# 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

# Structured Data

- Many Real-world tasks involve Sequential, Spatial, Structured data.
  - Eg. Hand-written character recognition: Image $\longrightarrow$ Word
  - NLP: Sentence $\longrightarrow$ Parse Tree
  - Bond prediction in Proteins: Amino acid Sequence $\longrightarrow$ Bond Structure
  - Terrain Segmentation: 3D Image $\longrightarrow$ 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 ($\mathrm{M^3N}$)

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

| **Characteristics** | **SVM** | **MN** | $\mathrm{M^3N}$ |
|---|---|---|---|
| High-dimensional Feature Space (Kernel) | $+$ | - | $+$ |
| Generalization Guarantees | $+$ | - | $+$ |
| Ability to deal with Structured Objects | - | $+$ | $+$ |

# Structured Classification - Framework

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

# Probabilistic Graphical Model for Structured Classification

- 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 $argmax_y P(\boldsymbol{y} \,|\, \boldsymbol{x})$
- 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: $\mathcal{G} = (\mathcal{Y}, \mathcal{E})$
  - edge $(i, j) \mapsto$ Potential $\psi_{ij}(x, y_i, y_j)$
- $P(\boldsymbol{y} \,|\, \boldsymbol{x})$: Joint Conditional Probability distribution encoded by the network
- $P(\boldsymbol{y} \,|\, \boldsymbol{x}) \propto \prod_{(i,j) \in E} \psi_{ij}(x, y_i, y_j)$
- MN: Compact Parametrization of a classifier
- $\mathcal{G} =$ Tree-structured network:
  - $\arg\max_{\mathrm{y}} P(\boldsymbol{y} \,|\, \boldsymbol{x}) =$ 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
- $f_k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{(i,j) \in E} f_k(x, y_i, y_j)$ - 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 $\mathrm{M^3N}$

# Loss function and Risk

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

# 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 $\gamma\, s.t.\ \|\boldsymbol{w}\| \leq 1;\ \boldsymbol{w}^\top \Delta \boldsymbol{f}_{\boldsymbol{x}}(\boldsymbol{y}) \geq \gamma,\ \forall \boldsymbol{x} \in S,\ \forall \boldsymbol{y} \neq \boldsymbol{t}(\boldsymbol{x})$
  - where $\Delta \boldsymbol{f}_{\boldsymbol{x}}(\boldsymbol{y}) := \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{t}(\boldsymbol{x})) - \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y})$
- $\arg\max_{\mathrm{y}} \boldsymbol{w}^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{t}(\boldsymbol{x})$ is a consequence of the above constraint

# Loss in Structured Problems

- Not a simple 0-1 loss
- 0-1 loss: $\mathrm{I}(\arg\max_{\mathrm{y}} \boldsymbol{w}^\top \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{t}(\boldsymbol{x}))$
- Loss is a per-label loss aka Proportion of incorrect labels predicted
- Margin between $\boldsymbol{t}(\boldsymbol{x}) and \boldsymbol{y}$ scales linearly with the number of wrong labels in $\boldsymbol{y}$, $\Delta \boldsymbol{t}_{\boldsymbol{x}}(\boldsymbol{y})$:
  - max $\gamma\, s.t\, \|\boldsymbol{w}\| \leq 1$, $\boldsymbol{w}^\top \Delta \boldsymbol{f}_{\boldsymbol{x}}(\boldsymbol{y}) \geq \gamma \Delta \boldsymbol{t}_{\boldsymbol{x}}(\boldsymbol{y})$, $\forall \boldsymbol{x} \in S$, $\forall \boldsymbol{y}$
  - $\Delta \boldsymbol{t}_{\boldsymbol{x}}(\boldsymbol{y}) := \sum_{i=1}^{l} \Delta \boldsymbol{t}_{\boldsymbol{x}}(y_i)$
  - $\Delta \boldsymbol{t}_{\boldsymbol{x}}(y_i) := \mathrm{I}(y_i \neq (\boldsymbol{t}(\boldsymbol{x})_i))$
- 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: $min\ \frac{1}{2}\|w\|^2\ s.t.\ \boldsymbol{w}^\top \Delta\boldsymbol{f_x}(\boldsymbol{y}) \geq \Delta\boldsymbol{t_x}(\boldsymbol{y}),\ \forall\boldsymbol{x} \in S,\ \forall\boldsymbol{y}$
- Introducing slack variables $\xi_{\boldsymbol{x}}$ to allow linearly inseparable data, we get the Primal (P) and Dual (D)
- (P): $min\ \ \frac{1}{2}\left\|w\right\|^2 + C\ \sum_{\boldsymbol{x}}\xi_{\boldsymbol{x}};$

