Maximum Margin Markov Networks

Krishnan Narayanan

27 March 2008
This talk is based on the paper “Max-Margin Markov Networks” by B. Taskar, C. Guestrin and D. Koller [TGK03], NIPS 2003
Many Real-world tasks involve Sequential, Spatial, Structured data.

- Eg. Hand-written character recognition: Image $\rightarrow$ Word
- NLP: Sentence $\rightarrow$ Parse Tree
- Bond prediction in Proteins: Amino acid Sequence $\rightarrow$ Bond Structure
- Terrain Segmentation: 3D Image $\rightarrow$ Segmented Objects

Common Characteristics:

- Correlated Labels, Multi-label, Multi-class classification
- Inference here is Global rather than Local
Structured Classification

- **Classification**: Find a function that assigns a label to an arbitrary object
- **Supervised Classification**: Given a sequence of labelled examples independently chosen from an arbitrary distribution, find a function that will assign labels to unseen objects
- **Structured Classification**: To jointly classify different objects in the supervised setting
Structured Classification and SVM

- SVM: Very effective classifier for a variety of applications
- SVM = Kernel + Generalization Bounds (Max-Margin)
- Kernel: Reduce arbitrary nonlinear classification in the input space to linear classification in the feature space.
- Generalization Bounds: Justification for Max-Margin
- SVM assign a single label to an object at a time, do not exploit correlation between labels.
- Running time of SVM: Polynomial in \# classes.
- To jointly classify objects with a joint label, an exponential number of classes required, so infeasible
Markov Networks (MN) and Structured Classification

- Can express correlation between labels
- Can exploit problem structure
- Cannot handle high-dimensional feature spaces
- No strong generalization bounds
Maximum-Margin Markov Networks (M³N)

- Combines the Kernel and Max-Margin concepts of SVM with the ability of MN to handle structured data
- For structured classification, $M^3N = SVM + MN$

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>SVM</th>
<th>MN</th>
<th>M³N</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-dimensional Feature Space (Kernel)</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Generalization Guarantees</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Ability to deal with Structured Objects</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
Structured Classification - Framework

- Task: Learn a function \( f : \mathcal{X} \longrightarrow \mathcal{Y} \)
- \( S = \{(x^{(i)}, y^{(i)} = t(x^{(i)}))\}_{i=1}^{m} \sim D_{\mathcal{X} \times \mathcal{Y}}^{m} \)
- \( \mathcal{H} \): A parameter family
- Classification function \( h \in \mathcal{H} \)
- Common choice: \( \mathcal{H} \) - linear family
- Given \( n \) basis functions \( \{f_j : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}\}_{j=1}^{n} \)
- A hypothesis \( h_w \in \mathcal{H} \) is defined by a set of \( n \) coefficients \( w_j \in \mathbb{R} \)
  - \( h_w(x) = \arg \max_y \sum_{i=1}^{n} w_j f_j(x, y) = \arg \max_y w^\top f(x, y) \)
  where \( f(x, y) \) are features (=basic functions)
Structured Classification - Framework (contd)

- Single-label case
  - $\mathcal{Y} = \{y_1, y_2, \cdots, y_l\}$
- Our focus: Multi-Label case
  - $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_k$
  - $\mathcal{Y}_i = \{y_1, y_2, \cdots, y_l\}$
- Eg: OCR where $\mathcal{Y}_i$: A character, $\mathcal{Y}$: A full word
- $|\mathcal{Y}| = l^k \sim \exp(k)$
- Infeasible to
  - Represent basis functions $f_j(x, y)$
  - Compute $\arg \max_y$
Examples

- Hidden Markov Model (HMM)
- Conditional Random Field (CRF)
- Markov Random Field (MRF) (aka Markov Network (MN))

Here the model defines (directly or indirectly) a conditional distribution $P(\mathcal{Y} | \mathcal{X})$

Goal: Select the label $\text{argmax}_y P(y | x)$

Advantage: Possible to exploit sparse label correlations

Eg. OCR task using Markov Network

$\mathcal{Y}_i \perp \perp \mathcal{Y}_j | \mathcal{Y}_{i-1}, \mathcal{Y}_{i+1}, j \neq i - 1, i + 1$
Encoding a Probability Structure in Markov Network

- Assumption: Pairwise interaction between labels
- MN: $G = (\mathcal{V}, \mathcal{E})$
  - edge $(i, j) \mapsto$ Potential $\psi_{ij}(x, y_i, y_j)$
- $P(y | x)$: Joint Conditional Probability distribution encoded by the network
  
