Lecture 2 - examples

Tuesday, January 19, 2021 14:14

IFT6132: reminder: fill survey http://bit.ly/IFT6132-W21 ASAP!

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

Examples:

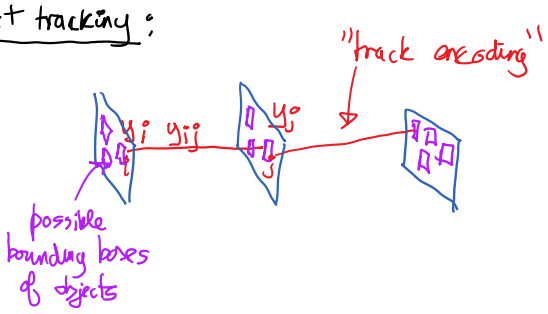
I) word alignment (continuation)

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

English words French words

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

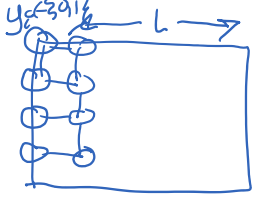
II) multi-object tracking:



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

encoding → network flow

III) image segmentation:



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

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

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

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

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

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..

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

$s(x, y; w) = \langle w, \varphi(x,y) \rangle = \sum_{i,j} y_{ij} \langle w, \psi(x_i^E, x_j^F) \rangle$ ← "French word x_j" | etc..

"score to match word i to j"

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

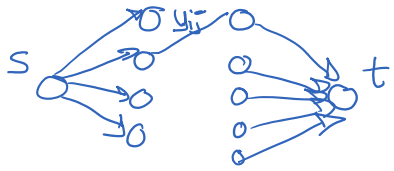
$\rightsquigarrow \max_y \sum_{i,j} y_{ij} s_{ij}(x)$ — can be solved exactly

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

(is min linear cost matching problem

e.g. Hungarian alg

or more generally min cost network flow alg.



[sidenote: integer program with LP relaxation]

14h23

Learning w?

I) structured perceptron:

- initialize w_0
- repeat for $t = 0, \dots,$
  - sample i_t
  - let $\hat{y}_t = h_{w_t}(x^{(i_t)}) = \underset{y \in \mathcal{Y}(x^{(i_t)})}{\operatorname{argmax}} \langle w, \varphi(x^{(i_t)}, y) \rangle$ "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; $\varphi(x^{(i_t)}, y^{(i_t)})$: $\Rightarrow$ boost score ground truth; $\varphi(x^{(i_t)}, \hat{y}_t)$: 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.:

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

$$\ell^{\text{percept.}}(x, u; w) \triangleq \left[\max_{\tilde{y}} \langle w, \varphi(x, \tilde{y}) \rangle - \langle w, \varphi(x, u) \rangle \right]$$

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

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

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

II) conditional random field

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

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

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

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

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

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

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

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

$R(w)$: $\frac{\lambda}{2} \|w\|^2$

$+ \frac{1}{n} \sum_i \xi_i$

QP with an exponential # of constraints

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

equivalent (non-smooth) formulation:

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

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

"loss augmented decoding"

"structural hinge loss"

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

OCR - optical character recognition example:

x: sequence of images of characters 

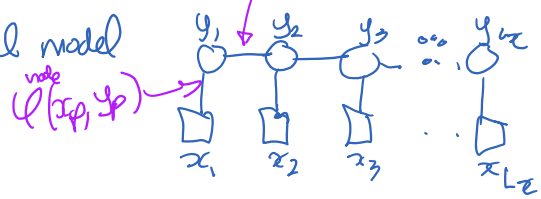
$x = (x_1, \dots, x_{L_x})$ $x_p \in \{0,1\}^{16\times 8}$

y B R A C E $\quad$ $y(x) = \Sigma^{L_x}$ $\Sigma = \{A, \dots, Z\}$

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



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

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

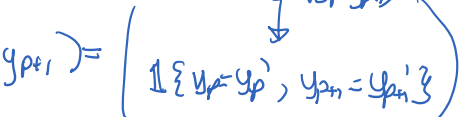
notation $y_c \triangleq (y_i)_{i \in c}$ $\Rightarrow$ Can compute $\arg\max_y \langle w, \varphi(x,y)\rangle$ using max product alg. aka. Viterbi alg. or max sum

node: $\varphi^{(node)}(x_p, y_p) =$ 

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

w_b  $\quad = 0 + 0 + \dots + \langle w_{y_p}, x_p\rangle + 0 + \dots$

a template for y_p

edge feature: $\varphi(y_p, y_{p+1}) =$ 

$\langle w^{(edge)}, \dots \rangle = w_{\dots}^{(edge)}$

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