Lecture 2 - Examples

Thursday, January 11, 2024 9:34 AM

IFT6132 reminder: fill survey http://bit.ly/IFT6132-W24 ASAP!

today: • examples of structured prediction
• structured perceptron & friends

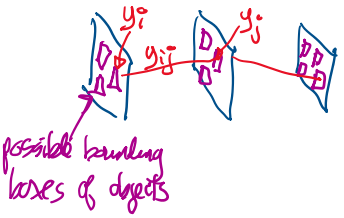
Examples:

I) word alignment (continuation)

here $x = (\underbrace{x_1^E, \dots, x_{L_E}^E}_{\text{English words}}; \underbrace{x_1^F, \dots, x_{L_F}^F}_{\text{French words}})$

$$\mathcal{Y}(x) = \{ y \in \{0,1\}^{L_E \times L_F} : \sum_j y_{ij} \le 1 \ \forall i, \ \sum_i y_{ij} \le 1 \ \forall j \}$$

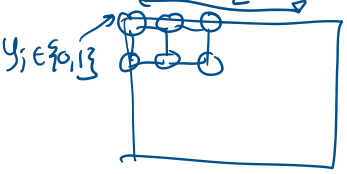
II) multi-object tracking:



$y_i := \sum_j y_{ij}$

encoding → network flow

III) image segmentation:



x = image of RGB values $L \times L$ pixels

$\mathcal{Y}(x) = \{0,1\}^{L \times L}$

background / foreground

prediction model $h_w(x)$:

standard: $h_w(x) \triangleq \underset{y \in \mathcal{Y}(x)}{\operatorname{argmax}}$ $s(x,y;w)$] compatibility score of y for x
$-E(x,y;w)$] energy fct. of E

linear model: $s(x,y;w) = \langle w, \varphi(x,y) \rangle$ $\varphi: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$

"joint feature" vector $\in \mathbb{R}^d$

word alignment: $\varphi(x,y) = \sum_{ij} y_{ij} \psi_{ij}(x_i^E, x_j^F)$

features defined on a pair of English word x_i^E French word x_j^F

- string edit distance(x_i^E, x_j^F)
- distance between i & j
- $\mathbb{1}[x_i^E, x_j^F]$ in dictionary
etc...

$s(x,y;w) = \langle w, \varphi(x,y) \rangle = \sum_{ij} y_{ij} \langle w, \psi_{ij}(x_i^E, x_j^F) \rangle$

"French word x_j^F"

$$s(x,y;w) = \langle w, \varphi(x,y)\rangle = \sum_{i,j} y_{ij} \langle w, \varphi(x_i^E, x_j^F)\rangle$$

"score to match word i to j"

$$h_w(x) = \operatorname*{argmax}_{y \in \mathcal{Y}(x)} s(x,y;w)$$

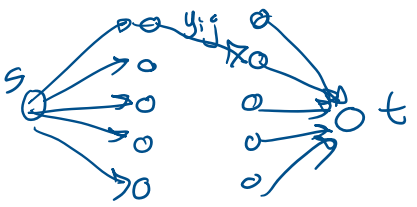
$$\rightsquigarrow \max_y \sum_{i,j} y_{ij} s_{ij}(x)$$

$$\text{s.t. } y_{ij} \in \{0,1\}, \quad \sum_j y_{ij} \le 1 \ \forall i, \quad \sum_i y_{ij} \le 1 \ \forall j$$

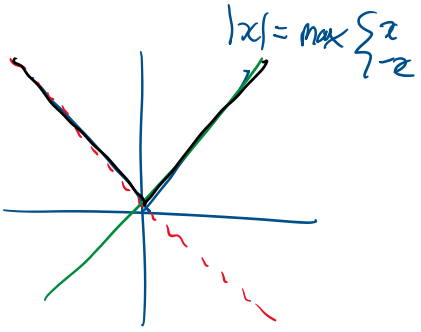
integer LP but can be solved exactly as min linear cost matching problem

eg Hungarian alg.

or more generally min cost network flow alg.



[side note: integer program with LP relaxation (exact)]



Learning w?

I) structured perceptron:

- initialize w_0
- repeat for $t = 0, \dots$
  - sample i_t
  - let $\hat{y}_t = h_{w_t}(x^{(i_t)}) = \operatorname*{argmax}_{y \in \mathcal{Y}(x^{(i_t)})} \langle w, \varphi(x^{(i_t)}, y)\rangle$ } "decoding oracle"
  - $w_{t+1} = w_t + \eta \left(\varphi(x^{(i_t)}, y^{(i_t)}) - \varphi(x^{(i_t)}, \hat{y}_t) \right)$

stepsize; $\Rightarrow$ boost score ground truth; penalize prediction

for stability: output $\hat{w}_T = \frac{1}{T+1} \sum_{t=0}^{T} w_t$ ← "Polyak averaging"