  
  $P(y | x) \propto \prod_{(i, j) \in E} \psi_{ij}(x, y_i, y_j)$

- MN: Compact Parametrization of a classifier
- $G =$ Tree-structured network:
  - $\arg \max_y P(y | x) =$ Viterbi Algorithm
  - Efficient, even if there are an exponential number of labels
  - This is a great advantage of graphical models over SVM
- In general, Approximate Inference algorithm that exploit structure
Markov Network Distribution - Log-Linear (LL) Model

- Belongs to the family of Generalized Linear Models
- $\psi_{ij}(x, y_i, y_j)$: Network potential
- $f_k(x, y_i, y_j)$: Basis functions, $k = 1, \cdots, n$
- MN can be parameterized by the Basis functions
- Assumption: All the edges in the graph denote the same type of interaction
- $f_k(x, y) = \sum_{(i, j) \in E} f_k(x, y_i, y_j)$ - features $k = 1, \cdots, n$
- $\log \psi_{ij}(x, y_i, y_j) = \sum_{k=i}^n w_k f_k(x, y_i, y_j)$
- $\psi_{ij}(x, y_i, y_j) = \exp \left[ \sum_{k=i}^n w_k f_k(x, y_i, y_j) \right] = \exp \left[ w^\top f(x, y_i, y_j) \right]$
- $w$ in LL model can be trained by Maximum Likelihood (ML) or Conditional Likelihood
- Alternative approach: To select $w$ by maximizing the margin is the approach in $M^3N$
Loss function and Risk

- Statistical Learning Theory provides the justification for maximal margin criterion
- Maximum-margin minimizes the generalization error bound
- Loss function \( L : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R} \)
- \( L(\mathbf{x}, \mathbf{y}, h(\mathbf{x})) \): Loss in assigning \( h(\mathbf{x}) \) to \( \mathbf{x} \) when the true label is \( \mathbf{y} \)
- \( L(\mathbf{x}, \mathbf{y}, \mathbf{y}) = 0 \)
- Goal: Minimize the total loss on the labels to predicted
- \( R[h] \): Expected risk in choosing classifier \( h \)
- \( R[h] := \int_{\mathcal{X} \times \mathcal{Y}} L(\mathbf{x}, \mathbf{y}, h(\mathbf{x})) \, d \mathcal{D}(\mathbf{x}, \mathbf{y}) \)
- \( R[h] \) cannot be computed but can be approximated by the empirical risk \( R_{emp}[h] \) [Vapnik]
- \( R_{emp}[h] := \frac{1}{n} \sum_{i=1}^{n} L(\mathbf{x}^{(i)}, \mathbf{t}^{(i)}, h(\mathbf{x}^{(i)})) \)
Generalization Error Bound

- Statistical Learning Theory informs the tradeoff between the choice of the model class (expressibility) and training error.
- Gives a theoretical bound on the generalization error of a classifier which is independent of the distribution $\mathcal{D}$ (SVM)
- Generalization error independent of the dimension of the feature space
- Freedom from the ‘curse of dimensionality’ (SVM)
- SVM learning rooted in Statistical Learning Theory
Margin-based Structured Classification

- SVM: Single label 2-classification
- (Single-label) m-classification extends 2-classification
- $\gamma$: margin
- $\max \gamma \text{ s.t. } \|w\| \leq 1; \ w^\top \Delta f_x(y) \geq \gamma, \ \forall x \in S, \ \forall y \neq t(x)$
  - where $\Delta f_x(y) := f(x, t(x)) - f(x, y)$
- $\arg \max_y w^\top f(x, y) = t(x)$ is a consequence of the above constraint
Loss in Structured Problems

- Not a simple 0-1 loss
- 0-1 loss: \( I(\arg\max_y w^T f(x, y) = t(x)) \)
- Loss is a per-label loss aka Proportion of incorrect labels predicted
- Margin between \( t(x) \) and \( y \) scales linearly with the number of wrong labels in \( y \), \( \Delta t_x(y) \):
  - \( \max \gamma s.t. \|w\| \leq 1, w^T \Delta f_x(y) \geq \gamma \Delta t_x(y) \), \( \forall x \in S, \forall y \)
  - \( \Delta t_x(y) := \sum_{i=1}^{l} \Delta t_x(y_i) \)
  - \( \Delta t_x(y_i) := I(y_i \neq (t(x)_i)) \)
- We tidy up the above by eliminating \( \gamma \) to get the Quadratic Program (QP)
The QP Problem for Margin-based Structured Classification and its Dual

- QP: \( \min \frac{1}{2} \| w \|^2 \) s.t. \( w^\top \Delta f_x(y) \geq \Delta t_x(y), \forall x \in S, \forall y \)

Introducing slack variables \( \xi_x \) to allow linearly inseparable data, we get the Primal (P) and Dual (D)

- (P): \( \min \frac{1}{2} \| w \|^2 + C \sum_x \xi_x \)
  
  s.t. \( w^\top \Delta f_x(y) \geq \Delta t_x(y) - \xi_x, \forall x, \forall y \).

