Lecture 1 - scribbles - intro

Tuesday, January 23, 2018
13:29

bit.ly/IFT6132-W18

Structured prediction basics & setup

Learning problem:

given a training dataset $D = (x^{(i)}, y^{(i)})_{i=1}^{n}$

$x^{(i)} \in \mathcal{X}$ $y^{(i)} \in \mathcal{Y}$

goal: learn a prediction mapping $h_w : \mathcal{X} \to \mathcal{Y}$ (w: parameter)

that has low generalization error $L(w; P) \triangleq \mathbb{E}_{(x,y)\sim P}[\ell(y, h_w(x))]$

P: 'test' distribution on (x,y)

ℓ: structured error function

"risk" in ML (Vapnik) / statistical decision loss

[See lecture 4 & 5 of my PGM → review of statistical decision theory]

regularized ERM (empirical risk minimization)

$$\hat{L}(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(y^{(i)}, h_w(x^{(i)})) + R(w)$$

R: regularizer

ℓ: not cts. in w; non-convex; messy ... NP hard to minimize in general

⇒ replace it with surrogate loss (ML) / contrast fct. (statistics)

$$\hat{f}(w) = \arg\min_{w} \sum_{i=1}^{n} \ell(x^{(i)}, y^{(i)}, w) + R(w)$$

ℓ: surrogate loss, e.g. convex in w

↳ examples: · structured hinge loss → structured SVM
· log-loss → CRF

$\hat{f}(w)$: M-estimator in statistics

---

generative vs discriminative learning continuum

generative modeling of (x, y) (input, output)

$p_w(x, y)$

↳ condition $p_w(y|x) \rightarrow h_w^{(p_w)}(x) = \arg\max_y p_w(y|x)$

"more discriminative"

$p_w(x|y)$	$p_w(y|x)$ $\rightarrow h_w^{(p_w)}$	$h_w : \mathcal{X} \rightarrow \mathcal{Y}$, $\ell(y, y')$
learn $\hat{w}$ by MCL	learn $\hat{w}$ by MCL	learn $\hat{w}$ by using surrogate loss minimization e.g. structured SVM } related to ERM

← more assumptions / less robust to classification task

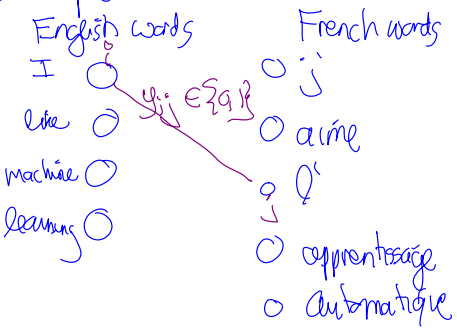
some important aspects of structured prediction (vs binary classification)

1) $\mathcal{Y}$ output space is usually exponentially big
2) structured error function $\ell(y, y')$

3) sometimes constraints on pieces of y

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

word alignment example:



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

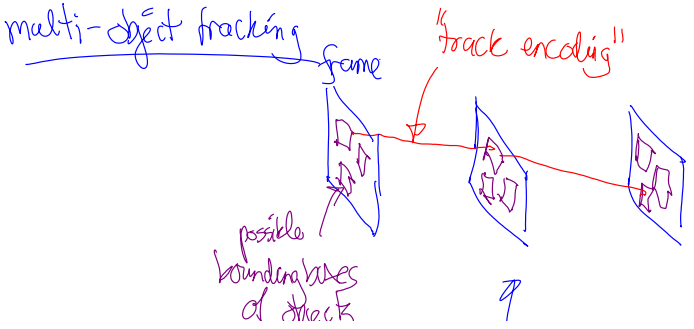
E = possible edges (cell pairs)

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

here $x = (x^E_1, \dots, x^E_{L_E}, x^F_1, \dots, x^F_{L_F})$ (English words, French words)

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

another example:

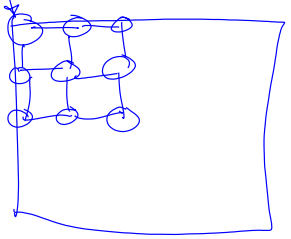


of objects

model this with network flow structure

Image Segmentation:

$y_i \in \{0,1\}$



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

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

background, foreground

standard hw(x):

"score"

$h_w(x) \triangleq \underset{y \in \mathcal{Y}(x)}{\operatorname{argmax}}$ $S(x,y;w)$ } compatibility ff. of y with x

$-E(x,y;w)$ } energy fct.

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 $\in \mathbb{R}^d$

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

$\psi(x_i^E, x_j^F) \in \mathbb{R}^d$

features defined on English word x_i^E and French " x_j^F

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

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

"score to match i to j"

$$h_w(x) = \operatorname*{argmax}_{y \in \mathcal{Y}(x)} s(x, y; w) \quad \rightarrow \quad \max_y \sum_{ij} y_{ij} s_{ij}(x)$$

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

$\sum_j y_{ij} \leq 1$

$\sum_i y_{ij} \leq 1$

Can be solved exactly as min cost matching problem

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

Learning w:

structured perceptron:

- initialize w_0
- repeat for $t = 0, \ldots,$
  - 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; $\varphi(x^{(i_t)}, y^{(i_t)})$: boost ground truth score; $\varphi(x^{(i_t)}, \hat{y}_t)$: 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 optimization on the following non-smooth objective

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

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

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

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

Conditional random field:

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

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

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

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

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

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

$\underbrace{\quad \text{yey} \quad}$

$\log Z_w(x)$

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

#P-complete for $\mathcal{Y}$ = set of matchings

Structured SVM:

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

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

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

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

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

$\lambda\|w\|^2 + \frac{1}{n}\sum_{i=1}^n \xi_i$

QP with exponential # of constraints

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

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

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"

it turns out that if $\ell(\cdot,\cdot) = \Delta(\cdot,\cdot)$ ($\ell^{SVM}(x^{(i)}, y^{(i)}; w) \leq \Delta(h_w(x^{(i)}), y^{(i)})$

it turns out that if $y^{(i)} \in \mathcal{Y}(x^{(i)})$, then $g^{SVM}(x^{(i)}, y^{(i)}; w) \geq \ell(y^{(i)}, h_w(x^{(i)}))$