  $s.t.\ \boldsymbol{w}^\top \Delta\boldsymbol{f_x}(\boldsymbol{y}) \geq \Delta\boldsymbol{t_x}(\boldsymbol{y}) - \xi_{\boldsymbol{x}},\ \forall\boldsymbol{x},\ \forall\boldsymbol{y}.$
- (D): $max\ \ \sum_{\boldsymbol{xy}}\alpha_{\boldsymbol{x}}(\boldsymbol{y})\Delta\boldsymbol{t_x}(\boldsymbol{y}) - \frac{1}{2}\left\|\sum_{\boldsymbol{x},\boldsymbol{y}}\alpha_{\boldsymbol{x}}(\boldsymbol{y})\Delta\boldsymbol{f_x}(\boldsymbol{y})\right\|^2;$

  $s.t.\ \sum_{\boldsymbol{y}}\alpha_{\boldsymbol{x}}(\boldsymbol{y}) = C,\ \forall\boldsymbol{x};\ \alpha_{\boldsymbol{x}}(\boldsymbol{y}) \geq 0,\ \forall\boldsymbol{x},\boldsymbol{y}.$

# Difficulty in Solving Primal and Dual

- (P): $min \ \frac{1}{2}\left\|w\right\|^2 + C \sum_{\boldsymbol{x}} \xi_{\boldsymbol{x}};$
  $s.t.\ \boldsymbol{w}^\top \Delta \boldsymbol{f}_{\boldsymbol{x}}(\boldsymbol{y}) \geq \Delta \boldsymbol{t}_{\boldsymbol{x}}(\boldsymbol{y}) - \xi_{\boldsymbol{x}},\ \forall \boldsymbol{x},\ \forall \boldsymbol{y}.$
- (D): $max \ \sum_{\boldsymbol{x}\boldsymbol{y}} \alpha_{\boldsymbol{x}}(\boldsymbol{y}) \Delta \boldsymbol{t}_{\boldsymbol{x}}(\boldsymbol{y}) - \frac{1}{2}\left\|\sum_{\boldsymbol{x},\boldsymbol{y}} \alpha_{\boldsymbol{x}}(\boldsymbol{y}) \Delta \boldsymbol{f}_{\boldsymbol{x}}(\boldsymbol{y})\right\|^2;$
  $s.t.\ \sum_{\boldsymbol{y}} \alpha_{\boldsymbol{x}}(\boldsymbol{y}) = C,\ \forall \boldsymbol{x};\ \alpha_{\boldsymbol{x}}(\boldsymbol{y}) \geq 0,\ \forall \boldsymbol{x}, \boldsymbol{y}.$
- # 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 $\alpha_{\boldsymbol{x}}(y)$ in (D) can be interpreted as an unnormalized density function over $y$ conditional on $x$:
  - $\sum_{y} \alpha_{\boldsymbol{x}}(y) = C, \ \alpha_{\boldsymbol{x}}(y) \geq 0$
- The dual objective is a function of
  - $\mathbb{E}\big[\Delta \boldsymbol{t}_{\boldsymbol{x}}(\boldsymbol{y})\big]$ and $\mathbb{E}\big[\Delta \boldsymbol{f}_{\boldsymbol{x}}(y)\big]$, where $\mathbb{E}$ expectation w.r.t $\alpha_{\boldsymbol{x}}(y)$
- $\Delta \boldsymbol{t}_{\boldsymbol{x}}(\boldsymbol{y}) := \sum_{i} \Delta \boldsymbol{t}_{\boldsymbol{x}}(y_i)$

  $\Delta \boldsymbol{f}_{\boldsymbol{x}}(\boldsymbol{y}) := \sum_{i,j} \Delta \boldsymbol{f}_{\boldsymbol{x}}(x_i, y_j)$

  are sums of functions over nodes and edges
- So only node and edge marginals of the measure $\alpha_{\boldsymbol{x}}(y)$ needed to compute the above expectations
- Here the sparse correlations in the feature representations is used.

