

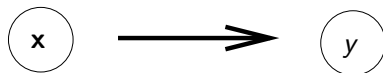
# CRFs in the dual

Roland Memisevic

December 10, 2005

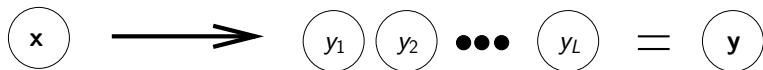
# Linear classification

## Standard



$$\max_y \mathbf{w}^T \phi(\mathbf{x}, y)$$

## Structured



$$\max_y \mathbf{w}^T \phi(\mathbf{x}, \mathbf{y})$$

# Motivation

## Structured predictions

- ▶ 'Margin' based (hinge-loss):
  - ▶ M3N, HMSVM
  - ▶ Optimization typically in the **dual** (SMO [Taskar, 2004], Cutting plane methods [Tsochantaridis et al., 2004])
- ▶ Probabilistic (log-loss):
  - ▶ CRF [Lafferty et al., 2001], [Altun et al., 2004]
  - ▶ Optimization typically in the **primal** (gradient based).

## CRFs in the dual

- ▶ SMO for binary kernel logistic regression [Keerthi et al., 2005]
- ▶ SMO for multinomial kernel logistic regression
- ▶ SMO for CRFs

# Multinomial logistic regression

- ▶ Model:

$$p(y|\mathbf{x}; \mathbf{w}) = \frac{\exp(\mathbf{w}^T \phi(\mathbf{x}; y))}{\sum_y \exp(\mathbf{w}^T \phi(\mathbf{x}; y))}$$

# Multinomial logistic regression

- ▶ Model:

$$p(y|\mathbf{x}; \mathbf{w}) = \frac{\exp(\mathbf{w}^T \phi(\mathbf{x}; y))}{\sum_y \exp(\mathbf{w}^T \phi(\mathbf{x}; y))}$$

- ▶ Use training data  $\{(\mathbf{x}^i, y^i)\}_{i=1, \dots, N}$

# Multinomial logistic regression

- ▶ Model:

$$p(y|\mathbf{x}; \mathbf{w}) = \frac{\exp(\mathbf{w}^T \phi(\mathbf{x}; y))}{\sum_y \exp(\mathbf{w}^T \phi(\mathbf{x}; y))}$$

- ▶ Use training data  $\{(\mathbf{x}^i, y^i)\}_{i=1, \dots, N}$
- ▶ Minimize

$$J(\mathbf{w}) = \frac{\lambda}{2} \|\mathbf{w}\|^2 + \sum_i \log \sum_y \exp(\mathbf{w}^T \phi(\mathbf{x}^i; y)) - \mathbf{w}^T \phi(\mathbf{x}^i; y^i)$$

# Multinomial logistic regression

- ▶ Model:

$$p(y|\mathbf{x}; \mathbf{w}) = \frac{\exp(\mathbf{w}^T \phi(\mathbf{x}; y))}{\sum_y \exp(\mathbf{w}^T \phi(\mathbf{x}; y))}$$

- ▶ Use training data  $\{(\mathbf{x}^i, y^i)\}_{i=1, \dots, N}$
- ▶ Minimize

$$J(\mathbf{w}) = \frac{\lambda}{2} \|\mathbf{w}\|^2 + \sum_i \underbrace{\log \sum_y \exp(\mathbf{w}^T \phi(\mathbf{x}^i; y)) - \mathbf{w}^T \phi(\mathbf{x}^i; y^i)}_{f^i(\xi^i)}$$

# Multinomial logistic regression

- ▶ Model:

$$p(y|\mathbf{x}; \mathbf{w}) = \frac{\exp(\mathbf{w}^T \phi(\mathbf{x}; y))}{\sum_y \exp(\mathbf{w}^T \phi(\mathbf{x}; y))}$$

