Top-down vs. bottom-up methods for hierarchical classification

Claudio Gentile
DICOM
Universita' dell'Insubria, Italy
claudio.gentile@uninsubria.it

joint work with:

N. Cesa-Bianchi

L. Zaniboni

State of the art:

Many hierarchical classification models/algorithms

[KS97, DC00, SL01, RS02, HCC03, Ro$^+$05, CGZ06,...]

Top-down vs bottom-up / local vs global / on-line vs batch ...

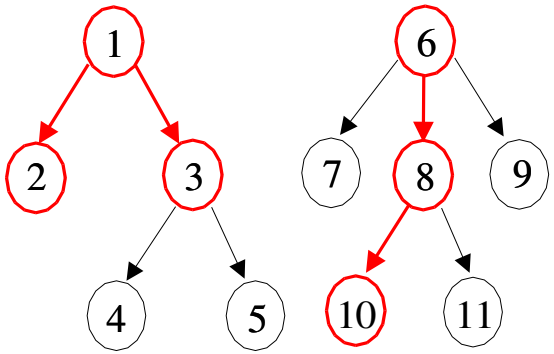
Outline:

- Hierarchical classification framework
- Bottom-up alg B-SVM:
  Combines Bayes optimal (labels are subtrees) with SVM
- Report experimental comparison between B-SVM and local top-down H-SVM on artificial and real-world medium-size datasets
- On-line learning: model, alg, regret analysis

Hierarchical classification model/1

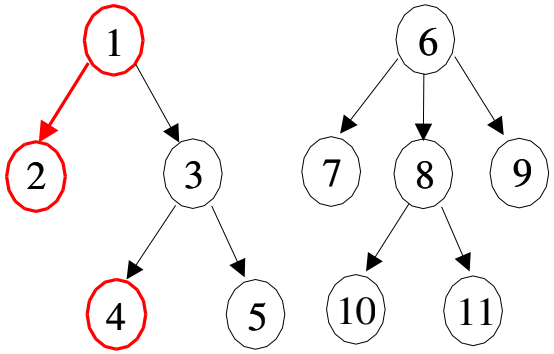
Taxonomy: tree forest ($N = 11$ nodes)

Legal multilabel: union of paths



$\boldsymbol{v}$=(1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0)

Illegal multilabel: rest



$\boldsymbol{v}$=(1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0)

Example $(\boldsymbol{x}, \boldsymbol{v}) \in \mathbb{R}^d \times \{0,1\}^N$

Hierarchical classification model/2

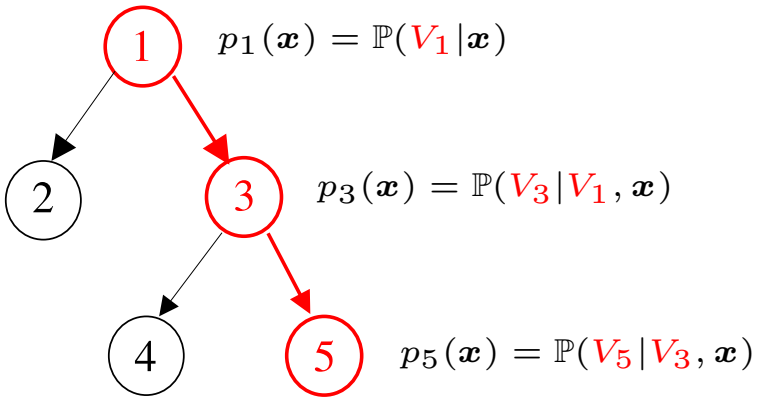
Generation of multilabels

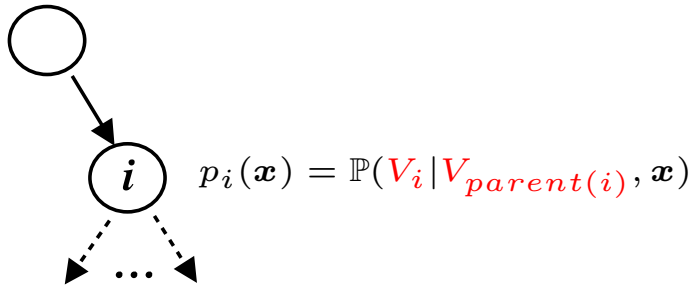
Instance $\boldsymbol{x} \in \mathbb{R}^d$