- (D): \( \max \sum_{x,y} \alpha_x(y) \Delta t_x(y) - \frac{1}{2} \left\| \sum_{x,y} \alpha_x(y) \Delta f_x(y) \right\|^2 \)

  s.t. \( \sum_y \alpha_x(y) = C, \forall x; \alpha_x(y) \geq 0, \forall x, y \).
Difficulty in Solving Primal and Dual

- (P): \( \min \frac{1}{2} \| w \|^2 + C \sum_x \xi_x \);
  \( s.t. \ w^T \Delta f_x(y) \geq \Delta t_x(y) - \xi_x, \forall x, \forall y. \)

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

- # of constraints in (P) and # variables in (D) are both exponential in # labels
- Infeasible computation
- Can we get around it?

TGK03 give an affirmative answer in this paper!
Key idea in the [TGK03] solution

- The variables $\alpha_x(y)$ in (D) can be interpreted as an unnormalized density function over $y$ conditional on $x$:
  - $\sum_y \alpha_x(y) = C$, $\alpha_x(y) \geq 0$

- The dual objective is a function of
  - $\mathbb{E}[\Delta t_x(y)]$ and $\mathbb{E}[\Delta f_x(y)]$, where $\mathbb{E}$ expectation w.r.t $\alpha_x(y)$

- $\Delta t_x(y) := \sum_i \Delta t_x(y_i)$
- $\Delta f_x(y) := \sum_{i,j} \Delta f_x(x_i, y_j)$

  are sums of functions over nodes and edges

- So only node and edge marginals of the measure $\alpha_x(y)$ needed to compute the above expectations

- Here the sparse correlations in the feature representations is used.
Marginal Dual Variables (MDV)

- MDV $\mu_x(y_i, y_j)$ and $\mu_x(y_i)$ are defined here
- $\mu_x(y_i, y_j) := \sum_{y \sim [y_i, y_j]} \alpha_{\mathbf{x} \mathbf{y}}, \forall (i, j) \in E, \forall y_i, y_j, \forall \mathbf{x}$;
- $\mu_x(y_i) := \sum_{y \sim [y_i]} \alpha_{\mathbf{x} \mathbf{y}}, \forall i, \forall y_i, \forall \mathbf{x}$;
- $\mathbf{y} \sim [y_i, y_j]$ denote a full assignment $\mathbf{y}$ consistent with partial assignments $y_i, y_j$
- We now reformulate the QP (D) via MDV
Reformulation of QP(D) via MDV

- The first term of the objective function in QP(D) can be written in terms of MDV as follows

\[ \sum_x \alpha_x(y) \Delta t_x(y) = \sum_y \sum_i \alpha_x(y) \Delta t_x(y_i) \]

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

- Similarly, the second term of the objective function in QP(D) via the edge marginals \( \mu_x(y_i, y_j) \)
Consistency Conditions to produce Equivalent QP

- To ensure that the MDV \( \mu_x(y_i, y_j), \mu_x(y_i) \) are marginals arising from a legal density \( \alpha(y) \):

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

- Now we can formulate the equivalent QP in MDV.
Factored Dual QP Equivalent to Original QP(D)

$$\begin{align*}
\max & \sum_{x \in K} \sum_{i,j} \mu_x(y_i) \Delta t_x(y_i) - \frac{1}{2} \sum_{x,y} \sum_{i,j} \sum_{r,s} \mu_x(y_i,y_j) \mu_y(y_r,y_s) f_x(y_i,y_j) \top f_y(y_r,y_s); \\
\text{s.t} & \quad \mu_x(y_i,y_j) = \mu_x(y_j); \, \mu_x(y_i) = C; \, \mu_x(y_i,y_j) \geq 0.
\end{align*}$$

- The objective function here depends only on a polynomial number of MDV Kernels can be used as the basis functions enter as dot products
- The solution of the Factored Dual is
  $$w = \sum_{x} \sum_{i,j} \sum_{y_i,y_j} \mu_x(y_i,y_j) \Delta f_x(y_i,y_j)$$
Conclusion

For Structured Data Classification, $M^3N = SVM + MN$
References

- Max-Margin Markov Networks. B. Taskar, C. Guestrin and D. Koller, NIPS 2003
- Pattern Classification. O. Duda, P.E. Hart and D.C. Stork
- Statistical Learning Theory. V. Vapnik