- ▶ Use training data  $\{(\mathbf{x}^i, y^i)\}_{i=1, \dots, N}$
- ▶ Minimize

$$J(\mathbf{w}) = \frac{\lambda}{2} \|\mathbf{w}\|^2 + \sum_i \underbrace{\log \sum_y \exp(\mathbf{w}^T \phi(\mathbf{x}^i; y)) - \mathbf{w}^T \phi(\mathbf{x}^i; y^i)}_{f^i(\xi^i)}$$

$$\xi_y^i := \mathbf{w}^T \phi(\mathbf{x}^i, y)$$

$$f^i(\xi^i) = \log \sum_y \exp(\xi_y^i) - \xi_{y^i}^i$$

# Multinomial logistic regression (primal)

- ▶ Constrained problem:

$$\begin{aligned} \min_{\mathbf{w}, \xi} \quad & \frac{\lambda}{2} \|\mathbf{w}\|^2 + \sum_i f^i(\xi^i) \\ \text{s.t.} \quad & \xi_y^i = \mathbf{w}^T \phi(\mathbf{x}^i; y) \quad \forall i, y \end{aligned}$$

- ▶ Lagrangian:

$$L(\mathbf{w}, \xi) = \frac{\lambda}{2} \|\mathbf{w}\|^2 + \sum_i f(\xi^i) + \sum_{i,y} \eta_y^i [\mathbf{w}^T \phi(\mathbf{x}^i, y) - \xi_y^i]$$

- ▶ (In the following we will replace  $\alpha_y^i := \eta_y^i + \delta_{y,y^i}$ )

# Multinomial logistic regression (dual)

- ▶ We get the dual:

$$\max_{\alpha} \quad -\frac{1}{2\lambda}\alpha^T K\alpha + \frac{1}{\lambda}\alpha^T K\delta + \sum_i H(\alpha^i)$$

$$\text{s.t.} \quad \alpha_y^i > 0 \quad \forall i, y; \quad \sum_y \alpha_y^i = 1 \quad \forall i,$$

with  $H(\cdot)$  the Shannon entropy and  $\delta = (\delta_{y,y^i})_{(i,y)}$ .

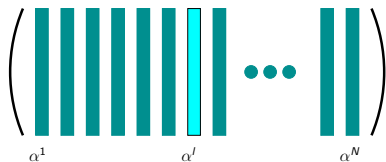
- ▶ We also get the *representation*:

$$\mathbf{w}^T \phi(\mathbf{x}, y) = -\frac{1}{\lambda} \sum_{i,y'} (\alpha_{y'}^i - \delta_{y',y^i}) k((\mathbf{x}, y), (\mathbf{x}^i, y'))$$

# Sequential minimal optimization

First idea: *Coordinate ascent*.

- ▶ Constraints couple only the  $\alpha_y^i$  across different  $y$ .
- ▶ Optimize wrt. a single  $\alpha^i := (\alpha_y^i)_y$ , one at a time:



Second idea: "*Shift weight*".

- ▶ Each  $\alpha^i$  forms a *distribution*.
- ▶ Use the updates:  $\alpha_{y_1}^i \leftarrow \alpha_{y_1}^i + \Delta$  and  $\alpha_{y_2}^i \leftarrow \alpha_{y_2}^i - \Delta$



- ▶ We get the one-dimensional problem:

$$\begin{aligned} \max_{\Delta} \quad & -a\Delta^2 - b\Delta \\ & -(\alpha_{y^1}^i + \Delta) \log(\alpha_{y^1}^i + \Delta) - (\alpha_{y^2}^i - \Delta) \log(\alpha_{y^2}^i - \Delta) \\ \text{s.t.} \quad & -\alpha_{y^1}^i < \Delta < \alpha_{y^2}^i \end{aligned}$$

- ▶ Can solve using Newton-Raphson/bisection.
- ▶ *Stopping and pair selection*: KKT-conditions satisfied when for each  $l$ , the quantity

$$\log(\alpha_y^i) + \frac{1}{\lambda} \sum_{j,y'} (\alpha_{y'}^j - \delta_{y'y^j}) k((\mathbf{x}^i, y)(\mathbf{x}^j, y')) =: g^i(y)$$

is equal across all  $y$ .

# Experiments (USPS)

▶  $\lambda = 0.1$

N	100	500	1000	3000
conj grad	0.54	11.41	49.50	474.17
smo	0.96	4.23	9.44	58.19

▶  $\lambda = 0.5$

conj grad	0.72	17.23	46.50	514.25
smo	1.40	4.01	9.35	56.27

▶  $\lambda = 1.0$

conj grad	0.70	14.82	53.92	555.76
smo	0.85	4.31	15.94	54.23

# Conditional random fields

$$y \rightarrow \mathbf{y}$$

- ▶ Instead of single class labels, consider label-*vectors*.
- ▶ Learning (and representation) generally **intractable**.
- ▶ Trick: Decompose the kernel according to the cliques of a graph:

$$k((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) = \sum_c k((\mathbf{x}, \mathbf{y}_c), (\mathbf{x}', \mathbf{y}'_c)).$$

# Conditional random fields

- ▶ We can express everything in terms of *marginal variables*  $\mu_c^i(\mathbf{y}_c)$ .

