Structured Perceptron & Structured SVM

Monday, November 21, 2022 1:25 PM

Multiclass Perceptron

Key Idea: If the prediction is incorrect ($\hat{y} \neq y^*$) then increase the score of y^* and decrease the score of $\hat{y}$

Algo: (Multiclass Perceptron) $\theta \in \mathbb{R}^M$ $f(\cdot,\cdot) \in \mathbb{R}^M$

$\theta \leftarrow 0$

while not converged:

for (x, y^*) in D:

$$\hat{y} = \underset{j \in \{1,\dots,K\}}{\text{argmax}}\ \theta^T f(\vec{x}, j)$$

if $\hat{y} \neq y^*$:

$$\theta \leftarrow \theta + \underline{f(\vec{x}, y^*)} - f(\vec{x}, \hat{y})$$

Q1: fill in the blank

$$(\theta + f(\vec{x}, y^*))^T f(\vec{x}, y^*) \geq \theta^T f(\vec{x}, y^*)$$

Algo. (Structured Per.)

"

"

"

$$\hat{y} = \underset{y \in \mathcal{Y}(\vec{x})}{\text{argmax}}\ \theta^T f(\vec{x}, y)$$

"

"

Structured Perceptron

- **Algo:**
- Same as Multiclass Perceptron w/ each possible output structure $\vec{y} \in \mathcal{Y}(\vec{x})$ as a "class"
- Training examples are $(\vec{x}, \vec{y})$ where $\vec{x} \in \mathcal{X}$ and $\vec{y} \in \mathcal{Y}(\vec{x})$
- Feature functions $f(\vec{x}, \vec{y}) \in \mathbb{R}^M \leftarrow$ repr. of $(\vec{x}, \vec{y})$ pair
- Predict $\hat{y} = \underset{\vec{y} \in \mathcal{Y}(\vec{x})}{\text{argmax}}\ \theta^T f(\vec{x}, \vec{y})$

Key Question: How to compute?

If $\exp(\theta^T f(\vec{x}, \vec{y}))$ decomposes multiplicatively according to some factor graph, then solve this "MAP inference problem" w/ MILP, e.g.

$$\exp(\theta^T f(\vec{x}, \vec{y})) = \exp\left(\sum_c \theta^T f_c(\vec{x}, \vec{y}_c)\right)$$

$$= \prod_c \exp(\theta^T f_c(\vec{x}, y_c))$$

$$= \prod_c \exp(\theta^T f_c(\vec{x}, y_c))$$
$$= \prod_c \psi_c(\vec{x}, y_c)$$

Structured SVM

<u>Data</u>: $\{(\vec{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ $\quad \vec{x}^{(i)} \in \mathcal{X} \quad \vec{y}^{(i)} \in \mathcal{Y}(\vec{x})$

<u>Model</u>: Linear $\hat{y} = h(\vec{x}) = \underset{\vec{y} \in \mathcal{Y}(\vec{x})}{\operatorname{argmax}} \underbrace{w^T f(\vec{x}, \vec{y})}_{s_w(x,y)}$

<u>Q.P.</u>:
$$\min_{w,e} \frac{1}{2}(\|w\|_2)^2 + C \sum_{i=1}^{N} \sum_{\hat{y} \in \mathcal{Y}(\vec{x}^{(i)})} e_{i,\hat{y}}$$
$$\text{s.t. } \left. \begin{array}{l} s_w(x^{(i)}, y^{(i)}) - s_w(x^{(i)}, \hat{y}) \geq \ell(y^{(i)}, \hat{y}) - e_{i,\hat{y}} \\ \\ e_{i,\hat{y}} \geq 0 \end{array} \right] \begin{array}{l} \forall i \\ \forall \hat{y} \in \mathcal{Y}(\vec{x}^{(i)}) \setminus y^{(i)} \end{array}$$

<u>Key Idea</u>:
- Score of true $y^{(i)}$ should be larger than the score of the predicted $\hat{y}$ by some margin
- margin to be scaled by the loss

<u>Ex</u>: Sorted by score

$\vec{x}^{(i)}$ = time flies fast	$s_w(\vec{x}^{(i)}, \cdot)$		$\ell(y^{(i)}, \cdot)$ Hamming Loss
$y^{(i)}$ = N V A	7		0
$\hat{y}_1$ = V V A	5	+2	1
$\hat{y}_2$ = N N A	4	+3	1
$\hat{y}_3$ = V N A	3.5	+3.5	2
⋮	0		⋮
	⋮		
$\hat{y}_{LAST}$ = ...	−7		

Q: How many constraints are in this Q.P.?

A: $O(N \cdot \max_i |\mathcal{Y}(\vec{x}^{(i)})|)$

Q.P. w/ fewer constraints

<u>Key Idea</u>: fold a maximization problem into constraint

<u>Q.P.</u>:
$$\min_{w,e} \frac{1}{2}\|w\|_2^2 + C \sum_{i=1}^{N} e_i$$
$$\text{s.t. } s_w(\vec{x}^{(i)}, \vec{y}^{(i)}) \geq \left[\max_{\hat{y} \in \mathcal{Y}(\vec{x})} s_w(\vec{x}^{(i)}, \hat{y}) + \ell(y^{(i)}, \hat{y})\right] - e_i \quad \forall i$$

$$\text{s.t.}\quad s_w(\vec{x}^{(i)}, \vec{y}^{(i)}) \geq \left[\max_{\hat{y}\in\mathcal{Y}(\vec{x})} s_w(\vec{x}^{(i)}, \hat{y}) + \ell(y^{(i)}, \hat{y})\right] - e_i \quad \forall i$$

$$e_i \geq 0$$

these N constraints and N slack vars replaced the $O(N \cdot \max_i |\mathcal{Y}(\vec{x}^{(i)})|)$ constraints and slack vars above

Q.P. w/ Hinge Loss

$$\min_w \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{N} \underbrace{\left[\max\left(0, \left[\max_{\hat{y}\in\mathcal{Y}(\vec{x})} s_w(\vec{x}^{(i)}, \hat{y}) + \ell(y^{(i)}, \hat{y})\right] - s_w(\vec{x}^{(i)}, y^{(i)})\right)\right]}$$

structured hinge loss $= \ell_w^{S.H.}(\vec{x}^{(i)}, y^{(i)})$

Sub-Gradient

$$\nabla \ell_w^{S.H.}(\vec{x}^{(i)}, y^{(i)}) = \begin{cases} 0 & \text{if } \ell_w^{S.H.}(\vec{x}^{(i)}, \vec{y}^{(i)}) = 0 \\ f(\vec{x}^{(i)}, y^{(i)}) - f(\vec{x}^{(i)}, \hat{y}) & \text{otherwise} \end{cases}$$

$$\text{where}\quad \hat{y} = \underbrace{\operatorname*{argmax}_{\hat{y}\in\mathcal{Y}(\vec{x}^{(i)})} s_w(\vec{x}^{(i)}, \hat{y}) + \ell(y^{(i)}, \hat{y})}$$

not your standard MAP Inference problem

Train w/ Stochastic Subgradient Gradient

OH

$p(z) = \phi_z$

ϕ

0.2	0.1	0.6	0.1

$p(z) = \phi_z$

ϕ

0.2	0.1	0.6	0.1

$$\sum_z p(z) = 1$$

z discrete

$$\int_z p(z)\, dz = 1$$

$z \in \mathbb{R}$

$$\int_z p(z|x)\, dz = 1$$