Lecture 2 - examples

Thursday, January 12, 2023 1:28 PM

Reminder: do not forget to fill survey http://bit.ly/IFT6132-W23 ASAP if not done!

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

Examples

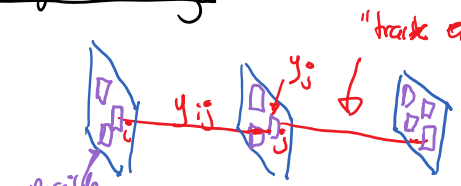
I) word alignment (continuation)

here $x = (x_1^E, x_2^E, \dots, x_{L_E}^E ; x_1^F, \dots, x_{L_F}^F)$

English words 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:

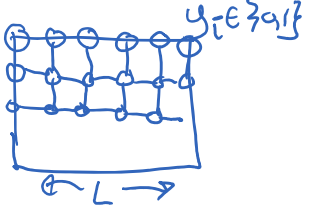


$$y_i = \sum_j y_{ij}$$

possible bounding boxes of objects

encoding → network flows

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. E

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

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

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

features defined on a pair of English word x_i^E French " 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.

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

"score to match i to j"

$h_w(x) = \arg\max_{y \in \mathcal{Y}(x)} s(x,y;w) \rightsquigarrow \max_y \sum_{ij} y_{ij} s_{ij}(x)$

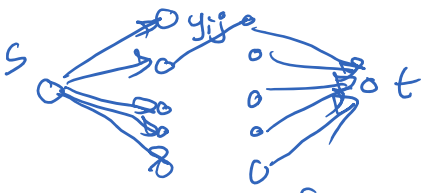
s.t. $y_{ij} \in \{0,1\}$

$\sum_j y_{ij} \le 1 \ \forall i$

$\sum_i y_{ij} \le 1 \ \forall j$

can be solved exactly as min linear cost matching problem

e.g. Hungarian alg or more generally min cost network flow alg



14h29

[redonte: integer program with LP relaxation]

good question about tradeoff : tractable search vs. more powerful models

Learning w?

I) structured perceptron :

- initialize w_0
- repeat for $t = 0, \dots$
  - sample i_t
  - let $\hat{y}_t = h_{w_t}(x^{(i_t)}) = \arg\max_{y \in \mathcal{Y}(x^{(i_t)})} \langle w_t, \varphi(x^{(i_t)}, y)\rangle$ 'decoding oracle'
  - $w_{t+1} = w_t + \eta\left(\varphi(x^{(i_t)}, y^{(i_t)}) - \varphi(x^{(i_t)}, \hat{y}_t)\right)$

η : step-size

$\varphi(x^{(i_t)}, y^{(i_t)})$: boost score of ground truth

$\varphi(x^{(i_t)}, \hat{y}_t)$: penalize incorrect prediction

$\text{score}_{t+1}(\tilde{y}) = \text{score}_t(\tilde{y}) + \eta\left[\langle \varphi(y^{(i_t)}), \varphi(\tilde{y})\rangle - \langle \varphi(\hat{y}_t), \varphi(\tilde{y})\rangle\right]$

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.:

$$\hat{\mathcal{J}}(w) = \frac{1}{n}\sum_{i=1}^{n} \ell^{percep.}(x^{(i)}, y^{(i)}; w)$$

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

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

II) conditional random field

define $p_w(y|x)$ as $\exp\left(\langle w, \phi(x,y)\rangle\right)$

$$h_w(x) = \operatorname*{argmax}_{y\in\mathcal{Y}(x)} p_w(y|x) = \operatorname*{argmax}_{y} \langle w, \phi(x,y)\rangle$$

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

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

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

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

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

• $\sum_{\tilde{y}\in\mathcal{Y}} \exp(\langle w, \phi(x,\tilde{y})\rangle)$ is often intractable

e.g. #complete problem 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: $y \in \{-1, +1\}$, $h_w(x) = \text{sgn}(\langle w, \phi(x)\rangle)$]

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

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

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

QP with an exponential # of constraints

$$\begin{cases} \min\limits_{w,\xi} & \frac{\lambda \|w\|^2}{2} + \frac{1}{n}\sum_i \xi_i \\ & \xi_i + \langle w, \varphi(x^{(i)}, y^{(i)}) \rangle \geq \langle w, \varphi(x^{(i)}, \tilde{y}) \rangle + \ell(y^{(i)}, \tilde{y}) \quad \forall \tilde{y} \in \mathcal{Y}_i, \forall i \end{cases}$$

equivalent (non-smooth) formulation:

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

"loss augmented decoding"

where $\ell^{\text{SVM}}(x, y; w) \triangleq \max\limits_{\tilde{y} \in \mathcal{Y}(x)} [\langle w, \varphi(x, \tilde{y}) \rangle + \ell(y, \tilde{y})] - \langle w, \varphi(x, y) \rangle$

"structured hinge loss" (suppose that $y \in \mathcal{Y}(x)$)