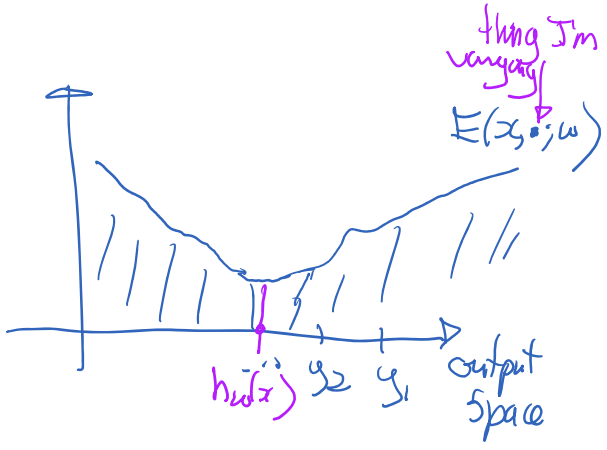
Lecture 3 - scribbes

Tuesday, January 15, 2019 14:12

today : energy based methods & surrogate losses
multiclass

energy based methods : [Lecun & al. 2006]



model : $h_w(x) = \underset{y \in \mathcal{Y}(x)}{\operatorname{argmin}} E(x, y; w)$ "energy fct."

$= \underset{y \in \mathcal{Y}(x)}{\operatorname{argmax}} S(x, y; w)$ "score / compatibility fct."

ingredients modeling :
1) what is $E(x, y; w)$? e.g. $S(x, y; w) = \langle w, \varphi(x, y) \rangle$
2) how do you compute $\underset{y \in \mathcal{Y}(x)}{\operatorname{argmin}} E(x, y; w)$? → "inference" / "decoding"

learning :
3) how to evaluate "quality" $E(x, y; w)$ on a training set? → surrogate loss $\hat{\mathcal{L}}(w)$
in general : $\tilde{\mathcal{L}}(x^{(i)}, y^{(i)}, E(\cdot, \cdot))$ "loss functional"
4) how to minimize $\hat{\mathcal{L}}(w)$ to learn w? → optimization tricks

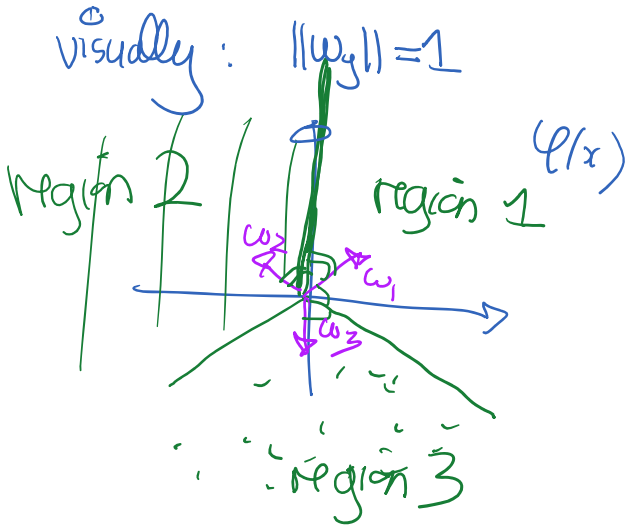
flat multiclass case. : $\in \mathbb{R}^d$

flat multiclass case:

"flat" setting $h_w(x) = \arg\max_y \langle w_y, \varphi(x) \rangle$ ($w_y \in \mathbb{R}^d$)

equivalent to $\varphi(x,y) = (0, 0, \varphi(x), 0, 0, 1)$ $\in \mathbb{R}^{d \cdot k}$ (k = # of classes), $\varphi(x)$ in the y^{th} position

visually: $\|w_y\| = 1$



OCR node vs. feature map

$\langle w, \varphi^{(node)}(x,y) \rangle = \sum_p \langle w, \varphi^{(node)}(x, y_p) \rangle$

$= \sum_p \sum_{y_p'} \mathbb{1}\{y_p = y_p'\} \langle w_{y_p'}, \varphi(x) \rangle$

→ here "sharing" of parameters between different pieces of the labels → "structure"

aside: in structured prediction, usually absorb "bias" in parameters

standard binary classification: $\text{sgn}(\langle w, x \rangle + b)$

$\tilde{\varphi}(x) = (\varphi(x), 1)$

$\langle \tilde{w}, \tilde{\varphi}(x) \rangle = \langle w, \varphi(x) \rangle + b$

$\tilde{w} = (w, b)$

open question: regularizing or not the bias

open question: regularizing or not the bias in structured prediction, does it matter?

surrogate losses:

