Lecture 3 - surrogate losses

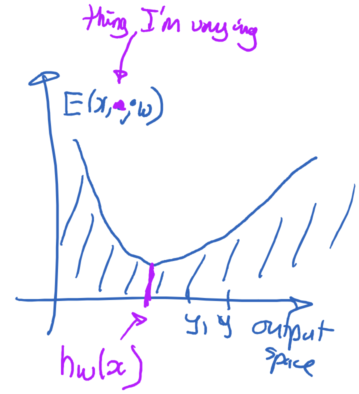
Thursday, January 21, 2021 13:39

today: energy based methods & surrogate losses
multiclass

energy based methods : [LeCun & al. 2006]

model: $h_w(x) = \underset{y \in \mathcal{Y}(x)}{\arg\min}\ E(x,y;w)$ "energy f.t."

$= \underset{y \in \mathcal{Y}(x)}{\arg\max}\ S(x,y;w)$ "score / compatibility"



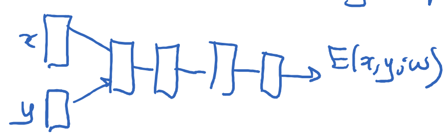
ingredients:

modeling:

1) what is $E(x,y;w)$?

e.g. $S(x,y;w) = \langle w, \varphi(x,y) \rangle$

or $E(x,y;w)$ out of a NN with x & y as input



2) how do you compute $\underset{y \in \mathcal{Y}(x)}{\arg\min}\ E(x,y;w)$? → "decoding" / "inference"
(computational)

learning:

3) how to evaluate $E(x,y;w)$ on a training set? → surrogate loss $\mathcal{J}(w)$
"quality"

in general: $\mathcal{L}(x^{(i)}, y^{(i)}; E(\cdot\,;\cdot\,;w))$
"loss function"

4) how to minimize $\mathcal{J}(w)$ to learn $\hat{w}$? → optimization tricks
(computational)

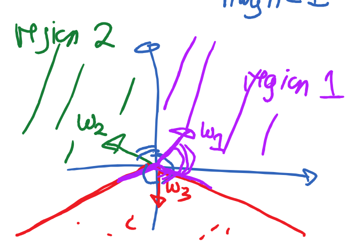
flat multiclass case

"flat" (ie. non structured) setting $h_w(y) = \underset{y}{\arg\max} \langle w_y, \underbrace{\varphi(x)}_{\in \mathbb{R}^d} \rangle$

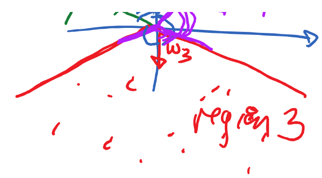
equivalent to $\varphi(x,y) = \begin{pmatrix} 0 \\ 0 \\ \varphi(x) \\ 0 \\ 0 \end{pmatrix} \in \mathbb{R}^{d \cdot k}$ ← y^{th} position; k = # of classes

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

visually: $\|w_y\| = 1$



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



contrast this flat case with structured case:

e.g. OCR node feature map $\langle w, \varphi^{(node)}(x,y)\rangle = \sum_p \underbrace{\langle w, \varphi^{(node)}(x_p, y_p)\rangle}_{\sum_{y_p'} \mathbb{1}\{y_p, y_p'\}\langle w_{y_p}, \varphi(x)\rangle}$

→ two "sharing" of parameters between pieces of the joint label → "structure"

aside: in structured prediction, usually absorb "bias" in parameters $\tilde{\varphi}(x)$

standard binary classification $\text{sgn}(\langle w, x\rangle + b)$ $\quad \tilde{\varphi}(x) = \begin{pmatrix}\varphi(x)\\ 1\end{pmatrix}$

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

open question: regularizing or not the bias does it matter in struct. pred.?

