Note Title 11/25/2006

SVM with MultiClass

Task: learn a classifier $h : \mathcal{X} \to \mathcal{Y}$

Hypothesis: $h_{\underline{\omega}}(\underline{x}) = \arg\max_{\underline{y}} \sum_{i=1}^{n} \omega_j \cdot f_j(\underline{x}, \underline{y}) = \arg\max_{\underline{y}} \underline{\omega}^T \underline{f}(\underline{x}, \underline{y})$

Task: select weights $\{\omega_j\}$ using training data.

$\{(\underline{x}^{(i)}, y^{(i)} = t(\underline{x}^{(i)})) : i = 1 \text{ to } N\}$ $\underline{t}(\underline{x})$ = true label of $\underline{x}$.

Define: $\underline{\Delta f_x}(\underline{y}) = \underline{f}(\underline{x}, \underline{t}(\underline{x})) - \underline{f}(\underline{x}, \underline{y})$

Minimize $\frac{1}{2} \|\underline{\omega}\|^2 + C \sum_{x} \xi_{\underline{x}}$

s.t. $\underline{\omega} \cdot \underline{\Delta f_x}(\underline{y}) \geq \Delta \underline{t}_x(\underline{y}) - \xi_{\underline{x}}$ $\forall \underline{x}, \underline{y}$

Condition $\Rightarrow \xi_{\underline{x}} \geq 0$, by setting $y = t(\underline{x})$, $\underline{\Delta f_x}(t(\underline{x})) = 0 = \Delta \underline{t_x}(t(\underline{x}))$.

$\Delta \underline{t}_x(\underline{y}) = 1$, if $y \neq \underline{t}(\underline{x})$
0, otherwise.

Primal Problem

$$L_p(\underline{\omega}, \underline{\xi_x} : \underline{\alpha_x})$$

$$= \frac{1}{2} \|\underline{\omega}\|^2 + C \sum_{x} \xi_{\underline{x}}$$

$$- \sum_{\underline{x}, \underline{y}} \underline{\alpha}_x(y)\left(\underline{\omega} \cdot \underline{\Delta f_x}(\underline{y}) - \underline{\Delta t_x}(\underline{y}) + \underline{\xi_x}\right)$$

Dual Problem

$$L_d(\underline{\alpha_x}) = \sum_{\underline{x}, \underline{y}} \alpha_{\underline{x}}(\underline{y}) \Delta t_x(\underline{y}) - \frac{1}{2} \left\| \sum_{\underline{x}, \underline{y}} \alpha_{\underline{x}}(\underline{y}) \underline{\Delta f_x}(\underline{y}) \right\|^2$$

s.t. $\sum_{\underline{y}} \alpha_x(\underline{y}) = C$, $\forall \underline{x}$. $\alpha_x(\underline{y}) \geq 0$.

E.G. | a | b | c | d | e | f |

(2) Special Case: Markov Network.

$$\mathcal{Y} = \mathcal{Y}_1 \times \dots \times \mathcal{Y}_\ell$$

$h : \mathcal{X} \to \mathcal{Y}$

Dataset $S = \{ (\underline{x}^i, \underline{y}^i) : i = 1, \dots, n \}$

Let $\underline{y}^i$ be multivalued.

in form: $\underline{y}^i = (y^i_1, \dots, y^i_k)$

with $y^i_a \in \{\pm 1\}$, for each a.

basis function:

$$f_j : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}.$$

classification function:

$$h_{\underline{w}}(\underline{x}) = \arg\max_{\underline{w}} \sum_{i=1}^{n} w_i f_i(\underline{x}, \underline{y})$$

E.G. OCR example

Each $\mathcal{Y}_i$ is a character

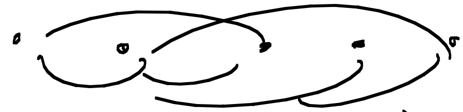
$\mathcal{Y}$ is a full word.

Webpage classification:

Each $\mathcal{Y}_i$ is a webpage label.

$\mathcal{Y}$ is a label for an entire website.

(3) Set of features



Edges.

$$f_h(\underline{x},\underline{y}) = \sum_{(i,j)\in E} f_h(\underline{x}, y_i, y_j)$$

E – edges of Graph.

penalty: $\Delta t_{\underline{x}}(\underline{y}) = \sum_{i=1}^{\ell} \Delta t_{\underline{x}}(y_i)$

with $\Delta t_{\underline{x}}(y_i) \equiv I(y_i \neq (t(x))_i)$.

