Max-Margin Markov Networks

Ben Taskar
Carlos Guestrin
Daphne Koller

Main Contribution

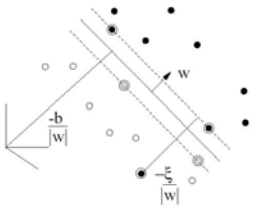
- The authors combine a graphic model and a discriminative model and apply it in a sequential learning setting.
  - Graphic models: better at interpreting data, worse performance
  - Discriminative models: better performance, unintelligible working mechanism

SVM

- SVM officially proposed as a QP problem

$$\begin{aligned} &\text{Find} && \mathbf{w}, \xi \\ &\text{Minimize} && \|\mathbf{w}\|^2 + C \sum_i \xi_i \\ &\text{Subject to} && \\ & && y_i \cdot (\mathbf{w} \cdot \mathbf{x}_i + b) + \xi_i \geq 1 \end{aligned}$$

- Schematic plot



SVM (2)

- Having learned w, our discriminant function is defined as

  $h(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$

- One way to extend binary to multiclass SVM is to train a weight vector ***w*** for each class, and

  $h(\mathbf{x}) = \text{argmax}_r\, (w_r * x + b_r),\ r = 1..k$

SVM (3)

- Multiclass SVM (Crammer & Singer)

$$\min_{M,\xi} \quad \frac{1}{2}\beta\|M\|_2^2 + \sum_{i=1}^{m}\xi_i$$

$$\text{subject to}: \quad \forall i, r \quad \bar{M}_{y_i}\cdot\bar{x}_i + \delta_{y_i,r} - \bar{M}_r\cdot\bar{x}_i \geq 1 - \xi_i$$

  where M is the matrix with w_r (M_r) as row vectors

- Scaling problem

  This QP problem might be much harder to solve. Platt proposed Sequential Minimal Optimization (SMO) to speed up the training.

Problem Setting

- Multi-class Sequential Supervised Learning
  - Training example: (X,Y) where
    - X = (x_1, …, x_T) is a sequence of feature vectors
    - Y = (y_1, …, y_T) is a matching sequence of class labels
  - Goal: Given new X, predict new Y
- We work on OCR data, e.g.

Problem Setting (2)

- The task is to learn a function $h : \mathcal{X} \mapsto \mathcal{Y}$ from a training set $S = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)} = \mathbf{t}(\mathbf{x}^{(i)}))\}_{i=1}^{m}$ where $\mathcal{Y} = \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_l$ with $\mathcal{Y}_i = \{y_1, \ldots, y_k\}$. Given *n* basis function $f_j : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$, $h_{\mathbf{w}}$ is defined as:

$$h_{\mathbf{w}}(\mathbf{x}) = \arg\max_{\mathbf{y}} \sum_{i=1}^{n} w_j f_j(\mathbf{x}, \mathbf{y}) = \arg\max_{\mathbf{y}} \mathbf{w}^{\top} \mathbf{f}(\mathbf{x}, \mathbf{y})$$

- Note that # of assignments to **y** is exponential (k^l). Both representing f_j and solving the above *argmax* are infeasible

Graphic Model

- Pairwise Markov network
  - Defined as a graph G = (Y, E); each edge (i,j) associated with a potential $\Psi_{ij}(x,yi,yj)$.
  - Encodes a joint cpd $P(\mathbf{y} \mid \mathbf{x}) \propto \prod_{(i,j) \in E} \psi_{ij}(\mathbf{x}, y_i, y_j)$
  - Captures interactions between Y's compactly
  - Given cpd, intuitively we want to take

    $\text{argmax}_{\mathbf{y}}\ P(\mathbf{y} \mid \mathbf{x})$

    as our prediction.

Unifying Markov Network and SVM

- Markov network distribution is a log-linear model
- Potential $\Psi_{ij}(x,y_i,y_j)$ can be represented (in log-space) a sum of basis functions over x, y_i and y_j.

