Max-Margin Markov Networks

Ben Taskar
Carlos Guestrin
Daphne Koller

Presenter: Nam Nguyen

Overview

- *Max-margin Markov networks captures the strengths of the two frameworks:*
  - *Kernel-based approaches (i.e. SVMs): the ability to use high-dimensional feature spaces; strong theoretical guarantees.*
  - *Graphical models (i.e. Markov networks): the ability to represent correlations between labels.*

Problem Formulation

- Multi-label supervised learning:
  - Input: $\mathbf{X} = (x_1, x_2, ..., x_l)$
  - Output: $\mathbf{Y} = (Y_1, Y_2, ..., Y_l)$
- Example problem:

$\mathbf{X}$ $\mathbf{Y}$

embraces

Discriminative vs. Generative Models

- Linear Discriminative Model:

$$h_{\mathbf{w}}(\mathbf{x}) = \arg\max_{\mathbf{y}} \sum_{i=1}^{n} w_i f_i(\mathbf{x}, \mathbf{y}) = \arg\max_{\mathbf{y}} \mathbf{w}^T \mathbf{f}(\mathbf{x}, \mathbf{y})$$

- Generative Model:

$$\bar{\mathbf{y}} = \arg\max_{\mathbf{y}} P(\mathbf{x}, \mathbf{y}) = \arg\max_{\mathbf{y}} P(\mathbf{y} \mid \mathbf{x}) P(\mathbf{x}) = \arg\max_{\mathbf{y}} P(\mathbf{y} \mid \mathbf{x})$$

Markov Network

- A pairwise Markov network
  - A graph G=(Y,E)
  - A potential function associated with each edge

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

- The joint conditional probability distribution

$$P(\mathbf{y} \mid \mathbf{x}) \propto \prod_{(i,j) \in \mathbf{E}} \psi_{ij}(\mathbf{x}, y_i, y_j) = \prod_{(i,j) \in \mathbf{E}} \exp[\sum_{k=1}^{n} w_k f_k(\mathbf{x}, y_i, y_j)]$$

$$= \exp[\sum_{(i,j) \in \mathbf{E}} \sum_{k=1}^{n} w_k f_k(\mathbf{x}, y_i, y_j)] = \exp[\sum_{k=1}^{n} w_k \underbrace{\sum_{(i,j) \in \mathbf{E}} f_k(\mathbf{x}, y_i, y_j)}]$$

$$= \exp[\sum_{k=1}^{n} w_k f_k(\mathbf{x}, \mathbf{y})] = \exp(\mathbf{w}^T \mathbf{f}(\mathbf{x}, \mathbf{y}))$$

Marginal-based Markov networks

- Goal:

$$\arg\max_{\mathbf{y}} P(\mathbf{y} \mid \mathbf{x}) = \arg\max_{\mathbf{y}} [\exp(\mathbf{w}^T \mathbf{f}(\mathbf{x}, \mathbf{y}))] = \arg\max_{\mathbf{y}} \mathbf{w}^T \mathbf{f}(\mathbf{x}, \mathbf{y})$$

- Start with formulation of SVMs for a *single-label* binary classification:

$$\begin{aligned} &\text{maximize} \quad \gamma \\ &\text{s.t.} \quad \|\mathbf{w}\| \leq 1; \quad \mathbf{w}^\top \Delta \mathbf{f}_{\mathbf{x}}(\mathbf{y}) \geq \gamma, \quad \forall \mathbf{x} \in S, \quad \forall \mathbf{y} \neq \mathbf{t}(\mathbf{x}) \end{aligned}$$

where $\Delta \mathbf{f}_{\mathbf{x}}(\mathbf{y}) = \mathbf{f}(\mathbf{x}, \mathbf{t}(\mathbf{x})) - \mathbf{f}(\mathbf{x}, \mathbf{y})$