Multilabel $\boldsymbol{V} = (V_1, V_2, ..., V_N) \in \{0,1\}^N$

$\boldsymbol{V} \,|\, \boldsymbol{x} \sim \prod_{i=1}^{N} \mathbb{P}(V_i | V_{parent(i)}, \boldsymbol{x})$

Children's labels are indep. given parent's (no correlation)

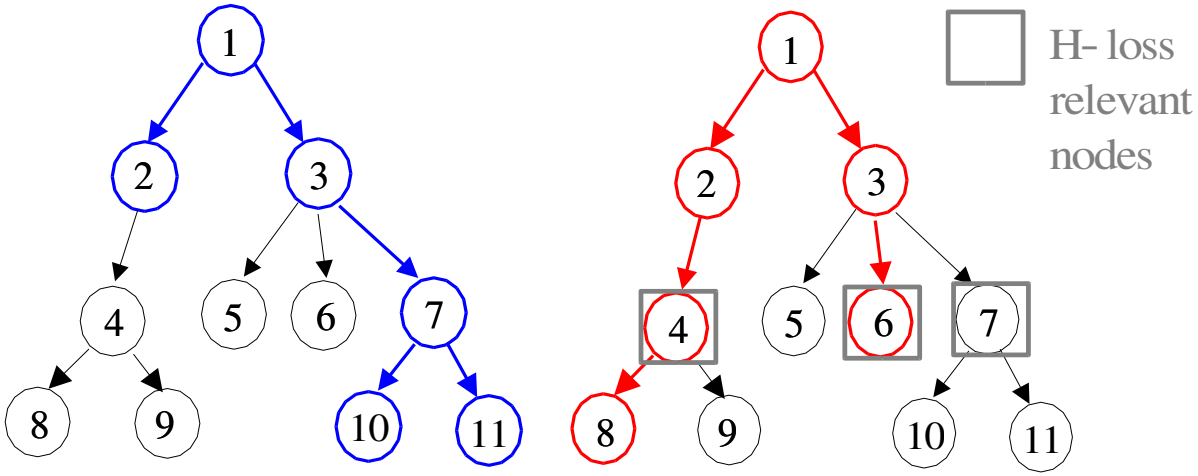




$\mathbb{P}(V_i = 1 \,|\, V_{parent(i)} = 0, \boldsymbol{x}) = 0 \quad \forall \boldsymbol{x}$

Hierarchical classification model/3

Hierarchical loss (H-loss) [CBGZ06]



$\widehat{\boldsymbol{Y}} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_N)$ $\quad \boldsymbol{V} = (V_1, V_2, ..., V_N)$

H-Loss:

(constant) cost coefficients

$$\ell_H(\widehat{\boldsymbol{Y}}, \boldsymbol{V}) = \sum_{i=1}^{N} c_i \{\hat{y}_i \neq V_i \wedge \forall j \in \text{ancest}(i) : \hat{y}_j = V_j\}$$