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

- If we define $f_k(\mathbf{x}, \mathbf{y}) = \sum_{(i,j)\in E} f_k(\mathbf{x}, y_i, y_j)$

  We end up with

  $\text{argmax}_{\mathbf{y}} P(\mathbf{y} \mid \mathbf{x}) = \text{argmax}_{\mathbf{y}} \mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y})$

Formulating SVM

- Single-label Multi-class SVM

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

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

- This is essentially the same as constraining the margin to be a constant and minimize ||**w**||

Formulating SVM (2)

- γ-multi-label margin: $\gamma \Delta \mathbf{t}_{\mathbf{x}}(\mathbf{y})$
  where $\Delta \mathbf{t}_{\mathbf{x}}(\mathbf{y}) = \sum_{i=1}^{l} I(y_i \neq (\mathbf{t}(\mathbf{x}))_i)$
- Multi-label SVM

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

- The result of using # of individual labeling errors as loss function.
- The QP form:

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

Formulating SVM (3)

- Final form (w/ slack variables)

$$\begin{aligned} \min \quad & \frac{\mathbf{w}^\top \mathbf{w}}{2} + C \sum_{\mathbf{x}} \xi_{\mathbf{x}} \, ; \\ \text{s.t.} \quad & \mathbf{w}^\top \Delta \mathbf{f}_{\mathbf{x}}(\mathbf{y}) \geq \Delta \mathbf{t}_{\mathbf{x}}(\mathbf{y}) - \xi_{\mathbf{x}}, \ \forall \mathbf{x}, \mathbf{y}, \\ & \xi_{\mathbf{x}} \geq 0, \ \forall \mathbf{x}. \end{aligned} \tag{5}$$

- Its dual formulation

$$\begin{aligned} \max \quad & \sum_{\mathbf{x},\mathbf{y}} \alpha_{\mathbf{x}}(\mathbf{y}) \Delta \mathbf{t}_{\mathbf{x}}(\mathbf{y}) \\ & -\frac{1}{2} \sum_{\mathbf{x},\mathbf{y}} \sum_{\hat{\mathbf{x}},\hat{\mathbf{y}}} \alpha_{\mathbf{x}}(\mathbf{y}) \alpha_{\hat{\mathbf{x}}}(\hat{\mathbf{y}}) \Delta \mathbf{f}_{\mathbf{x}}(\mathbf{y})^\top \Delta \mathbf{f}_{\hat{\mathbf{x}}}(\hat{\mathbf{y}}) \, ; \\ \text{s.t.} \quad & \sum_{\mathbf{y}} \alpha_{\mathbf{x}}(\mathbf{y}) = C, \forall \mathbf{x}; \ \ \alpha_{\mathbf{x}}(\mathbf{y}) \geq 0 \, , \forall \mathbf{x}, \mathbf{y}. \end{aligned} \tag{6}$$

SMO learning of M^3 Networks

- SMO is an efficient algorithm solving QP problems, it has three components
  - An analytic method to solve two Lagrangian multipliers subproblems
  - A heuristic for choosing which multipliers to optimize
  - A method for computing *b*
- We explore the structure of the dual form and propose how to do SMO learning on M^3 networks

Generalization Error Bound

- A theoretical analysis to relate training error to testing (generalization) error.
- Average per label loss

$$\mathcal{L}(\mathbf{w}, \mathbf{x}) = \frac{1}{l}\Delta\mathbf{t}_{\mathbf{x}}(\arg\max_{\mathbf{y}} \mathbf{w}^{\top}\mathbf{f}_{\mathbf{x}}(\mathbf{y}))$$

- γ-margin per-label loss

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

- Theorem 6.1 …there exists a constant *K*, the following holds with probability 1- δ

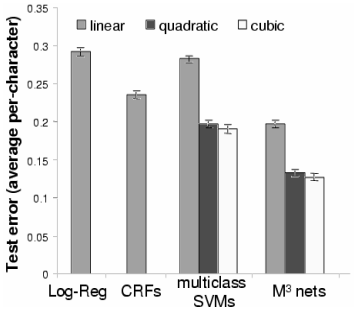
$$E_{\mathbf{x}}\mathcal{L}(\mathbf{w}, \mathbf{x}) \le E_S\mathcal{L}^{\gamma}(\mathbf{w}, \mathbf{x}) + \sqrt{\frac{K}{m}\left[\frac{R_{edge}^2 \|\mathbf{w}\|_2^2 q^2}{\gamma^2}[\ln m + \ln l + \ln q + \ln k] + \ln\frac{1}{\delta}\right]}$$

Experiments

- We select a subset of ~6100 handwritten words, with average length of ~8 characters, from 150 human subjects
- Each word is divided into characters, rasterized into 16x8 images
- 26-class problem: {a..z}

Experiments (2)

- Result



  - LR – independent-labeling; trained on conditional likelihood
  - CRF – sequential-labeling; links between yi and yi+1
  - SVMs – linear, quadratic and cubic kernels
  - Multi-class SVM – independent-labeling