Discriminative Unsupervised Learning of Structured Predictors

Linli Xu, Dana Wilkinson, Finnegan Southey, Dale Shuurmans

presentation by Thomas Finley for CS 778, November 7, 2006

Semidefinite Programming

- *M* is **symmetric positive semidefinite** (SPSD) ⇔ for any vector *x*, $x^T Mx \geq 0$.
- **Semidefinite Programming**
  - Linear **objective** and **constraints** with *X* optimized over SPD matrices.
  - Seems primitive, but subsumes many classes of optimization problems!
  - Expressible in either primal or dual forms.
- **One Major Trick**
  - Want to find *B*, not necessarily SPD.
  - Rephrase equations in terms of $X=BB^T$. (BB^T must be SPD!)
  - When *X* found, use matrix decomposition to find approximate *B*.

$$\begin{aligned} \max_{X \succeq 0} \quad & \langle C, X \rangle \\ \text{subject to} \quad & \langle D_1, X \rangle = d_1 \\ & \vdots \\ & \langle D_k, X \rangle = d_k \end{aligned}$$

$$\begin{aligned} \min \quad & c^T x \\ \text{subject to} \quad & x_1 A_1 + \cdots \\ & + x_n A_n - B \succeq 0 \end{aligned}$$

Adult Supervision

- **Supervised setting**: Examples **x** accompanied by fully specified label **y**.
- **Semi-supervised setting**: Examples **x** accompanied by perhaps incomplete label **y**.
- **Unsupervised setting**: Examples **x** are unaccompanied.

Unsupervised SVMs

- **Unsupervised**: Assign labels to examples.
- **SVMs**: Find "simplest" hypothesis that still clears some margin between the correct labeling and incorrect labeling(s)!
- **Unsupervised SVMs**: Label examples (under strong constraints) such that the model is as simple as possible.

Getting Unsupervised Structural SVMs

- Unsupervised 2-class SVMs.
  (De Bie & Cristianini, 2003; Xu et al., 2004)
- Unsupervised multiclass SVMs.
  (Xu & Schuurmans, 2005)
- Unsupervised structural SVMs.

Max Margin Clustering

$$\omega(\mathbf{y}) = \min_{\mathbf{w}} \frac{\beta}{2}\|\mathbf{w}\|^2 + \sum_i [1 - y_i \boldsymbol{\phi}(\mathbf{x}_i)^\top \mathbf{w}]_+$$

$$= \max_{0 \le \boldsymbol{\lambda} \le 1} \boldsymbol{\lambda}^\top \mathbf{e} - \frac{1}{2\beta} \langle K \circ \boldsymbol{\lambda}\boldsymbol{\lambda}^\top, \mathbf{y}\mathbf{y}^\top \rangle$$

- **Recall**: SVM **primal** and **dual**
- **Supervised case**: **y** known, find dual **λ** to maximize margin
- **Unsupervised case**: **y** unknown, find **y** and **λ** simultaneously (w/**balance constraints**)

