Lecture 2 - examples

Thursday, January 9, 2020 13:33

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

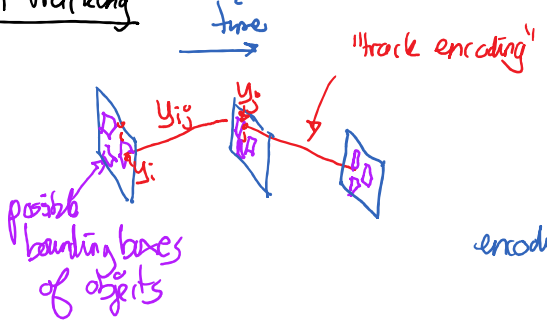
Examples:

I) word alignment (continuation)

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

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

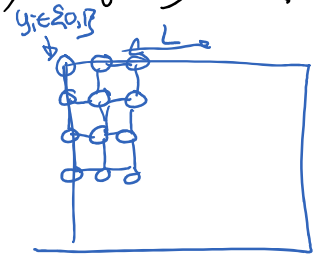
II) multi-object tracking



$y_i = \sum_j y_{ij}$

encoding → network flow

III) image segmentation



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

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

---

prediction model $h_w(x)$

standard: $h_w(x) \triangleq \underset{y \in \mathcal{Y}(x)}{\operatorname{argmax}}$ "score" $s(x,y;w)$] compatibility score of y with x

$-E(x,y;w)$] E → energy fct.

linear model: $s(x,y;w) = \langle w, \underbrace{\varphi(x,y)}_{\text{"joint feature" vector} \in \mathbb{R}^d} \rangle$ $\varphi: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$

Word alignment : $\varphi(x,y) = \sum_{ij} y_{ij} \underbrace{\psi(x_i^E, x_j^F)}_{\in \mathbb{R}^d}$

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

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

etc...

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

"score to match i to j"

$$h_w(x) = \operatorname*{argmax}_{y \in \mathcal{Y}(x)} s(x,y;w) \rightsquigarrow \max_y \sum_{ij} y_{ij} s_{ij}(x) \quad \text{s.t. } y_{ij} \in \{0,1\}, \ \sum_i y_{ij} \le 1, \ \sum_j y_{ij} \le 1$$

can be solved exactly (linear) as min cost matching problem

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

[sidenote: integer program with tight LP relaxation]

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"
  - $w_{t+1} = w_t + \eta \left(\varphi(x^{(i_t)}, y^{(i_t)}) - \varphi(x^{(i_t)}, \hat{y}_t) \right)$

η: step-size; $\Rightarrow$ boost score of ground truth; penalize prediction score

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 optim. on the following non-smooth objective

doing stochastic subgradient optim. on the following non-smooth objective

$$\hat{\mathcal{L}}(w) = \frac{1}{n}\sum_{i=1}^{n} \ell^{\text{percept}}(x^{(i)}, y^{(i)}; w)$$

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

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

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

[4h3]

II) conditional random fields

define $p_w(y|x) \propto \exp(\langle w, \phi(x, y) \rangle)$

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

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

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

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

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

issues:
- $\ell(y, y')$ doesn't appear in it
- $\sum_{\tilde{y} \in \mathcal{Y}} \exp(\langle w, \phi(x, \tilde{y}) \rangle)$ can be difficult

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

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

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

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

$$\text{QP with exponential \# of constraints} \begin{cases} \min_{w,\xi} & \underbrace{\frac{\lambda\|w\|^2}{2}}_{R(w)} + \frac{1}{n}\sum_{i=1}^n \xi_i \\ & \xi_i + \langle w, \varphi(x^{(i)}, y^{(i)})\rangle \geq \langle w, \varphi(x^{(i)}, \tilde{y})\rangle + \ell(y^{(i)}, \tilde{y}) \quad \forall \tilde{y} \in \mathcal{Y}_i, \forall i \end{cases}$$

equivalent (non-smooth) formulation:

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

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

("loss-augmented decoding")

("structured hinge loss")

(suppose that $y \in \mathcal{Y}(x)$)

OCR - optical character recognition example

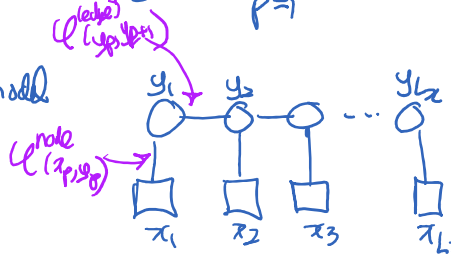
x: sequence of images of characters  $x = (x_1, \ldots, x_{L_x})$ $x_p \in \{0,1\}^{16\times 8}$

y B R A C E $\quad \mathcal{Y}(x) = \mathcal{E}^{L_x}$ $\mathcal{E} = \{A, \ldots, Z\}$

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



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

notation: $y_C \triangleq (y_i)_{i \in C}$

$\Rightarrow$ can compute $\arg\max_y \langle w, \varphi(x, y) \rangle$ using max-product aka Viterbi alg. or max sum

node: $\varphi^{(node)}(x_p, y_p) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \text{Vector}(x_p) \\ \vdots \\ 0 \\ 0 \end{pmatrix}$ ← y_pth position

16×8 (size of Vector(x_p))

$16 \times 8 \cdot 26$ (26 = # of characters)

w_a [a]

w_b [b]

$\langle w, \varphi(x_p, y_p) \rangle$
$= 0 + \dots + 0$
$+ \langle w_{y_p}, x_p \rangle + 0 \dots$

(w_{y_p}: a template for character y_p)

edge feature: $\varphi(y_p, y_{p+1}) = \begin{pmatrix} \mathbb{1}\{y_p = y_p', y_{p+1} = y_{p+1}'\} \end{pmatrix}$ ↕ 26^2 (indexed by (y_p', y_{p+1}'))

$\langle w^{(edge)}, \varphi(y_p, y_{p+1}) \rangle = w^{(edge)}_{y_p, y_{p+1}}$