mathrm{t}_x(\mathrm{y}) = \sum_{i=1}^{l} \Delta\mathrm{t}_x(y_i)$
- Dual formulation

$$\max \sum_{x,y}\alpha_x(\mathrm{y})\Delta\mathrm{t}_x(\mathrm{y}) - \frac{1}{2}\|\sum_{x,y}\alpha_x(\mathrm{y})\Delta\mathrm{f}_x(\mathrm{y})\|^2$$
$$\mathrm{s.t.} \sum_{\mathrm{y}}\alpha_x(\mathrm{y}) = C, \forall \mathrm{x}; \alpha_x(\mathrm{y}) \ge 0, \forall \mathrm{x},\mathrm{y}$$

Margin-based structured classification



Taskar 05

Exploiting structure in M3 networks(1/7)

- Reconsider the dual formulation

$$\max \sum_{x,y} \alpha_x(y)\Delta t_x(y) - \frac{1}{2}\|\sum_{x,y}\alpha_x(y)\Delta f_x(y)\|^2$$
$$\text{s.t.} \sum_y \alpha_x(y) = C, \forall x; \alpha_x(y) \geq 0, \forall x,y$$

  - If we interpret the variables $\alpha_x(y)$ as a density function over y conditional on x, the dual objective is a function of expectations of $\Delta t_x(y)$ and $\Delta f_x(y)$

Exploiting structure in M3 networks(2/7)

- Find an instrument
  - Since $\Delta t_x(y)=\sum_{i=1}^{l}\Delta t_x(y_i)$ and $\Delta f_x(y)=\sum_{(i.j)}\Delta f_x(y_i,y_j)$ are sums of functions over nodes and edges, we only need node and edge marginals of the measure $\alpha_x(y)$ to compute their expectations
  - Define

$$\mu_x(y_i,y_j)=\sum_{y\sim[y_i,y_j]}\alpha_x(y), \forall(i,j)\in E, \forall y_i,y_j, \forall x$$
$$\mu_x(y_i)=\sum_{y\sim[y_i]}\alpha_x(y), \forall i, \forall y_i, \forall x$$

Exploiting structure in M3 networks(3/7)

- Reform the dual formulation

$$\max \sum_{x,y} \alpha_x(y)\Delta t_x(y) - \frac{1}{2}\|\sum_{x,y}\alpha_x(y)\Delta f_x(y)\|^2$$
$$\text{s.t.} \sum_y \alpha_x(y) = C, \forall x; \alpha_x(y) \geq 0, \forall x,y$$

Magic process

- As to the first term

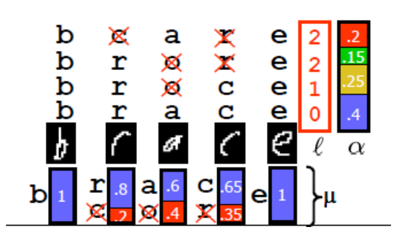
$$\sum_y \alpha_x(y)\Delta t_x(y)=\sum_y\sum_i \alpha_x(y)\Delta t_x(y_i)=\sum_{i,y_i}\Delta t_x(y_i)\sum_{y\sim[y_i]}\alpha_x(y)=\sum_{i,y_i}\mu_x(y_i)\Delta t_x(y_i)$$

- As to the second term

$$\sum_y \alpha_x(y)\Delta f_x(y)=\sum_y\sum_{i,j}\alpha_x(y)\Delta f_x(y_i,y_j)=\sum_{\substack{(i,j)\\ y_i,y_j}}\Delta f_x(y_i,y_j)\sum_{y\sim[y_i,y_j]}\alpha_x(y)=\sum_{\substack{(i,j)\\ y_i,y_j}}\mu_x(y_i,y_j)\Delta f_x(y_i,y_j)$$

Exploiting structure in M3 networks(4/7)

- The connection



Taskar 05

Exploiting structure in M3 networks(5/7)

Attention

We must enforce consistency between the pairwise and singleton marginals, that is,

$$\sum_{y_i}\mu_x(y_i,y_j)=\mu_x(y_j), \forall y_j, \forall(i,j)\in E, \forall x$$
$$\sum_{y_i}\mu_x(y_i)=C$$

Exploiting structure in M³ networks(6/7)