$$\mathbf{y} \in \{-1, 1\}^n \qquad M = \mathbf{y}\mathbf{y}^T$$

M_{ij}=1 if i and j in same class

M is binary equivalence relation if and only if $M \succcurlyeq 0$

Relax M from $\{-1,1\}^{n \times n}$ to $[-1,1]^{n \times n}$

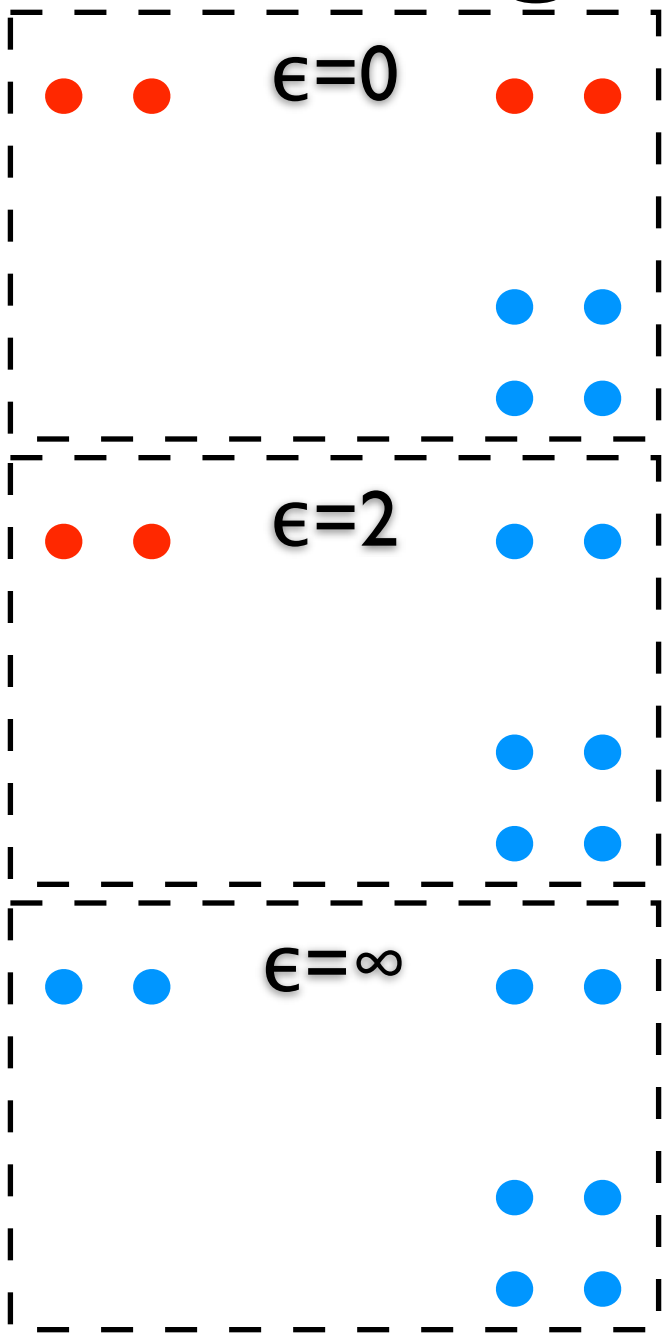
$$\omega(M) = \max_{0 \le \boldsymbol{\lambda} \le 1} \boldsymbol{\lambda}^\top \mathbf{e} - \frac{1}{2\beta} \langle K \circ \boldsymbol{\lambda}\boldsymbol{\lambda}^\top, M \rangle$$

$$\min_{M \succeq 0,\ \mathrm{diag}(M) = \mathbf{e}} \omega(M) \qquad \text{subject to} \quad -\epsilon\mathbf{e} \le M\mathbf{e} \le \epsilon\mathbf{e}$$

Max Margin Clustering

- Classify with balanced classes so that were one to subsequently run an SVM on it, margin would be maxed.

- Examples of how max margin clustering assigns elements to **one class** or **the other**.



$$\omega(M) = \max_{0 \le \boldsymbol{\lambda} \le 1} \boldsymbol{\lambda}^\top \mathbf{e} - \frac{1}{2\beta} \langle K \circ \boldsymbol{\lambda}\boldsymbol{\lambda}^\top, M \rangle$$

$$\min_{M \succeq 0,\ \mathrm{diag}(M) = \mathbf{e}} \omega(M) \quad \text{subject to} \quad -\epsilon\mathbf{e} \le M\mathbf{e} \le \epsilon\mathbf{e}$$

Multiclass MM-Clustering

$$\omega(D) = \min_{\mathbf{w}} \left(\frac{\beta}{2} \|\mathbf{w}\|^2 + \sum_i \max_u \left(1 - D_{iu} - \mathbf{w}^T \left(\phi(\mathbf{x}_i, y_i) - \phi(\mathbf{x}_i, u)\right)\right) \right)$$

$$= \max_{\Lambda \geq 0, \Lambda \mathbf{e} = \mathbf{e}} \left(n - \langle D, \Lambda \rangle - \frac{1}{2\beta} \langle K, DD^T \rangle + \frac{1}{\beta} \langle KD, \Lambda \rangle - \frac{1}{2\beta} \langle \Lambda\Lambda^T, K \rangle \right)$$

- With n examples, κ classes.
- Typical multiclass formulation: $\phi(\mathbf{x},\mathbf{y})$ "offsets" $\mathbf{x}$ to select out the portion of $\mathbf{w}$ corresponding to class $\mathbf{y}$.
- $D \in \{0,1\}^{n \times \kappa}$ indicator matrix: example $\mathbf{x}_i$ in class $y_i \in \{1..\kappa\}$ has $D_{i,y_i}=1$, and 0 otherwise.

Unsupervised Version

$$\omega(D)=\min_{\mathbf{w}}\left(\frac{\beta}{2}\|\mathbf{w}\|^2+\sum_i\max_u\left(1-D_{iu}-\mathbf{w}^T\left(\phi(\mathbf{x}_i,y_i)-\phi(\mathbf{x}_i,u)\right)\right)\right)$$

$$=\max_{\Lambda\geq 0,\Lambda\mathbf{e}=\mathbf{e}}\left(n-\langle D,\Lambda\rangle-\frac{1}{2\beta}\langle K,DD^T\rangle+\frac{1}{\beta}\langle KD,\Lambda\rangle-\frac{1}{2\beta}\langle\Lambda\Lambda^T,K\rangle\right)$$