Margin-based Markov networks

- In the multi-label setting, the margin between **t**(**x**) and **y** scales linearly with the number of wrong labels in **y**, $\Delta\mathbf{t}_\mathbf{x}(\mathbf{y})$:

$$\begin{aligned}&\text{maximize} \quad \gamma \\ &\text{s.t.} \quad \|\mathbf{w}\| \le 1; \quad \mathbf{w}^\top \Delta\mathbf{f}_\mathbf{x}(\mathbf{y}) \ge \gamma\, \Delta\mathbf{t}_\mathbf{x}(\mathbf{y}), \quad \forall \mathbf{x} \in S, \quad \forall \mathbf{y}\end{aligned}$$

where $\Delta\mathbf{t}_\mathbf{x}(\mathbf{y}) = \sum_{i=1}^{l} \Delta\mathbf{t}_\mathbf{x}(y_i)$ and $\Delta\mathbf{t}_\mathbf{x}(y_i) \equiv I(y_i \neq (\mathbf{t}(\mathbf{x}))_i)$

- The **QP** form:

$$\begin{aligned}&\text{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{s.t.} \quad \mathbf{w}^\top \Delta\mathbf{f}_\mathbf{x}(\mathbf{y}) \ge \Delta\mathbf{t}_\mathbf{x}(\mathbf{y}), \quad \forall \mathbf{x} \in S, \ \forall\, \mathbf{y}\end{aligned}$$

Margin-based Markov networks

- Introduction of slack variables, the primal formulation:

$$\begin{aligned}&\min \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_\mathbf{x} \xi_\mathbf{x}\,; \\ &\text{s.t.} \quad \mathbf{w}^\top \Delta\mathbf{f}_\mathbf{x}(\mathbf{y}) \ge \Delta\mathbf{t}_\mathbf{x}(\mathbf{y}) - \xi_\mathbf{x}, \ \forall \mathbf{x}, \mathbf{y}\end{aligned}$$

- The dual formulation:

$$\begin{aligned}&\max \sum_{\mathbf{x},\mathbf{y}} \alpha_\mathbf{x}(\mathbf{y})\Delta\mathbf{t}_\mathbf{x}(\mathbf{y}) - \frac{1}{2}\left\|\sum_{\mathbf{x},\mathbf{y}} \alpha_\mathbf{x}(\mathbf{y})\Delta\mathbf{f}_\mathbf{x}(\mathbf{y})\right\|^2; \\ &\text{s.t.} \quad \sum_\mathbf{y} \alpha_\mathbf{x}(\mathbf{y}) = C, \forall \mathbf{x}; \quad \alpha_\mathbf{x}(\mathbf{y}) \ge 0\,, \forall \mathbf{x}, \mathbf{y}.\end{aligned}$$

Exploiting Structure in M^3 networks

- Note: the number of constraints in the primal QP and the number of variables in the dual QP are exponential in the number of labels.
- Main Idea: the variables $\alpha_\mathbf{x}(\mathbf{y})$ in the dual formulation can be interpreted as a density function over **y** conditional on **x**.

Exploiting Structure in M3 networks

- Define the marginal dual variables:

$$\begin{aligned}\mu_\mathbf{x}(y_i, y_j) &= \textstyle\sum_{\mathbf{y}\sim[y_i,y_j]} \alpha_\mathbf{x}(\mathbf{y}), \quad \forall\,(i,j) \in E, \forall y_i, y_j, \ \forall\,\mathbf{x}; \\ \mu_\mathbf{x}(y_i) &= \textstyle\sum_{\mathbf{y}\sim[y_i]} \alpha_\mathbf{x}(\mathbf{y}), \quad \forall\, i, \ \forall y_i, \ \forall\,\mathbf{x};\end{aligned}$$

