Lecture 2 - scribbles

Thursday, January 10, 2019 13:32

today: structured prediction examples in more details
structured perceptron & friends

<u>some important aspects of structured prediction</u> (vs. binary classification)

1) $\mathcal{Y}$ output space is usually exponentially big

2) structured error function $\ell(y, y')$

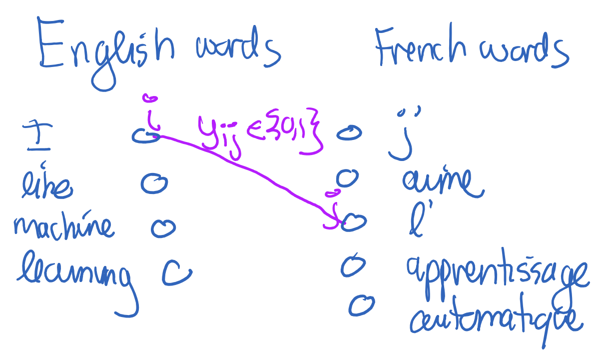
3) sometimes <u>constraints</u> on pieces of y

need an "encoding function" for $y = (y_1, \dots, y_p) \in \mathbb{R}^p$

↳ mapping from abstract structured output to their representation ...

<u>examples</u>:

a) word alignment:



$$y = (y_{ij})_{(i,j) \in E} \in \{0,1\}^{|E|}$$

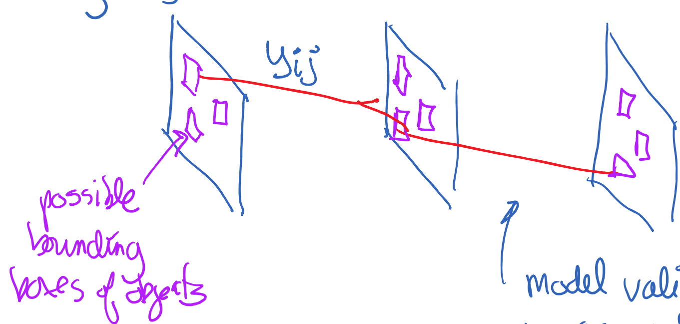
all possible edges (all pairs from E − F) matching constraints $\begin{cases} \sum_j y_{ij} \leq 1 & \forall i \\ \sum_i y_{ij} \leq 1 & \forall j \end{cases}$

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

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

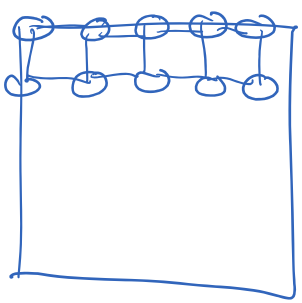
b) multi-object tracking

frame



model valid "tracks" using network flow structure

c) image segmentation



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

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

background (0), foreground (1)

standard $h_w(x)$:

"score"

$$h_w(x) \triangleq \underset{y \in \mathcal{Y}(x)}{\operatorname{argmax}} \; s(x,y;w) \quad] \text{ compatibility fct. of } y \text{ for } x$$

$$-E(x,y;w) \quad] \text{ energy fct.}$$

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

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

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

$$s(x,y;w) = \langle w, \varphi(x,y) \rangle = \sum_{i,j} y_{ij} \underbrace{\langle w, \psi(x_i^E, x_j^F) \rangle}_{\text{"score to match } i \text{ to } j\text{"}}$$

$$h_w(x) = \underset{y \in \mathcal{Y}(x)}{\operatorname{argmax}} \; s(x,y;w) \longrightarrow \max_y \sum_{i,j} y_{ij} s_{ij}(x,w)$$

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

$\sum_i y_{ij} \leq 1$

$\sum_j y_{ij} \leq 1$

can be solved exactly as min cost matching problem

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

aside: about integer program

aside: about integer program has a tight LP relaxation

- or more generally min cost network flow algorithm

Learning w

I) structured perceptron

- initialize at w_0
- repeat for $t=0,\dots,$
  - sample i_t
  - let $\hat{y}_t = h_{w_t}(x^{(i_t)}) = \underset{y \in \mathcal{Y}(x^{(i_t)})}{\arg\max} \langle w, \varphi(x^{(i_t)}, y)\rangle$ } decoding oracle
  - $w_{t+1} = w_t + \underbrace{\eta}_{\text{step-size}} \left(\underbrace{\varphi(x^{(i_t)}, y^{(i_t)})}_{\text{boost ground truth score}} - \underbrace{\varphi(x^{(i_t)}, \hat{y}_t)}_{\text{penalize prediction (mistake) score}} \right)$

for stability:

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

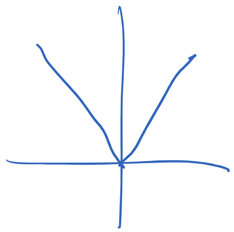
(*) structured perceptron can be interpreted as doing stochastic subgradient optimization on the following non-smooth objective

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

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

$$\mathcal{L}^{percept}(x,y;w) \triangleq \Big[\max_{\tilde{y}\in\mathcal{Y}} \langle w, \varphi(x,\tilde{y})\rangle - \langle w, \varphi(x,y)\rangle\Big]_+$$

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



$|x| = \max(\{x, -x\})$

[aside: when $y^{(i)} \in \mathcal{Y}$, then this is always ≥ 0 and $[\cdot]_+$ is <u>not needed</u>]

II) <u>conditional random field</u>:

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

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

then do <u>maximum conditional likelihood</u> on training set to learn $\hat{w}$

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

$$\mathcal{L}^{CRF}(x,y;w) = -\log p_w(y|x)$$

$$= \underbrace{\log\Big(\sum_{\tilde y \in \mathcal{Y}} \exp(\langle w, \varphi(x,\tilde y)\rangle)\Big)}_{\log Z_w(x)} - \langle w, \varphi(x,y)\rangle$$

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 matching!

III) structured SVM

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

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

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

$y^{(i)}\langle w, \varphi(x^{(i)})\rangle \geq 1$ hard margin constraint

soft-margin SVM: $\min_{\xi, w} \underbrace{\frac{\lambda \|w\|^2}{2}}_{R(w)} + \underbrace{\frac{1}{n}\sum_i \xi_i}_{\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}^{SVM}(x^{(i)}, y^{(i)}; w)}$

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

QP with exponential # of constraints

n

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

where $\quad$ "structured hinge loss"

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

exercise to reader: $\quad$ show if $y^{(i)} \in \mathcal{Y}(x^{(i)})$;

then $\quad \ell^{SVM}(x^{(i)}, y^{(i)}; w) \geq \ell(y^{(i)}, h_w(x^{(i)}))$

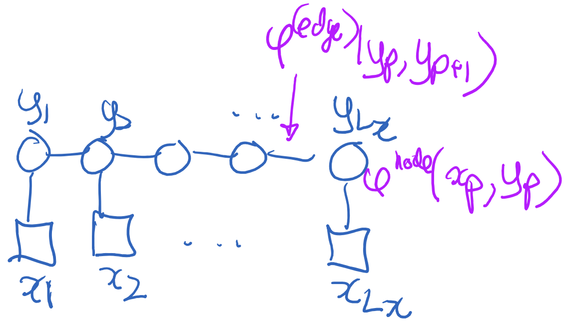
OCR - optical character recognition example

x: sequence of image of characters $\quad$ $\boxed{b}\,\boxed{r}\,\boxed{a}\,\boxed{c}\,\boxed{e}$ $\quad$ $x = (x_1, \dots, x_{L_x})$ $\quad$ $x_p \in \{0,1\}^{16 \times 8}$

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

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

graphical model



$\underset{x_1}{\square}\quad \underset{x_2}{\square}\quad \cdots \quad \underset{x_{L_x}}{\square}$

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

$$p(y|x) = \frac{1}{Z_w(x)} \exp(\langle w, \varphi(x,y)\rangle) = \frac{1}{Z} \prod_{C\in\mathcal{C}} \psi_C(x, y_C) \qquad \text{here } C = \{p, p+1\}$$

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

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

node: $$\varphi(x_p, y_p) = \begin{pmatrix} 0 \\ \vdots \\ \text{vector}(x_p) \\ \vdots \\ 0 \\ 0 \end{pmatrix} \begin{matrix} \}\, 16\times 8 \\ \\ \leftarrow y_p^{th} \text{ position} \\ \\ \\ \\ \end{matrix}$$

$16\times 8 \cdot 26$

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

w_a $\boxed{a}$

w_b $\boxed{b}$

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

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

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