- Establish equiv $M=DD^T$. $M_{ij}=1$ if $y_i=y_j$, else 0.
- Not convex. Relax to $M\succcurlyeq DD^T$, diag(M)=**e**.
- Establish **class balance** constraints.

$$\omega(D,M)=\max_{\Lambda\geq 0,\Lambda\mathbf{e}=\mathbf{e}}\left(n-\langle D,\Lambda\rangle-\frac{1}{2\beta}\langle K,M\rangle+\frac{1}{\beta}\langle KD,\Lambda\rangle-\frac{1}{2\beta}\langle\Lambda\Lambda^T,K\rangle\right)$$

$$\min_{M\succeq 0,\mathrm{diag}(M)=\mathbf{e},D\geq 0}\ \omega(M)\ \text{s.t.}\ M\succeq DD^T,\ \left(\frac{1}{\kappa}-\epsilon\right)n\mathbf{e}\leq M\mathbf{e}\leq\left(\frac{1}{\kappa}+\epsilon\right)n\mathbf{e}$$

Supervised M³N

$$\omega(\mathbf{y}_1,\ldots,\mathbf{y}_n) = \min_{\mathbf{w}} \frac{\beta}{2}\|\mathbf{w}\|^2 + \sum_i \max_{\mathbf{u}_i} \left(\Delta(\mathbf{u}_i,\mathbf{y}_i) - \mathbf{w}^T(\phi(\mathbf{x}_i,\mathbf{y}_i) - \phi(\mathbf{x}_i,\mathbf{u}_i))\right)$$

- **Sequences**: Decomposes nicely into independent pieces across adjacent nodes!

$$\phi(\mathbf{x}_i,\mathbf{y}_i) = \sum_{k=1}^{L} \phi(x_{i,k}, y_{i,k}, y_{i,k-1})$$

- Define some convenient Δϕ notation:

$$\Delta\phi_{ik}(uu') = (\phi(x_{i,k}, y_{i,k}, y_{i,k-1}) - \phi(x_{i,k}, u, u'))$$

- Rephrase equivalent QP with dual vars over "pieces" of the sequence. (Taskar '03)

$$\omega(\mathbf{y}_1,\ldots,\mathbf{y}_n) = \max_{\boldsymbol{\mu},\boldsymbol{\nu}} \sum_{i,k,u} \mu_{ik}(u) 1_{(u\neq y_{ik})} - \frac{1}{2\beta} \sum_{ik,k\ell,uu',vv'} \nu_{ik}(uu')\nu_{j\ell}(vv')\Delta\phi_{ik}(uu')^T\Delta\phi_{j\ell}(vv')$$

$$\text{subject to} \quad \mu_{ik}(u) \geq 0, \nu_{ik}(uu') \geq 0, \sum_{u'} \nu_{ik}(uu') = \mu_{ik}(u), \sum_u \mu_{ik}(u) = 1$$

Rephrase w/ Ind. Matrices

- *i,j* index examples, *k*, ℓ index sequence positions
- *C* indicator matrix for positions & labels, $M=CC^T$ indicates if two positions have the same label. *D* indicator matrix for edges & pairs of labels, $N=DD^T$ indicates if two edges have same pair of labels.

$$M_{ik,j\ell} = 1_{(y_{ik}=y_{j\ell})} \quad N_{ikk-1,j\ell\ell-1} = 1_{(y_{ikk-1}=y_{j\ell\ell-1})}$$
$$C_{ik,u} = 1_{(y_{ik}=u)} \quad D_{ikk-1,uu'} = 1_{(y_{ikk-1}=uu')}$$

- *K* matrix with $K_{ik,j\ell}$ inner product between subfeature vectors that omit transition model features with current state values equal!!
- *E* matrix of all ones, p_1 number of singleton positions in training set.
- $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are matrices with $\boldsymbol{\mu}_{ik,u} = \mu_{ik}(u)$, $\boldsymbol{\nu}_{ik,uu'} = \nu_{ik}(uu')$.

$$\omega(M,N,C,D) = p_1 - \langle \boldsymbol{\mu}, C\rangle + \tfrac{1}{\beta}(\langle \boldsymbol{\mu}, KC\rangle + \langle \boldsymbol{\nu}, ED\rangle) - \tfrac{1}{2\beta}(\langle M, K\rangle + \langle N, E\rangle + \langle \boldsymbol{\mu}\boldsymbol{\mu}^T, K\rangle + \langle \boldsymbol{\nu}\boldsymbol{\nu}^T, E\rangle)$$

$$\text{subject to } \sum_{u'} \nu_{ik,uu'} = \mu_{ik,u} \quad \forall iku,$$
$$\boldsymbol{\mu} \geq 0, \boldsymbol{\nu} \geq 0, \boldsymbol{\nu}\mathbf{e} = \mathbf{e}, M = CC^T, N = DD^T,$$
$$N_{ikk-1,j\ell\ell-1} = M_{ik,j\ell} M_{ik-1,j\ell-1} \quad \forall ijk\ell$$