Primal Problems

$$\min \frac{1}{2}\|\underline{\omega}\|^2 + C\sum_{\underline{x}} \underline{\xi}_x$$

s.t. $\underline{\omega}^T \underline{\Delta f}_{\underline{x}}(\underline{y}) \geq \Delta t_{\underline{x}}(\underline{y}) - \underline{\xi}_{\underline{x}}$, $\forall\, \underline{x},\underline{y}$.

Dual Problem

$$\max \sum_{\underline{x},\underline{y}} \alpha_{\underline{x}}(\underline{y})\Delta t_{\underline{x}}(\underline{y}) - \frac{1}{2}\Big\|\sum_{\underline{x},\underline{y}} \alpha_{\underline{x}}(\underline{y})\,\underline{\Delta f}_{\underline{x}}(\underline{y})\Big\|^2$$

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

Now we can interpret $\alpha_{\underline{x}}(\underline{y})$ as a probability distribution $\rightarrow$ with normalization C.

The dual function is a function of the expectation of $\Delta t_{\underline{x}}(\underline{y})$ & $\underline{\Delta f}_{\underline{x}}(\underline{y})$.

$$\Delta t_{\underline{x}}(\underline{y}) = \sum_i \Delta t_{\underline{x}}(y_i), \quad \underline{\Delta f}_{\underline{x}}(\underline{y}) = \sum_{(i,j)\in E} \underline{\Delta f}_{\underline{x}}(y_i, y_j)$$

(4) The dual problem reduces to a formulation on the marginal unary and pairwise distributions

$$\mu_{\underline{x}}(y_i, y_j) = \sum_{\underline{y} \sim [y_i, y_j]} \alpha_{\underline{x}}(\underline{y}), \quad \forall (i,j) \in E, \ \forall y_i, y_j, \forall \underline{x}$$

$$\mu_{\underline{x}}(y_i) = \sum_{\underline{y} \sim [y_i]} \alpha_{\underline{x}}(\underline{y}), \quad \forall i, \forall y_i, \forall \underline{x}.$$

$\underline{y} \sim [y_i, y_j]$
full assignment
$\underline{y}$ consistent with
partial assignment
y_i, y_j.

Must enforce consistency constraints between the $\mu_{\underline{x}}(y_i, y_j)$ and $\mu_{\underline{x}}(y_i)$

$$\sum_{y_i} \mu_{\underline{x}}(y_i, y_j) = \mu_{\underline{x}}(y_j), \quad \forall y_j, \ \forall (i,j) \in E.$$

Because:

$$\sum_{y} \alpha_x(y) \Delta_x(y) = \sum_{y} \sum_{i} \alpha_{\underline{x}}(y) \Delta t_{\underline{x}}(y_i) = \sum_{i, y_i} \Delta t_x(y_i) \sum_{\underline{y} \sim [y_i]} \alpha_x(\underline{y})$$

$$= \sum_{i, y_i} \mu_x(y_i) \Delta t_x(y_i)$$

similarly, the second term can be expressed in form

$$-\frac{1}{2} \sum_{x, x'} \sum_{\substack{(i,j) \\ y_i, y_j}} \sum_{\substack{r,s \\ y_r, y_s}} \mu_{\underline{x}}(y_i, y_j) \mu_{\underline{\hat{x}}}(y_r, y_s) \Delta f_{\underline{x}}(y_i, y_j)^T \Delta f_{\underline{x}}(y_r, y_s)$$

$$+ \sum_{\underline{x}} \sum_{i, y_i} \mu_{\underline{x}}(y_i) \Delta t_{\underline{x}}(y_i).$$

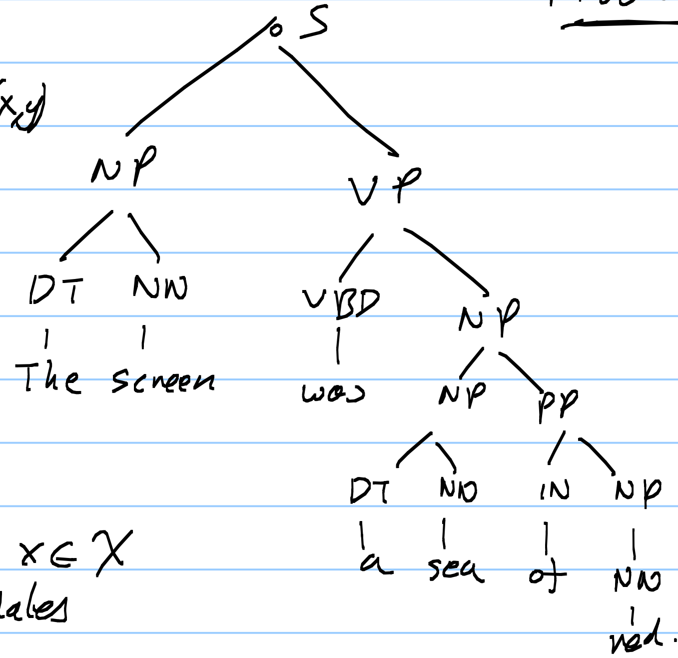
Algorithm can maximize this, by variant of Belief Propagation