- Then, we get the equivalent factored dual QP

$$\max \sum_{x}\sum_{i,y_i} \mu_x(y_i)\Delta t_x(y_i) - \frac{1}{2}\sum_{x,\hat{x}}\sum_{i,j}\sum_{r,s} \mu_x(y_i,y_j)\mu_{\hat{x}}(y_r,y_s)\Delta f_x(y_i,y_j)^T \Delta f_{\hat{x}}(y_r,y_s)$$

$$\text{s.t.} \sum_{y_i} \mu_x(y_i,y_j) = \mu_x(y_j); \sum_{y_i} \mu_x(y_i) = C; \mu_x(y_i,y_j) \ge 0$$

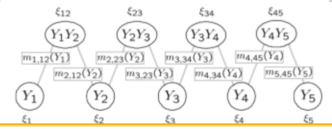
- And the factored primal

$$\min \ \frac{1}{2}\|w\|^2 + C\sum_{x}\sum_{i}\xi_{x,i} + C\sum_{x}\sum_{(i,j)}\xi_{x,ij}$$

$$s.t. \ w^T \Delta f_x(y_i,y_j) + \sum_{(i',j):i'\ne i} m_{x,i'}(y_j) + \sum_{(j',i):j'\ne j} m_{x,j'}(y_i) \ge -\xi_{x,ij};$$

$$\sum_{(i,j)} m_{x,j}(y_i) \ge \Delta t_x(y_i) - \xi_{x,i}; \xi_{x,ij} \ge 0, \xi_{x,i} \ge 0$$

Exploiting structure in M³ networks(7/7)



$$\mathbf{w}^\top \mathbf{f}_{12}(y_1^{(*)}, y_2^{(*)}) + \xi_{12} \ \ge \ \mathbf{w}^\top \mathbf{f}_{12}(y_1, y_2) + m_{1,12}(y_1) + m_{2,12}(y_2), \ \forall y_1, y_2;$$

$$\mathbf{w}^\top \mathbf{f}_{23}(y_2^{(*)}, y_3^{(*)}) + \xi_{23} \ \ge \ \mathbf{w}^\top \mathbf{f}_{23}(y_2, y_3) + m_{2,23}(y_2) + m_{3,23}(y_3), \ \forall y_2, y_3;$$

$$\mathbf{w}^\top \mathbf{f}_{34}(y_3^{(*)}, y_4^{(*)}) + \xi_{34} \ \ge \ \mathbf{w}^\top \mathbf{f}_{34}(y_3, y_4) + m_{3,34}(y_3) + m_{4,34}(y_4), \ \forall y_3, y_4;$$

$$\mathbf{w}^\top \mathbf{f}_{45}(y_4^{(*)}, y_5^{(*)}) + \xi_{45} \ \ge \ \mathbf{w}^\top \mathbf{f}_{45}(y_4, y_5) + m_{4,45}(y_4) + m_{5,45}(y_5), \ \forall y_4, y_5.$$

$$m_{1,12}(y_1) \ge I(y_1 \ne y_1^{(*)}) - \xi_1, \forall y_1$$

$$m_{2,12}(y_2) + m_{2,23}(y_2) \ge I(y_2 \ne y_2^{(*)}) - \xi_2, \forall y_2$$

$$m_{3,23}(y_3) + m_{3,34}(y_3) \ge I(y_3 \ne y_3^{(*)}) - \xi_3, \forall y_3$$

$$m_{4,34}(y_4) + m_{4,45}(y_4) \ge I(y_4 \ne y_4^{(*)}) - \xi_4, \forall y_4$$

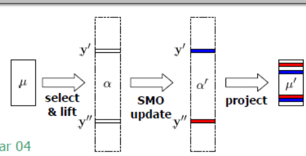
$$m_{5,45}(y_5) \ge I(y_5 \ne y_5^{(*)}) - \xi_5, \forall y_5$$

SMO learning of M³ networks

- The SMO approach solves this QP by analytically optimizing two-variable subproblems.
- Take any two variables $\alpha_x(y^1), \alpha_x(y^2)$ and move weight from one to another

$$\lambda = \alpha'_x(y^1) - \alpha_x(y^1) = \alpha_x(y^2) - \alpha'_x(y^2)$$

$$\mu'_x(y_i,y_j) = \mu_x(y_i,y_j) + \lambda I(y_i = y_i^1, y_j = y_j^1) - \lambda I(y_i = y_i^2, y_j = y_j^2)$$



*Taskar 04

Summary

- Max-Margin Markov Networks
  - integrates the kernel methods with the graphical models
- Reduce exponential constraints and variables to polynomial by
  - Using marginal dual variables
- Solve the QP by
  - SMO approach, specifically, by analytically optimizing two-variable subproblems

The End

Thanks!