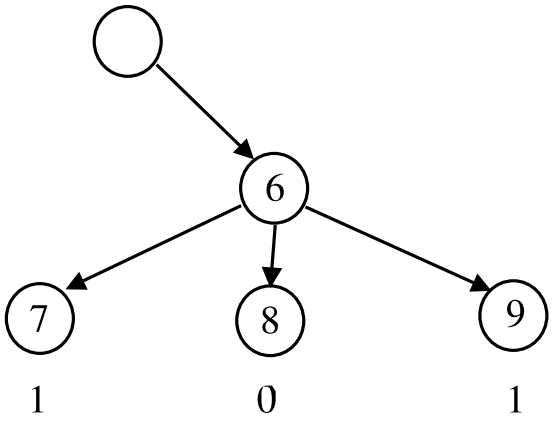
Hierarchical classification model/4

Bayes optimal classifier

$$\boldsymbol{y}^* = (y_1^*, ..., y_N^*) = \underset{\boldsymbol{y} \in \{0,1\}^N}{\operatorname{argmin}} \mathbb{E}\left[\, \ell_H(\boldsymbol{y}, \boldsymbol{V}) \,|\, \boldsymbol{x} \,\right]$$

Bottom-up message passing

Given $p_i = p_i(\boldsymbol{x})$



Leaves: $y_i^* = \{p_i \geq 1/2\}$

Hierarchical classification model/4

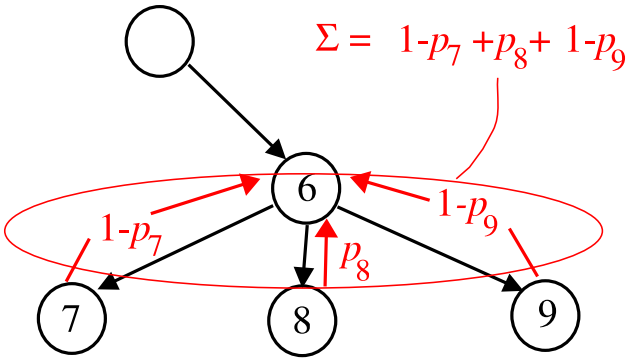
Bayes optimal classifier

$$\boldsymbol{y}^* = (y_1^*, ..., y_N^*) = \underset{\boldsymbol{y} \in \{0,1\}^N}{\operatorname{argmin}} \mathbb{E}\left[\, \ell_H(\boldsymbol{y}, \boldsymbol{V}) \,|\, \boldsymbol{x} \,\right]$$

Bottom-up message passing

Given $p_i = p_i(\boldsymbol{x})$

Messages: $y_i^*(1\text{-}p_i) + (1\text{-}y_i^*)\, p_i$



Hierarchical classification model/4

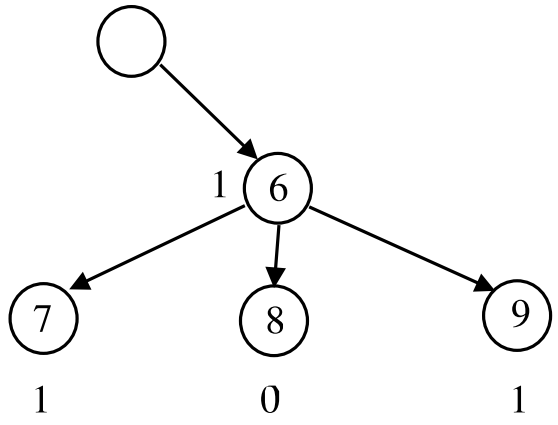
Bayes optimal classifier

$$\boldsymbol{y}^* = (y_1^*, ..., y_N^*) = \underset{\boldsymbol{y} \in \{0,1\}^N}{\operatorname{argmin}} \mathbb{E}\left[\, \ell_H(\boldsymbol{y}, \boldsymbol{V}) \,|\, \boldsymbol{x} \,\right]$$

Bottom-up message passing

Given $p_i = p_i(\boldsymbol{x})$

Nodes: $y_i^* = \{p_i \geq 1/(2\text{-}\Sigma)\}$



If $y_i^* = 0$ then $y_j^* = 0 \;\; \forall\, j \in$ subtree(i)