Go to 7

(5) Max-Margin Parsing.

$f_{\underline{w}}(x) = \underset{y \in G(\underline{x})}{\arg\max}\ \underline{w} \cdot \underline{\Phi}(x,y)$

Input $\underline{x}$,
$G(\underline{x})$ is a set of parses.



Production Rules

$S \to (NP, VP)$

$\underline{G}$ maps an input $x \in X$ to a set of candidates parses $\underline{G}(x) \subseteq Y$

$L : X \times Y \times Y \to \mathbb{R}^+$

$L(x, y, \hat{y})$ is the penalty for making parse y on data x when true parse is $\hat{y}$.

Want to estimate a linear discriminant form:

$$f_{\underline{w}}(x) = \underset{y \in G(\underline{x})}{\arg\max}\ \underline{w} \cdot \underline{\Phi}(x,y)$$

The arg max can be performed by dynamic programming.

(6)

$$y_i = \arg\max_{y \in G(x_i)} \underline{w} \cdot \underline{\Phi}(x_i, y)$$

The margin of the parameters $\underline{w}$ on example i and parse y

$$\underline{w} \cdot \underline{\Phi}(x_i, y_i) - \underline{w} \cdot \underline{\Phi}(x_i, y) = \underline{w} \cdot (\underline{\Phi}_{i,y_i} - \underline{\Phi}_{i,y})$$

y_i is the true parse.

Primal.

$$\min \frac{1}{2} \|\underline{w}\|^2 + C \sum_i \xi_i$$

s.t. $\underline{w} \cdot (\underline{\Phi}_{i,y_i} - \underline{\Phi}_{i,y}) \geq L_{i,y} - \xi_i \quad \forall y \in G(x_i)$

$L_{i,y} = L(x_i, y_i, y)$

Dual.

$$\max C \sum_{i,y} \alpha_{i,y} L_{i,y} - \frac{1}{2} \| C \sum_{i,y} (I_{i,y} - \alpha_{i,y}) \underline{\Phi}_{i,y} \|^2$$

s.t. $\sum_y \alpha_{i,y} = 1, \ \forall i, \quad \alpha_{i,y} \geq 0, \ \forall i, y$

$I_{i,y} = I(x_i, y_i, y)$ indicates whether y is the true parse y_i.

$$\underline{w}^* = C \sum_{i,y} (I_{i,y} - \alpha^*_{i,y}) \underline{\Phi}_{i,y}$$

(At 5)

(7) PCFG.   Chomsky Normal Form (CNF)

$A \to B C$ A, B, C non-terminal symbols

$A \to a$. a terminal symbol.

Two types of parts

First type $\langle A, s, e, i \rangle$

non-terminal A E.G.

start-point s $NP, 3, 5$ 'a sea'

end-point e

sentence i.   E.G.

$S \to NP, VP, 0, 2, 7$

Second type: $A \to B C, s, m, e, i$

split point m

Countable set of parts R,

$R(x,y)$ is the set of parts belonging to a particular parse

All rules are in binary-branching form

so $|R(x,y)|$ is constant across different derivations y for the same input x

$$\Phi(x,y) = \sum_{r \in R(x,y)} \phi(x,r)$$

$$L(x,y,\hat{y}) = \sum_{r \in R(x,y)} l(x,y,r)$$

One possibility – define $l(x,y,r) = 0$ only if the non-terminal A spans words s…e in the derivation y and 1 otherwise.

Indicator variables $I(x,y,r) = 1$, if $r \in R(x,y)$
$= 0$, otherwise

Dual:

$$C \sum_i E_{\alpha_i}[L_{i,y}] - \frac{1}{2} \left\| C \sum_i \Phi_{i,y_i} - E_{\alpha_i}[\Phi_{i,y}] \right\|^2$$