(*) structured perceptron can be interpreted as doing stochastic subgradient method (opt.) on the following non-smooth obj.:

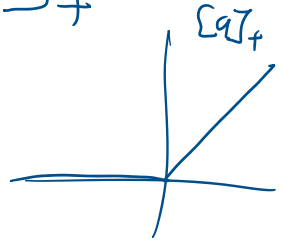
$$\hat{g}(w) = \frac{1}{n} \sum_{i=1}^{n} g^{\text{percep}}(x^{(i)}, y^{(i)}; w)$$

$$g^{\text{percept.}}(x,y;w) \triangleq \left[\max_{\tilde{y} \in \mathcal{Y}} \langle w, \varphi(x,\tilde{y})\rangle - \langle w, \varphi(x,y)\rangle \right]_+$$

$[a]_+$

$$\ell^{hinge}(x,y;w) = \left[\max_{\tilde{y} \in \mathcal{Y}} \langle w, \varphi(x,\tilde{y}) \rangle - \langle w, \varphi(x,y) \rangle \right]_+$$

where $[a]_+ \triangleq \begin{cases} a & \text{if } a \geq 0 \\ 0 & \text{o.w.} \end{cases}$



(if $y^{(i)} \in \mathcal{Y}$, then this is always ≥ 0 and $[\cdot]_+$ is not needed)

10h32

II) conditional random field

defines $p_w(y|x) \propto \exp(\langle w, \varphi(x,y) \rangle)$ $\quad \frac{(\exp(\cdot)}{Z(x)}$

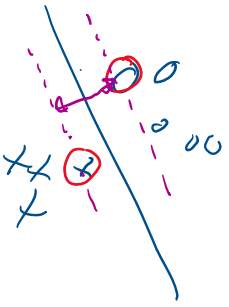
$$h_w(x) = \arg\max_{y \in \mathcal{Y}(x)} p_w(y|x) = \arg\max_{y} \langle w, \varphi(x,y) \rangle$$

then maximum conditional likelihood on training set to learn $\hat{w}$

$$\hat{w}^{CRF} = \frac{1}{n} \sum_{i=1}^{n} \ell^{CRF}(x^{(i)}, y^{(i)}; w) + \underbrace{\lambda \|w\|^2}_{\text{regularizer}}$$

$$\ell^{CRF}(x,y;w) \triangleq -\log p_w(y|x)$$

$$= \log \left(\underbrace{\sum_{\tilde{y}} \exp(\langle w, \varphi(x,\tilde{y}) \rangle)}_{Z_w(x)} \right) - \langle w, \varphi(x,y) \rangle$$



Issues: • $\ell(y,y')$ doesn't appear in it

• $\sum_{\tilde{y} \in \mathcal{Y}} \exp(\langle w, \varphi(x,\tilde{y}) \rangle)$ can be difficult

e.g. #P-complete for $\mathcal{Y}$ = set of all matchings ?

III) structured SVM

intuition: want $s(x^{(i)}, y^{(i)}; w) \geq s(x^{(i)}, \tilde{y}; w) + \ell(y^{(i)}, \tilde{y}) \quad \forall \tilde{y} \in \mathcal{Y}_i \triangleq \mathcal{Y}(x^{(i)})$

$\min \|w\|^2$ s.t. ↑ "hard margin structured SVM"

(binary SVM: $y \in \{-1,+1\}$ $h_w(x) = \text{sgn}(\langle w, \varphi(x) \rangle)$

$y_i \langle w, \varphi(x^{(i)}) \rangle \geq 1$

hinge loss: $\left[1 - y_i \langle w, \varphi(x^{(i)}) \rangle \right]_+$

soft-margin structured SVM: $R(w)$ ↑ $\frac{1}{n} \sum_{i=1}^{n} \ell^{svm}(x^{(i)}, y^{(i)}; w)$

soft-margin structured SVM:

$$\min_{w,\xi} \quad \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_i \xi_i$$

$$\xi_i + \langle w, \varphi(x^{(i)}, y^{(i)})\rangle \geq \langle w, \varphi(x^{(i)}, \tilde{y})\rangle + \ell(y^{(i)}, \tilde{y}) \quad \forall \tilde{y} \in \mathcal{Y}, \forall i$$

QP with an exponential # of constraints

equivalent (non-smooth) formulation:

$$\min_w \quad \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^n \ell^{SVM}(x^{(i)}, y^{(i)}; w)$$

$\xi_i \geq 0$

where $\ell^{SVM}(x, y; w) = \left[\max_{\tilde{y}\in\mathcal{Y}(x)} \left[\langle w, \varphi(x, \tilde{y})\rangle + \ell(y, \tilde{y})\right] - \langle w, \varphi(x, y)\rangle\right]_+$

"structured hinge loss"

