today: energy-based methods & surrogate losses

energy-based methods: [Secan & al. 2006]

\[ h_w(x) = \arg\max_y E(x, y; w) \]  
\[ = \log \frac{1}{P(x, y; w)} \]

ingredients:

model: \[ h_w(x) = \arg\max_y E(x, y; w) \]  

**modeling**

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

\[ S(x, y; w) = \langle w, (y | x) \rangle \]

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

2) how do you compute \( \arg\min_y E(x, y; w) \)?

(computational)

3) how to evaluate \( E(x, y; w) \) on a training set?

4) how to minimize \( J(w) \) to learn \( \hat{w} \)?

(computational)

flat multiclass are

"flat" (i.e. non structured) setting \( h_w(y) = \arg\max_y < w_y, (y | x) > \)

equivalent to

\[ (a | y) = \begin{pmatrix} 0 \\ (y | x) \end{pmatrix} \in R^k \]

\[ < w, (a | y) > = < w_y, (y | x) > \]
\[ \langle \mathbf{w}, \phi(x, y) \rangle = \langle \mathbf{w}_y, \phi(x) \rangle \]

contrast this flat case with structured case:

e.g. OCR node feature map
\[ \langle \mathbf{w}, \phi(x, y) \rangle \leq \sum_{y} \mathbf{w}_y \phi(x, y) \]

\[ \mathbf{w}_y = \sum_{y} \mathbf{w}_y \phi(y, x) \]

heu "sharing" of parameters between pairs of the joint label “structure"

aside: in structured prediction, usually absorb "bias" in parameters \( \mathbf{w}_y \)

standard binary classification
\[ \text{sign} \left( \mathbf{w}_y x + b \right) \]

\[ \langle \mathbf{w}_y \phi(y, x) \rangle \mathbf{w}_y \]

Open question: regularize or not the bias
can it matter in struc pred?

14ht7

**Surrogate losses**

\[ \hat{L}(\mathbf{w}) = \left( \frac{1}{n} \sum_{i=1}^{n} \mathbf{s}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w}) \right) + R(\mathbf{w}) \]

1) **perceptron loss**: [Collins et al., 2002, EXACT]

\[ \mathbf{L}_{\text{ perceptron }}(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \left[ \max_{y \in \mathcal{Y}} \mathbf{s}(\mathbf{x}, y, \mathbf{w}) - \mathbf{s}(\mathbf{x}, \mathbf{y}, \mathbf{w}) \right] \]

\[ \mathbf{s}(\mathbf{x}, \mathbf{y}; \mathbf{w}) = \langle \mathbf{w}, \phi(x, y) \rangle \]

\[ \max_{y} \langle \mathbf{w}, \phi(x, y) \rangle - \langle \mathbf{w}, \phi(x, y) \rangle \geq 0 \]

by \( y \mathbf{w}_y \)

observations:

1) degenerate solution to \( \hat{L}(\mathbf{w}) \) with \( \mathbf{w} = 0 \) or constant score over \( y \)

2) averaged perceptron alg:

- arranges to run accurate constant-step-size stochastic subgradient
- output \( \mathbf{w}_t = \frac{1}{T+1} \sum_{t=0}^{T} \mathbf{w}_t \) (Polyak avg.)
Comments:
1) Collins ppm → he gives error bound and generalization error guarantee for perception
2) (Aside) connection with the "heating" alg. by Welling et al.
   "3rd way to learn" [ICML 2012]

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

\[ \text{Suppose } p(y|x;w) = \exp(\beta s(x,y;w)) \]

\[ \beta = \frac{1}{kT} \text{ (temp)} \]

MCL → log-loss

\[ s(x,y;w) = \frac{1}{\beta} \log \left( \frac{\exp(\beta s(y))}{\sum_y \exp(\beta s(y))} \right) \]

"log-sum-exp" à "soft max" why?