- Require conditions:
  - Consistency: $\sum_{y_i} \mu_\mathbf{x}(y_i, y_j) = \mu_\mathbf{x}(y_j), \ \forall y_j, \quad \forall (i,j) \in E, \ \forall \mathbf{x}$
  - The Markov network is a forest.
- The equivalent dual QP:

$$\max \quad \sum_\mathbf{x}\sum_{i,y_i} \tilde{\mu}_\mathbf{x}(y_i)\Delta\mathbf{t}_\mathbf{x}(y_i) - \frac{1}{2}\sum_{\mathbf{x},\tilde{\mathbf{x}}}\sum_{\substack{(i,j)\\(r,s)}}\sum_{\substack{y_i,y_j\\y_r,y_s}} \mu_\mathbf{x}(y_i,y_j)\tilde{\mu}_{\tilde{\mathbf{x}}}(y_r,y_s)\mathbf{f}_\mathbf{x}(y_i,y_j)^\top \mathbf{f}_{\tilde{\mathbf{x}}}(y_r,y_s);$$

$$\text{s.t.} \quad \sum_{y_i} \mu_\mathbf{x}(y_i,y_j) = \mu_\mathbf{x}(y_j); \quad \sum_{y_i} \mu_\mathbf{x}(y_i) = C; \quad \mu_\mathbf{x}(y_i,y_j) \ge 0. \qquad (10)$$

Generalization Error Bound

- A γ-*margin per-label loss* measures the worst per-label loss on **x** made by any classifier **z** which is perturbed from $\mathbf{w}^\top\mathbf{f}_\mathbf{x}$ by at most a γ-margin per-label.

$$\mathcal{L}^\gamma(\mathbf{w},\mathbf{x}) = \sup_{\mathbf{z}:\ |\mathbf{z}(\mathbf{y}) - \mathbf{w}^\top\mathbf{f}_\mathbf{x}(\mathbf{y})| \le \gamma\Delta\mathbf{t}_\mathbf{x}(\mathbf{y});\ \forall\mathbf{y}} \tfrac{1}{7}\Delta\mathbf{t}_\mathbf{x}(\arg\max_\mathbf{y} \mathbf{z}(\mathbf{y}))$$

Theorem 6.1 *If the edge features have bounded 2-norm,* $\max_{(i,j),y_i,y_j} \|\mathbf{f}_\mathbf{x}(y_i,y_j)\|_2 \le R_{edge}$*, then for a family of hyperplanes parameterized by* $\mathbf{w}$*, and any* $\delta > 0$*, there exists a constant* K *such that for any* $\gamma > 0$ *per-label margin, and* $m > 1$ *samples, the per-label loss is bounded by:*

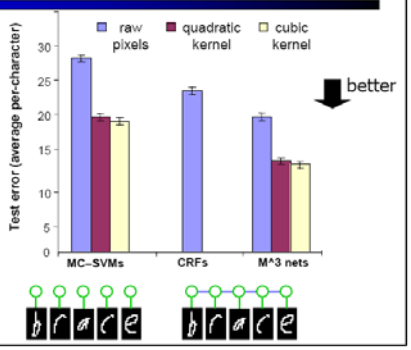
$$E_\mathbf{x}\mathcal{L}(\mathbf{w},\mathbf{x}) \ \le\ E_S\mathcal{L}^\gamma(\mathbf{w},\mathbf{x}) + \sqrt{\frac{K}{m}\left[\frac{R_{edge}^2\|\mathbf{w}\|_2^2 q^2}{\gamma^2}[\ln m + \ln l + \ln q + \ln k] + \ln\frac{1}{\delta}\right]}\ ;$$

with probability at least $1-\delta$*, where* $q = \max_i |\{(i,j) \in E\}|$ *is the maximum edge degree in the network,* k *is the number of classes in a label, and* l *is the number of labels.* ∎

Handwriting Recognition

Length: ~8 chars
Letter: 16x8 pixels
10-fold Train/Test
5000/50000 letters
600/6000 words

Models:
Multiclass-SVMs
CRFs
M³ nets



*Taskar et al 03

Conclusion

- M3 networks integrate kernel methods with graphical models.
- M3 networks effectively train by exploiting the network structure, i.e. a forest.
- Authors provide theoretical guarantees on the average *per-label* generalization error of the model in term of training set margin.