(suppose that $y \in \mathcal{Y}(x)$) → not needed

OCR – optical character recognition example

x: sequence of images of characters

y: characters → B R A C E

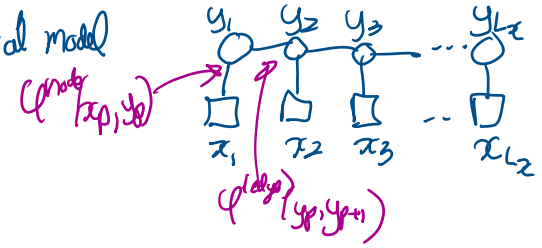
$x = (x_1, \ldots, x_{L_x})$ $x_p \in \{0,1\}^{16\times 8}$

$y(x) = \Sigma^{L_x}$ $\Sigma = \{A, B, \ldots, Z\}$

in max-margin Markov network (M^3-net) paper:

$$\langle w, \varphi(x,y)\rangle = \sum_{p=1}^{L_x} \langle w^{(node)}, \varphi^{(node)}(x_p, y_p)\rangle + \sum_{p=1}^{L_x-1} \langle w^{(edge)}, \varphi^{(edge)}(y_p, y_{p+1})\rangle$$

graphical model



$$p_w(y|x) = \frac{1}{Z_w(x)} \exp(\langle w, \varphi(x,y)\rangle) = \frac{1}{Z_w(x)} \prod_{c\in\mathcal{C}} \psi_c(x, y_c)$$

where $\mathcal{C} = \{\{p, p+1\}\}$ (edges)

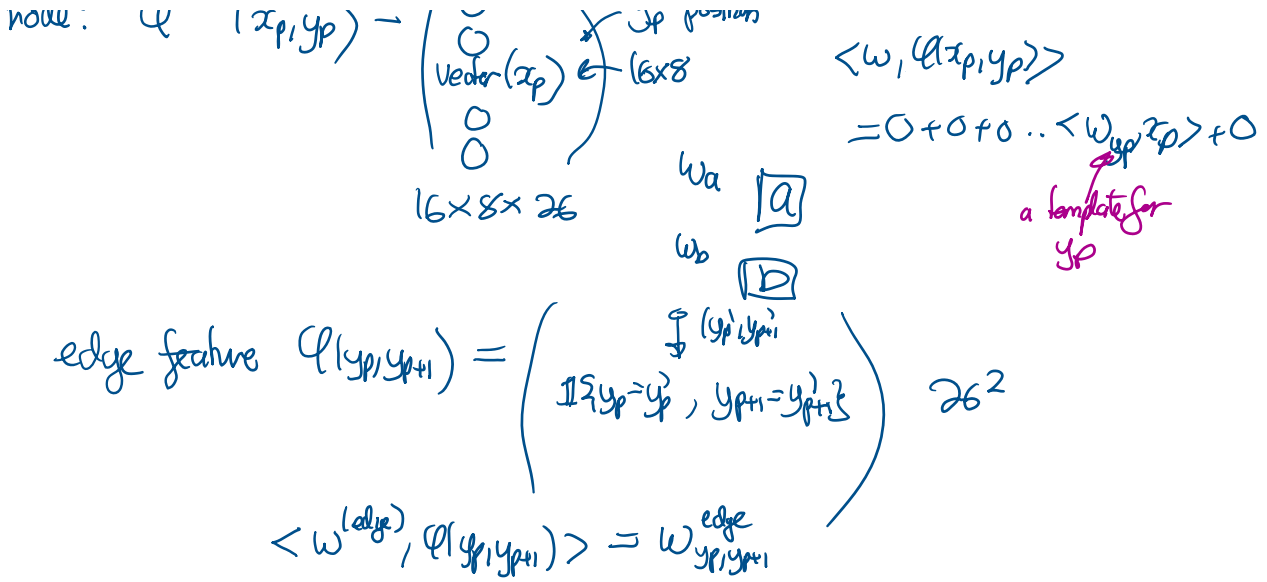
(chain structure)

notation: $y_c \triangleq (y_i)_{i\in c}$

⇒ can compute $\arg\max_{y\in\mathcal{Y}(x)} \langle w, \varphi(x,y)\rangle$ using max product / sum alg. aka. Viterbi alg.

node: $\varphi^{(node)}(x_p, y_p) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \text{vec}(x_p) \\ 0 \\ \vdots \end{pmatrix} \in 16\times 8$ ← y_pth position

$\langle w, \varphi(x_p, y_p)\rangle$


now: $\varphi(x_p, y_p) \rightarrow$
vector(x_p)
16×8
16×8×26
$\langle w, \varphi(x_p, y_p) \rangle$
$= 0 + 0 + 0 \ldots \langle w_{y_p}, x_p \rangle + 0$
a template for y_p
w_a
w_b
edge feature $\varphi(y_p, y_{p+1}) =$
$\mathbb{I}\{y_p = y'_p, y_{p+1} = y'_{p+1}\}$
26^2
$\langle w^{(edge)}, \varphi(y_p, y_{p+1}) \rangle = w^{edge}_{y_p, y_{p+1}}$