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

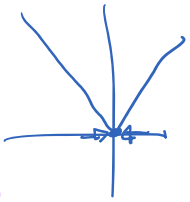
1) perceptron loss [Collins & al. 2002 EMNLP]

$$\ell(x, y; w) = \max_{\tilde{y} \in \mathcal{Y}(x)} s(x, \tilde{y}; w) - \underbrace{s(x, y; w)}_{\text{score of ground truth}} \qquad (\text{assume } y \in \mathcal{Y}(x))$$

$$s(x, y; w) = \langle w, \varphi(x, y) \rangle$$

$$\max_{\tilde{y}} \langle w, \underbrace{\varphi(x, \tilde{y}) - \varphi(x, y)}_{-\Psi_x(\tilde{y})} \rangle \geq 0$$

by using $\tilde{y} = y$



observations: 1) degenerate solution $w = 0$ or constant score over y

2) averaged perceptron alg.:

- run constant step-size stochastic subgradient method on $\hat{\ell}(w)$ → does not converge in general
- output $\hat{w}_T = \frac{1}{T+1} \sum_{t=0}^{T} w_t$ (Polyak avg.) → will converge to $w^* = 0$ when data is not separable

Comments: 1) Collins' paper → he gives error bound and generalization error guarantees for perceptron

2) (aside) connection with the "herding" alg. by Welling et al. "3rd way to learn" [see ICML 2012]

data $\xrightarrow{\text{learn}}$ $\hat{w}_{ML}$

herding → $y^{(t)}$ ← sampling

15h33

II) log-loss (CRF) (probabilistic interpretation)

Boltzmann dist. in physics ($\beta = \frac{1}{k_B T}$)

suppose $p(y|x;w) \propto \exp(\beta s(x,y;w))$

β: inverse temperature parameter

MCL → log-loss

$s(x,y;w)$

$$\ell(x,y;w) = -\frac{1}{\beta} \log p(y|x;w) = -\frac{1}{\beta} \log \left(\frac{\exp(\beta s(y))}{\sum_{\tilde{y}} \exp(\beta s(\tilde{y}))} \right)$$

β

$\sum_{\tilde{y}} \exp(\beta s(\tilde{y}))$ } $Z_\beta(x;w)$ partition

β ↳ rescaling

$\beta \left\{ \sum_{\tilde{y}} \exp(\beta s(\tilde{y}) \right\} Z_\beta(x;w)$ partition fct.

$$= \frac{1}{\beta} \log\left(\sum_{\tilde{y}} \exp(\beta s(\tilde{y})) \right) - \frac{\beta}{\beta} s(y)$$

"log-sum-exp" ⇝ "soft-max" why?

Def $\hat{y} = \operatorname{argmax}_{\tilde{y}} s(\tilde{y})$

NOTE: in deep learning book

$$\left(\frac{\exp(s(y))}{\sum_{\tilde{y}} \exp(s(\tilde{y}))} \right)_{y \in \mathcal{Y}}$$

I call this "soft-argmax"

$$\frac{1}{\beta} \log \left[\exp(\beta s(\hat{y})) \underbrace{\left[\sum_{\tilde{y}} \exp(\beta(s(\tilde{y}) - s(\hat{y}))) \right]}_{\leq |\mathcal{Y}|} \right]$$ ($s(\tilde{y}) - s(\hat{y}) \leq 0$)

$$= s(\hat{y}) + \frac{1}{\beta} \log(\underbrace{\text{stuff}}_{\leq |\mathcal{Y}|})$$

$\beta \to \infty$ (i.e. zero temp limit)

$$\frac{1}{\beta} \log\left(\sum_{\tilde{y}} \exp(\beta s(\tilde{y})) \right) \xrightarrow{\beta \to \infty} \max_{\tilde{y}} s(\tilde{y})$$

thus: $\lim_{\beta \to \infty}$ log-loss(β) ⟶ perceptron loss?

III) structured hinge loss

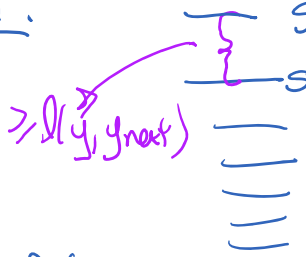
$$\ell(x, y; w) = \max_{\tilde{y} \in \mathcal{Y}(x)} \left[s(x, \tilde{y}; w) + \ell(y, \tilde{y}) \right] - s(x, y; w)$$

$$\ell^{hinge}(x,y,w) = \max_{\tilde{y}\in\mathcal{Y}(x)} \left[s(x,\tilde{y};w) + \ell(y,\tilde{y}) \right] - s(x,y;w)$$

"loss-augmented decoding"

a)

cartoon:



$\Rightarrow \ell^{hinge}(x,y;w) = 0$

b) $\ell(x,y;w) \geq \ell(y, h_w(x))$

why? $\ell(x,y;w) = \max_{\tilde{y}} [s(\tilde{y}) + \ell(\tilde{y})] - s(y)$

$\geq s(\hat{y}) + \ell(\hat{y}) - s(y)$ using $\hat{y} = \operatorname{argmax}_{\tilde{y}\in\mathcal{Y}(x)} s(\tilde{y}) = h_w(x)$

if $y \in \mathcal{Y}(x) \Rightarrow s(\hat{y}) \geq s(y)$

$\geq \ell(\tilde{y}) = \ell(y, \underbrace{h_w(x)}_{\tilde{y}})$ //

binary case:

$y \in \{-1, +1\}$ $\quad w = \begin{pmatrix} w_+ \\ w_- \end{pmatrix}$ $\quad \varphi(x,+1) = \begin{pmatrix} \varphi(x) \\ 0 \end{pmatrix}$

$h_w(x) = \operatorname{argmax} \{ \langle w_+, x \rangle, \langle w_-, x \rangle \}$

predict +1 if $\langle w_+, x\rangle \geq \langle w_-, x\rangle$

$\Leftrightarrow \langle w_+ - w_-, x\rangle \geq 0$

$\tilde{w} \triangleq w_+ - w_-$

$h_w(x) = \text{sgn}(\langle \tilde{w}, x\rangle)$

$$\ell^{SVM}(\hat{r}, y; w) = \max\{\langle w_+, x\rangle + \underbrace{\ell(y,+)}_{\mathbb{1}\{y\neq+1\}}, \langle w_-, x\rangle + \underbrace{\ell(y,-)}_{1-\mathbb{1}\{y\neq+1\}}\} - \langle w_y, x\rangle$$

$$\max\{\langle \tilde{w}, x\rangle + \langle w_-, x\rangle + \mathbb{1}\{y\neq+1\}, \langle w_-, x\rangle + 1 - \mathbb{1}\{y\neq+1\}\} - \langle w_y, x\rangle$$

$$= \max\{\langle \tilde{w}, x\rangle + \mathbb{1}, 1 - \mathbb{1}\} + \langle w_-, x\rangle - \langle w_y, x\rangle$$

case $y = +1$: $\max\{\langle \tilde{w}, x\rangle, 1\} - \langle \tilde{w}, x\rangle = [1 - \underbrace{y}_{(+1)}\langle \tilde{w}, x\rangle]_+$

$y = -1$: $\max\{\langle \tilde{w}, x\rangle + 1, 0\} + 0 = \max\{0, 1 - y\langle \tilde{w}, x\rangle\}$

overall: $[1 - y\langle \tilde{w}, x\rangle]_+$

structured SVM

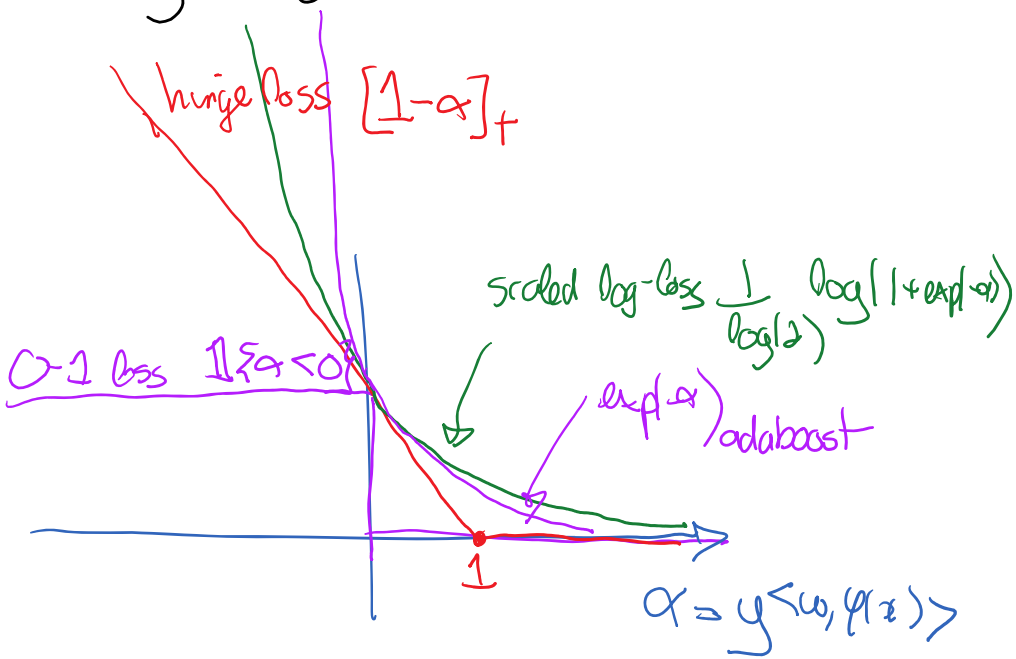
$$\ell(x, y; w) = [1 - y\langle \tilde{w}, x\rangle]_+$$

where $\tilde{w} = w_+ - w_-$

i.e. structured hinge loss reduces to binary SVM hinge loss when using $\Delta(y, \hat{y}) = \mathbb{1}\{y \neq \hat{y}\}$ and $\mathcal{Y} = \{-1, +1\}$

Binary surrogate losses

Binary surrogate losses



$y\langle w, \varphi(x)\rangle$ "margin" $\geq 0 \Rightarrow$ make no mistake

$\rightarrow \in \{-1, +1\}$

log loss: $\log(1+\exp(-\alpha))$

[logistic regression]

[Bartlett et al. 2006] → showed all these methods are consistent