14h27

surrogate losses

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

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

$$\ell^{percep.}(x,y;w) = \Big[\max_{\tilde{y}\in\mathcal{Y}(x)} s(x,\tilde{y},w) - \underbrace{s(x,y;w)}_{\text{score of ground truth}}\Big]_+$$

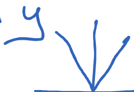
$+$ not needed if assume $y \in \mathcal{Y}(x)$

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

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

by using $\tilde{y} = y$

observations: 1) degenerate solution to $\hat{\ell}(w)$ with $w = 0$ or constant score over y



2) averaged perceptron alg:

- amounts to running 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.
[ICML 2012]
"3rd way to learn"

$$\text{data} \xrightarrow{\text{learn}} \hat{w}_{ML}$$
$$\downarrow \text{sampling}$$
$$y^{(t)}$$

herding

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

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

β "inverse temperature" parameter

$$\beta = \frac{1}{k_B T}$$ temp

MCL → log-loss

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

$s(y)$: $s(x,y;w)$

rescaling

$Z_\beta(x;w)$ partition function

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

"log-sum-exp" ⇝ "soft max" why?

let $\hat{y} = \arg\max_{\tilde{y}} s(\tilde{y})$

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

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

Note:
in deep learning book,
they call "soft-max"

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

as $\beta \to \infty$ (i.e. zero temp. limit) $\frac{1}{\beta}\log\left(\sum_{\tilde{y}} \exp(\beta s(\tilde{y}))\right) \xrightarrow{\beta\to\infty} \max_{\tilde{y}} s(\tilde{y})$

I call this "soft-argmax" thus $\lim_{\beta \to \infty}$ log-loss$(\beta) \longrightarrow$ perceptron loss

III) structured hinge loss

$$\ell^{SVM}(x, y; w) = \underbrace{\max_{\tilde{y} \in \mathcal{Y}(x)} \left[s(x, \tilde{y}; w) + \ell(y, \tilde{y}) \right]}_{\hookrightarrow \text{ loss-augmented decoding}} - s(x, y; w)$$

a) cartoon:

$s(y)$

$\geq \ell(y, y_{next})$

$s(y_{next})$

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

b) $\boxed{\ell^{SVM}(x, y; w) \geq \ell(y, h_w(x))}$

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

let $\hat{y} = \operatorname{argmax}_{\tilde{y}} s(\tilde{y})$
$= h_w(x)$

$\geq s(\hat{y}) + \ell(\hat{y}) - s(y)$

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

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

binary case: for structured hinge loss → see notes from last year

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

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

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

$\Leftrightarrow \langle \underbrace{w_+ - w_-}, \varphi(x) \rangle \geq 0$

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

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

(see notes to show)

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

(recopied from last year:)

structured hinge loss

$$\tilde{w} = w_+ - w_- \qquad h_w(x) = \mathrm{sgn}(\langle \tilde{w}, x\rangle)$$

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

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

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

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

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

case $y=-1$ $\max\{\langle \tilde{w}, x\rangle + 1, 0\} + 0 = [1 - y\langle \tilde{w}, x\rangle]_+$

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

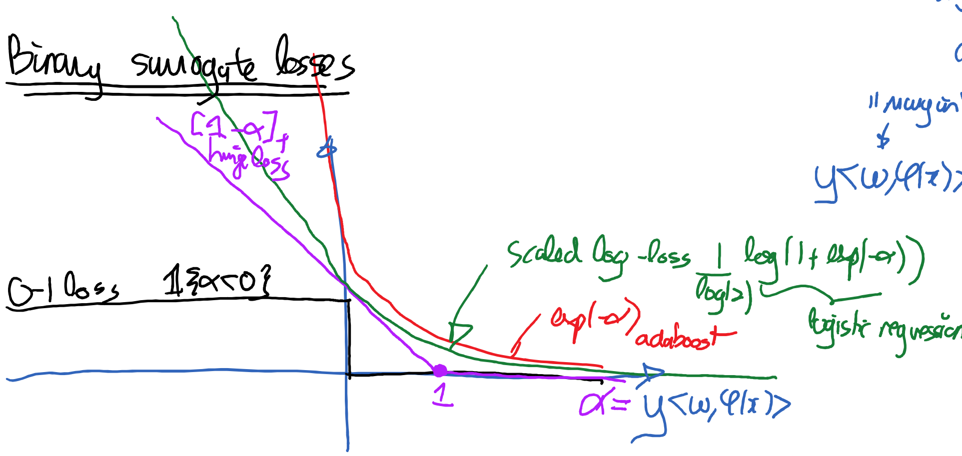
$$\ell^{SVM}(x,y;w) = [1 - y\langle \tilde{w}, x\rangle]_+$$

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

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

"margin"

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



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