Unsupervised M^{3}N

$$\omega(M,N,C,D) = p_1 - \langle \boldsymbol{\mu}, C\rangle + \tfrac{1}{\beta}(\langle \boldsymbol{\mu}, KC\rangle + \langle \boldsymbol{\nu}, ED\rangle) - \tfrac{1}{2\beta}(\langle M, K\rangle + \langle N, E\rangle + \langle \boldsymbol{\mu}\boldsymbol{\mu}^T, K\rangle + \langle \boldsymbol{\nu}\boldsymbol{\nu}^T, E\rangle)$$

$$\text{subject to } \sum_{u'} \nu_{ik,uu'} = \mu_{ik,u} \quad \forall iku,$$

$$\boldsymbol{\mu} \geq 0, \boldsymbol{\nu} \geq 0, \boldsymbol{\nu}\mathbf{e} = \mathbf{e}, M = CC^T, N = DD^T,$$

$$N_{ikk-1,j\ell\ell-1} = M_{ik,j\ell} M_{ik-1,j\ell-1} \quad \forall ijk\ell$$

- Impose class balance constraints.

$$\left(\tfrac{1}{\kappa} - \epsilon\right) p_1 \mathbf{e} \leq M\mathbf{e} \leq \left(\tfrac{1}{\kappa} + \epsilon\right) p_1 \mathbf{e} \qquad \left(\tfrac{1}{\kappa^2} - \epsilon\right) p_2 \mathbf{e} \leq N\mathbf{e} \leq \left(\tfrac{1}{\kappa^2} + \epsilon\right) p_2 \mathbf{e}$$

- Relax equality constraints between *M* & *C*, *N* & *D*.

$$M \succeq CC^T, N \succeq DD^T, \text{diag}(M) = \mathbf{e}, \text{diag}(N) = \mathbf{e}$$

- Relax consistency constraints between *M* & *N*.

$$N_{ikk-1,j\ell\ell-1} \leq M_{ik,j\ell} \qquad N_{ikk-1,j\ell\ell-1} \leq M_{ik-1,j\ell-1}$$

$$N_{ikk-1,j\ell\ell-1} \geq M_{ik,j\ell} + M_{ik-1,j\ell-1} - 1$$

- Relax *M* & *N* from binary {0,1} to real [0,1].

Experimental Comparisons

- Comparisons among three methods
  - **CDHMM**: Convex Discriminative HMM.
  - **ACDHMM**: Alternating CDHMM, which iteratively predicts labels, retrains model, until no labels change. (Local search.)
  - **EMHMM**: Conventional Baum-Welch EM training.

Synthetic & Small Datasets

Data set	CDHMM	ACDHMM	EM
synth1	3.38 ±0.75	14.46 ±1.78	15.09 ±1.92
synth2	8.12 ±1.57	17.34 ±1.52	17.49 ±1.81
synth3	22.12 ±1.40	26.56 ±1.06	30.06 ±1.24
synth4	31.50 ±1.46	38.58 ±0.96	39.90 ±0.86
prot1	51.75 ±1.80	56.67 ±0.47	58.11 ±0.47
prot2	50.38 ±2.04	53.65 ±0.57	57.23 ±0.39

- Four synthetic datasets.
  - Generated 10 samples of length 8 with 2-state HMM.
  - Percentage given is: (1) chance to stay in current state, and (2) chance of having emission noise (e.g., chance in state 0 of emitting 1 instead of 0).
- Two protein datasets.
  - UCI repository (protein-secondary-structure).
  - Samples subsequences & removed labels.
- Accuracy found through maximum mapping through predicted and possible state labels.

Larger Datasets

Data set	ACDHMM	EM
20×2-seq	43.12 ±2.20	46.27 ±1.51
10×5-seq	44.33 ±2.30	48.67 ±1.51
5×10-seq	46.44 ±2.12	48.67 ±1.82

- Use full sequences from protein secondary structure.
  - 20 samples over 2 seqs.
  - 10 samples over 5 seqs.
  - 5 samples over 10 seqs.
- Each observation x_i a window over 7 adjacent amino acids.
- Comparison of alternating method versus standard EM.
- The global CDHMM method is too slow for the complete sequences.

Conclusions & Future Work

- Convex discriminative for unsupervised training of predictors.
- Two approaches: Exact but expensive (CDHMM), inexact but cheap (ACDHMM).
- Future directions:
  - Decreasing computational time.
  - Extension to semi-supervised setting.