Hierarchical classification model/4

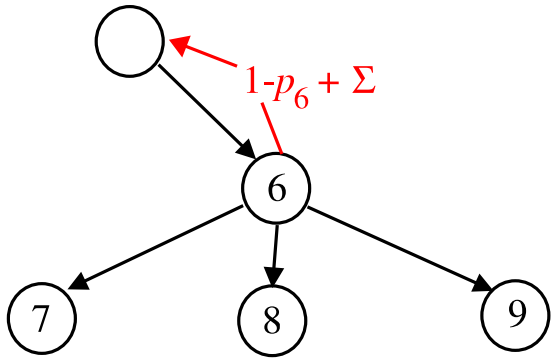
Bayes optimal classifier

$$\boldsymbol{y}^* = (y_1^*, ..., y_N^*) = \underset{\boldsymbol{y} \in \{0,1\}^N}{\operatorname{argmin}} \mathbb{E}\left[\, \ell_H(\boldsymbol{y}, \boldsymbol{V}) \,|\, \boldsymbol{x} \,\right]$$

Bottom-up message passing

Given $p_i = p_i(\boldsymbol{x})$

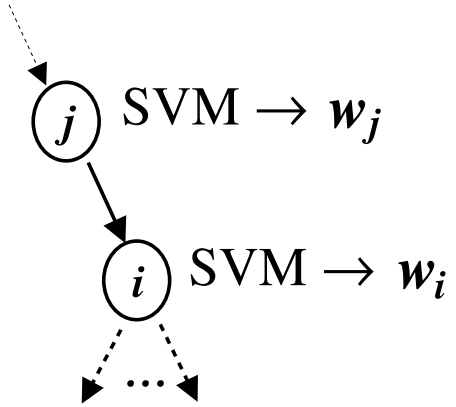
Messages: $y_i^*(1\text{-}p_i) + (1\text{-}y_i^*)\, p_i + \Sigma$



B-SVM algorithm (bottom-up)/1

SVM sitting at each node

$\boldsymbol{w}_i \leftarrow \text{SVM}(\{(\boldsymbol{x}, \boldsymbol{v}) : v_{parent(i)} = 1\})$



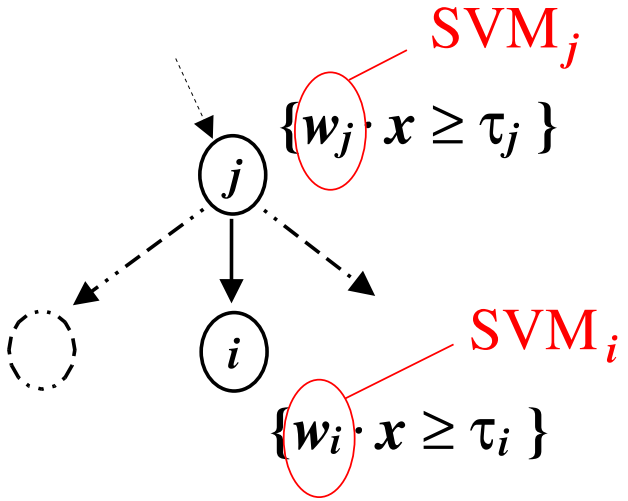
$$p_i(\boldsymbol{x}) \simeq \hat{p}_i(\boldsymbol{x}) = \frac{1}{(1 + e^{\alpha_i \boldsymbol{w}_i \cdot \boldsymbol{x}})}$$

Platt's method [Pl99] to fit parameters $\alpha_1, ..., \alpha_N$ via cross-validation on training set

Then bottom-up label assignment

B-SVM algorithm (bottom-up)/2

Results in SVM with
modified thresholds $\tau_j > 0$
$\{\boldsymbol{w}_j \cdot \boldsymbol{x} \geq \tau_j\}$



$$\hat{p}_i(\boldsymbol{x}) = \frac{1}{(1 + e^{\alpha_i \boldsymbol{w}_i \cdot \boldsymbol{x}})}$$

τ_j depends on behavior of children i:

- $\tau_j >> 0$ when $\hat{p}_i(\boldsymbol{x}) \approx 1/2$
- $\tau_j \approx 0$ when $|\hat{p}_i(\boldsymbol{x}) - 1/2| \approx 1/2$

