Lecture 1 - scribbles

Friday, January 6, 2017
14:18

gen./discriminative continuum

prediction → learn a mapping $h: \mathcal{X} \to \mathcal{Y}$

$p_w(x, y)$ ← output
↑ input
[generative model on $\mathcal{X} \times \mathcal{Y}$]

↳ $p_w(y|x) \to h_w(x) = \arg\max_{y \in \mathcal{Y}} p_w(y|x)$

"more discriminative" →

$p_w(x,y)$ → learn $\hat{w}$ by ML

$p_w(y|x)$ → learn $\hat{w}$ by MCL

$h_w: \mathcal{X} \to \mathcal{Y}$, $\ell(y, y')$ → learn $h_{\hat{w}}$ using surrogate loss minimization } related empirical risk minimization
structured SVM

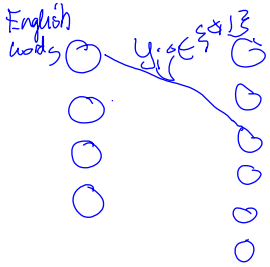
← more assumptions / less robust

some important aspects of structured prediction:

1) $\mathcal{Y}$ output space is usually exponentially big
2) error function $\ell(y, y')$
3) constraints on pieces of y

need an "encoding function": e.g. $y = (y_1, \ldots, y_p) \in \mathbb{R}^p$

word alignment

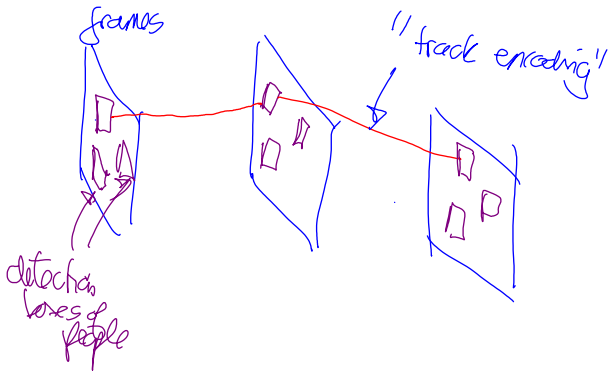


$$y = (y_{ij})_{ij \in E} \in \{0,1\}^{|E|}$$

matching constraint
$$\begin{cases} \sum_j y_{ij} \leq 1 & \forall i \\ \sum_i y_{ij} \leq 1 & \forall j \end{cases}$$

another example:

multi-object tracking



Simple structured prediction models

1) $h_w(x) \triangleq \underset{y \in \mathcal{Y}}{\operatorname{argmax}}$ "score" $sc(x,y;w)$] "compatibility function"

$-E(x,y;w)$] energy function

example: $sc(x,y;w) = \langle w, \underbrace{\varphi(x,y)}_{\in \mathbb{R}^d} \rangle$

joint feature vector

word alignment example: recall $y = (y_{ij}) \in \{0,1\}^{|E|}$

word alignment example: recall $y = (y_{ij}) \in \{0,1\}^{|E|}$

$$\varphi(x,y) = \sum_{ij} y_{ij} \psi(x_{ij})$$

features defined on English word i and French j

($\psi(x_{ij})$ = (string edit distance(i,j), $\mathbb{1}\{(i,j)$ are in dictionary, distance index i & j, ⋮))

here, you can do: $\operatorname{argmax}_{y \in \text{matchings}} s(x,y,w)$ using min-cost network algorithms

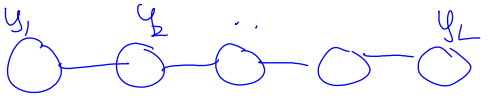
$$\langle w, \varphi(x,y) \rangle = \sum_{ij} y_{ij} \underbrace{\langle w, \psi(x_{ij}) \rangle}_{\text{represents how much we want to align } i \text{ & } j}$$

another example from a undirected graphical model

$$p_w(y|x) = \frac{1}{Z_w(x)} \prod_c \psi_c(y_c, x)$$

$$= \exp\Big(\sum_c \underbrace{\log \psi_c(y_c, x)}_{\langle w, \varphi_c(y_c, x_c) \rangle} - A(w;x) \Big)$$

sequence model $\log p_w(y|x) = \sum_t \langle w, \varphi(y_t, y_{t+1}; x) \rangle + \text{const.}$



argmax here → Viterbi algorithm

1) how do we learn w, given some training data $(x_i, y_i)_{i=1}^n$?

simplest algorithm:

Structured perceptron:

go over training set:

- sample i_t
- [predict on x_{i_t}]
- let $\hat{y}_t = h_{w_t}(x_{i_t}) = \arg\max_{y \in \mathcal{Y}} \langle w, \varphi(x_{i_t}, y) \rangle$

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

step-size; boost ground truth score; penalize prediction

for stability:

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

this can be interpreted as doing stochastic subgradient optimization on the following non-smooth objective:

$$L(w) = \frac{1}{n} \sum_{i=1}^{n} \left[\max_{y \in \mathcal{Y}} \langle w, \varphi(x_i, y) \rangle - \langle w, \varphi(x_i, y_i) \rangle \right]_+$$

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

if $y_i \in \mathcal{Y}$, then always positive and so $[\cdot]_+$ is not needed

Conditional random fields (CRF)

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

then do MCL on the training set

caveat: need (implicitly) to be able to sum over $\mathcal{Y}$ i.e. $\sum_{y \in \mathcal{Y}} \exp(\langle w, \varphi(x,y)\rangle)$

#P-complete for set of matchings!

structured SVM:

intuition: want $\langle w, \varphi(x_i, y_i)\rangle \geq \langle w, \varphi(x_i, y)\rangle + \ell(y_i, y) \quad \forall y \in \mathcal{Y}_i$

$\min \|w\|^2$

s.t. ——

(hard margin structured svm)

(I) $\min_{w,\xi} \quad \|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} \xi_i$

$\xi_i + \langle w, \varphi(x_i, y_i)\rangle \geq \langle w, \varphi(x_i, y)\rangle + \ell(y_i, y) \quad \forall y \in \mathcal{Y}_i, \forall i$] QP with exponential # of constraints

equivalently

$\xi_i + \langle w, \varphi(x_i, y_i)\rangle \geq \max_{y \in \mathcal{Y}} \{\langle w, \varphi(x_i, y)\rangle + \ell(y_i, y)\}$] non-linear constraint

equivalent problem

(II) $\min_{w} \quad \|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} \left[\max_{y \in \mathcal{Y}_i} \{\langle w, \varphi(x_i, y)\rangle + \ell(y_i, y)\} - \langle w, \varphi(x_i, y_i)\rangle\right]$

structured hinge loss