- ▶ For example:

$$\begin{aligned} & \mathbf{w}^T \phi(\mathbf{x}, \mathbf{y}) \\ &= -\frac{1}{\lambda} \sum_j \sum_{\mathbf{y}'} (\alpha_{\mathbf{y}'}^j - \delta_{\mathbf{y}^j \mathbf{y}'}) k((\mathbf{x}, \mathbf{y}), (\mathbf{x}^j, \mathbf{y}')) \end{aligned}$$

# Conditional random fields

- ▶ We can express everything in terms of *marginal variables*  $\mu_c^i(\mathbf{y}_c)$ .

- ▶ For example:

$$\begin{aligned} & \mathbf{w}^T \phi(\mathbf{x}, \mathbf{y}) \\ &= -\frac{1}{\lambda} \sum_j \sum_{\mathbf{y}'} (\alpha_{\mathbf{y}'}^j - \delta_{\mathbf{y}^j \mathbf{y}'}) \sum_c k_c((\mathbf{x}, \mathbf{y}_c), (\mathbf{x}^j, \mathbf{y}'_c)) \end{aligned}$$

## Conditional random fields

- ▶ We can express everything in terms of *marginal variables*  $\mu_c^i(\mathbf{y}_c)$ .

- ▶ For example:

$$\begin{aligned} & \mathbf{w}^T \phi(\mathbf{x}, \mathbf{y}) \\ &= -\frac{1}{\lambda} \sum_j \sum_c \sum_{\mathbf{y}'} (\alpha_{\mathbf{y}'}^j - \delta_{\mathbf{y}^j \mathbf{y}'}) k_c((\mathbf{x}, \mathbf{y}_c), (\mathbf{x}^j, \mathbf{y}'_c)) \end{aligned}$$

## Conditional random fields

- ▶ We can express everything in terms of *marginal variables*  $\mu_c^i(\mathbf{y}_c)$ .

- ▶ For example:

$$\begin{aligned} & \mathbf{w}^T \phi(\mathbf{x}, \mathbf{y}) \\ &= -\frac{1}{\lambda} \sum_j \sum_c \sum_{\mathbf{y}'_c} k_c((\mathbf{x}, \mathbf{y}_c), (\mathbf{x}^j, \mathbf{y}'_c)) (\alpha_{\mathbf{y}'_c}^j - \delta_{\mathbf{y}^j \mathbf{y}'_c}) \end{aligned}$$

# Conditional random fields

- ▶ We can express everything in terms of *marginal variables*  $\mu_c^i(\mathbf{y}_c)$ .

- ▶ For example:

$$\begin{aligned} & \mathbf{w}^T \phi(\mathbf{x}, \mathbf{y}) \\ &= -\frac{1}{\lambda} \sum_j \sum_c \sum_{\mathbf{y}_c} k_c((\mathbf{x}, \mathbf{y}_c), (\mathbf{x}^j, \mathbf{y}_c^j)) \sum_{\sim \mathbf{y}_c'} (\alpha_{\mathbf{y}_c'}^j - \delta_{\mathbf{y}_c \mathbf{y}_c'}) \end{aligned}$$

## Conditional random fields

- ▶ We can express everything in terms of *marginal variables*  $\mu_c^i(\mathbf{y}_c)$ .

- ▶ For example:

$$\begin{aligned} & \mathbf{w}^T \phi(\mathbf{x}, \mathbf{y}) \\ &= -\frac{1}{\lambda} \sum_j \sum_c \sum_{\mathbf{y}_c} k_c((\mathbf{x}, \mathbf{y}_c), (\mathbf{x}^j, \mathbf{y}_c^j)) (\sum_{\sim \mathbf{y}_c'} \alpha_{\mathbf{y}_c'}^j - \sum_{\sim \mathbf{y}_c'} \delta_{\mathbf{y}_c^j \mathbf{y}_c'}) \end{aligned}$$

# Conditional random fields

- ▶ We can express everything in terms of *marginal variables*  $\mu_c^i(\mathbf{y}_c)$ .