\[ \hat{y} = \arg \max_y \beta s(y) \]

\[ \frac{1}{\beta} \log \left( \frac{\exp(\beta s(y))}{\sum_y \exp(\beta s(y))} \right) = \frac{1}{\beta} \beta s(y) + \frac{1}{\beta} \log(\frac{\exp(\beta s(y))}{\sum_y \exp(\beta s(y))}) \]

\[ = 1 \frac{\exp(s(y))}{\sum_y \exp(s(y))} \] as \( \beta \to \infty \) (i.e. zero temp limit)

\[ \frac{1}{\beta} \log(\frac{\exp(\beta s(y))}{\sum_y \exp(\beta s(y))}) \to \max_y \frac{s(y)}{\beta} \]

Note: in deep learning book, they call "soft max"
I call this "soft-argmax" thus limit log-loss ($\rho \to \infty$) → perceptron loss

III) Structured hinge loss

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

→ loss-augmented decoding

a) \[ \text{canberra} \]

\[ \frac{s(y)}{s(y_{\text{new}})} \]

\[ \Rightarrow \]

\[ \sum_{y} f(x,y;w) = 0 \]

b) \[ f(x,y;w) \geq \log y, \text{h}_0(x) \]

Why? \[ f(x,y;w) = \max_{\hat{y}} \left[ s(\hat{y}) + \log y, \hat{y} \right] - s(y) \]

\[ \geq s(\hat{y}) + \log y - s(y) \]

\[ \text{if } y \in \mathcal{Y}(x) \Rightarrow s(\hat{y}) > s(y) \]

\[ \Rightarrow \log y = \log y, \text{h}_0(x) \]

Binary case: for structured hinge loss 

\[ y \in \mathcal{Y}_{\pm}, w = \begin{bmatrix} w_+ \\ w_- \end{bmatrix}, \quad \varphi(x;+i) = \begin{pmatrix} q_{i+1} \\ 0 \end{pmatrix} \]

\[ h_0(x) = \arg \max \left\{ <w_+, \varphi(x); x>, <w_-, \varphi(x); x> \right\} \]

Predict + if \[ <w_+, \varphi(x); x> \geq <w_-, \varphi(x); x> \]

\[ \Leftrightarrow <w_+ - w_-, \varphi(x); x> \geq 0 \]

\[ h_0(x) = \text{sign}(\hat{w}_0, x) \]

(See notes to show)

\[ g_{\text{sum}}(x,y;w) = \left[ 1 - y \hat{w}^T x \right]^+ \]

(recopied from last year)
\[ \sum (x_i, y_i; \hat{w}) = \max \left\{ \langle \hat{w}, x \rangle + 1 - \langle \hat{w}, x \rangle \right\} \]

\[ \hat{w} = w_+ - w_- \quad \text{and} \quad \text{sgn}(1 - \langle \hat{w}, x \rangle) \]

\[ \sum (x_i, y_i; \hat{w}) = \max \left\{ \langle \hat{w}, x \rangle + \langle \hat{w}, x \rangle + 1 - \langle \hat{w}, x \rangle \right\} \]

\[ \hat{w} = \text{sgn}(1 - \langle \hat{w}, x \rangle) \]

\[ \sum (x_i, y_i; \hat{w}) = \left[ 1 - \langle \hat{w}, x \rangle \right]_+ \]

where \( \hat{w} = w_+ - w_- \)

\[ \left[ 1 - \langle \hat{w}, x \rangle \right]_+ \]

\[ \text{Binary surrogate losses} \]

\[ \text{Squared log-loss} \]

\[ \text{logistic regression} \]

\[ \text{boosting} \]

\[ \text{Combination} \]

\[ \text{Bagging} \]

\[ \text{AdaBoost} \]

\[ \text{Hierarchical regression} \]

\[ \text{Scalable log-loss} \]

\[ \text{log}(1 + \exp(-\alpha y)) \]

\[ \alpha = y < \langle w, \phi(x) \rangle > \]

\[ \text{[Bookett et al., 2006] showed all these methods are consistent} \]