# Marginal Dual Variables (MDV)

- MDV $\mu_{\boldsymbol{x}}(y_i, y_j)$ and $\mu_{\boldsymbol{x}}(y_i)$ are defined here
- $\mu_{\boldsymbol{x}}(y_i, y_j) := \sum_{\boldsymbol{y} \sim [y_i, y_j]} \alpha_{\boldsymbol{x}\boldsymbol{y}}, \ \forall (i,j) \in E, \forall y_i, y_j, \ \forall \boldsymbol{x};$
- $\mu_{\boldsymbol{x}}(y_i) \quad := \sum_{\boldsymbol{y} \sim [y_i]} \alpha_{\boldsymbol{x}\boldsymbol{y}}, \quad \forall i, \forall y_i, \ \forall \boldsymbol{x};$
- $\boldsymbol{y} \sim [y_i, y_j]$ denote a full assignment $\boldsymbol{y}$ 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
- $\sum_{\boldsymbol{x}} \alpha_{\boldsymbol{x}}(\boldsymbol{y}) \Delta \boldsymbol{t}_{\boldsymbol{x}}(\boldsymbol{y}) = \sum_{\boldsymbol{y}} \sum_{\boldsymbol{i}} \alpha_{\boldsymbol{x}}(\boldsymbol{y}) \Delta \boldsymbol{t}_{\boldsymbol{x}}(y_i)$
  $= \sum_{i, y_i} \Delta \boldsymbol{t}_{\boldsymbol{x}}(y_i) \sum_{\boldsymbol{y} \sim [y_i]} \alpha_{\boldsymbol{x}}(\boldsymbol{y}) = \sum_{i, y_i} \mu_{\boldsymbol{x}}(y_i) \Delta \boldsymbol{t}_{\boldsymbol{x}}(y_i)$
- Similarly the second term of the objective function in QP(D) via the edge marginals $\mu_{\boldsymbol{x}}(y_i, y_j)$

# Consistency Conditions to produce Equivalent QP

- To ensure that the MDV $\mu_{\boldsymbol{x}}(y_i, y_j)$, $\mu_{\boldsymbol{x}}(y_i)$ are marginals arising from a legal density $\alpha(y)$:
  - $\sum_{y_i} \mu_{\boldsymbol{x}}(y_i, y_j) = \mu_{\boldsymbol{x}}(y_j), \forall y_j, \ \forall (i,j) \in E, \forall \boldsymbol{x}$
- Now we can formulate the equivalent QP in MDV.

# Factored Dual QP Equivalent to Original QP(D)

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

$$\text{s.t} \quad \mu_{\boldsymbol{x}}(y_i, y_j) = \mu_{\boldsymbol{x}}(y_j); \;\; \mu_{\boldsymbol{x}}(y_i) = C; \;\; \mu_{\boldsymbol{x}}(y_i, y_j) \geq 0.$$

- 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
  - $\boldsymbol{w} = \sum_{\boldsymbol{x}} \sum_{i,j} \sum_{y_i,y_j} \mu_{\boldsymbol{x}}(y_i, y_j) \, \Delta \boldsymbol{f}_{\boldsymbol{x}}(y_i, y_j)$

# Conclusion

For Structured Data Classification, $\mathrm{M}^3\mathrm{N} = \mathrm{SVM} + \mathrm{MN}$

# References


- Max-Margin Markov Networks. B. Taskar, C. Guestrin and D. Koller, NIPS 2003
- Combining SVM with Graphical Models for Supervised Classification: an Introduction to Max-Margin Markov Networks. Simon Lacoste-Julien, Report December, 2003
- Information Theory, Inference and Learning Algorithms. D.J.C. MacKay (book online)
- Pattern Classification. O. Duda, P.E. Hart and D.C. Stork
- Statistical Learning Theory. V. Vapnik