- ▶ For example:

$$\begin{aligned} & \mathbf{w}^T \phi(\mathbf{x}, \mathbf{y}) \\ &= -\frac{1}{\lambda} \sum_j \sum_c \sum_{\mathbf{y}_c} k_c((\mathbf{x}, \mathbf{y}_c), (\mathbf{x}^j, \mathbf{y}_c^j)) (\mu_c^j(\mathbf{y}_c^j) - \delta_{\mathbf{y}_c^j \mathbf{y}_c}) \end{aligned}$$

## Conditional random fields

- ▶ We can express everything in terms of *marginal variables*  $\mu_c^i(\mathbf{y}_c)$ .
- ▶ For example:  
$$\mathbf{w}^T \phi(\mathbf{x}, \mathbf{y})$$
$$= -\frac{1}{\lambda} \sum_j \sum_c \sum_{\mathbf{y}_c} k_c((\mathbf{x}, \mathbf{y}_c), (\mathbf{x}^j, \mathbf{y}_c^j)) (\mu_c^j(\mathbf{y}_c^j) - \delta_{\mathbf{y}_c^j \mathbf{y}_c'})$$
- ▶ Other expressions similarly.

## Conditional random fields

- ▶ We can express everything in terms of *marginal variables*  $\mu_c^i(\mathbf{y}_c)$ .
- ▶ For example:  
$$\mathbf{w}^T \phi(\mathbf{x}, \mathbf{y}) = -\frac{1}{\lambda} \sum_j \sum_c \sum_{\mathbf{y}_c} k_c((\mathbf{x}, \mathbf{y}_c), (\mathbf{x}^j, \mathbf{y}_c')) (\mu_c^j(\mathbf{y}_c') - \delta_{\mathbf{y}_c^i \mathbf{y}_c'})$$
- ▶ Other expressions similarly.
- ▶ Use belief propagation to maximize and minimize  $g^i(\mathbf{y})$ .

## Conditional random fields

- ▶ We can express everything in terms of *marginal variables*  $\mu_c^i(\mathbf{y}_c)$ .

- ▶ For example:

$$\begin{aligned} & \mathbf{w}^T \phi(\mathbf{x}, \mathbf{y}) \\ &= -\frac{1}{\lambda} \sum_j \sum_c \sum_{\mathbf{y}'_c} k_c((\mathbf{x}, \mathbf{y}_c), (\mathbf{x}^j, \mathbf{y}'_c)) (\mu_c^j(\mathbf{y}'_c) - \delta_{\mathbf{y}_c^i \mathbf{y}'_c}) \end{aligned}$$

- ▶ Other expressions similarly.
- ▶ Use belief propagation to maximize and minimize  $g^i(\mathbf{y})$ .
- ▶ (Toy experiment: Synthetic data (HMM, 3 states, Gaussian outputs). Test Errors: Independent predictions – 0.1081; chain CRF – 0.0384)

# Conclusions

## CRFs in the dual

- ▶ Dual optimization instead of representer theorem.
- ▶ 'Kernel adatron'.

## Future work

- ▶ Numerical issues.
- ▶ (Probabilistic) variational methods.
- ▶ Adapt other methods (cutting planes ?)
- ▶ Evaluation

# References



Altun, Y., Hofmann, T., and Smola, A. J. (2004).

Gaussian process classification for segmenting and annotating sequences.

In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 4, New York, NY, USA. ACM Press.



Keerthi, S. S., Duan, K., Shevade, S. K., and Poo, A. N. (2005).

A fast dual algorithm for kernel logistic regression.

*Machine Learning*.



Lafferty, J., McCallum, A., and Pereira, F. (2001).

Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA.



Platt, J. (1998).

Sequential minimal optimization: A fast algorithm for training support vector machines.



Taskar, B. (2004).

*Learning Structured Prediction Models: A Large Margin Approach*.

PhD thesis, CA.



Tsochantaridis, I., Hofmann, T., Joachims, T., and Altun, Y. (2004).

Support vector machine learning for interdependent and structured output spaces.

In *ICML '04: Twenty-first international conference on Machine learning*, New York, NY, USA. ACM Press.