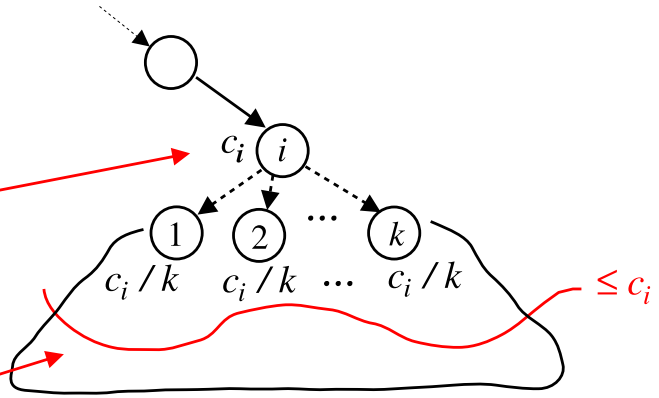
B-SVM algorithm (bottom-up)/3: Choice of H-loss coefficients c_i

B-SVM training

Cost of node i

$\geq$

Cost of subtree rooted at i



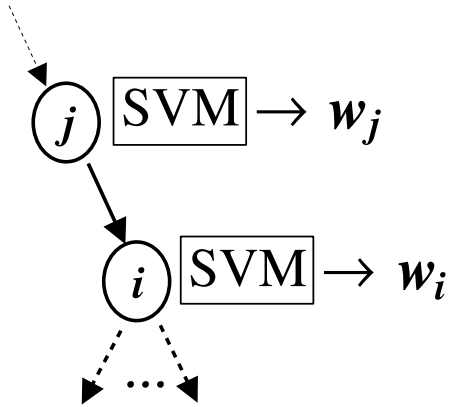
Removes bias towards short paths labelling

H-SVM algorithm (top-down) [HCC03,CGZ06], ...

SVM sitting at each node

$\boldsymbol{w}_i \leftarrow \text{SVM}(\{(\boldsymbol{x}, \boldsymbol{v}) : v_{parent(i)} = 1\})$

(**same** training as B-SVM)



Top-down label assignment

$$\hat{y}_i = \begin{cases} \{\boldsymbol{w}_i^\top \boldsymbol{x} \geq 0\} & \text{if } i \text{ is root} \\ \{\boldsymbol{w}_i^\top \boldsymbol{x} \geq 0\} & \text{if } i \text{ is } \textbf{not} \text{ root } \& \ \hat{y}_j = 1 \\ 0 & \text{otherwise} \end{cases}$$

Independent of c_i !

Experimental results/1: Datasets

- Reuters Corpus Volume 1 (RCV1), first $100,000$ docs: 101 nodes, 4 trees, height 3
  Avg # of paths/label: 1.5
  5 ordered chunks
- OHSUMED corpus of medical abstracts, subtree rooted in "Quality of Health Care": $55,503$ docs, 94 nodes, height 4
  Avg # of paths/label: 1.53
  5 random splits $40,000/15,503$
- Two synthetic datasets: $40,000$ examples, 3 complete ternary trees, 39 nodes, height 2
  Avg # of paths/label: SYNTH1: 2.66; SYNTH2: 1.28
  4 ordered chunks

Experimental results/2: Avg test error (H-loss)

Dataset	SYNTH1	SYNTH2
H-SVM	1.454 ($\pm$0.007)	0.350 ($\pm$0.007)
B-SVM	**1.269** ($\pm$0.008)	**0.322** ($\pm$0.003)

Dataset	RCV1	OHSUMED
H-SVM	0.716 ($\pm$0.024)	1.171 ($\pm$0.005)
B-SVM	**0.712** ($\pm$0.023)	**1.158** ($\pm$0.005)

Experimental results/3:
Test error (H-loss) per chunck

RCV1

CHUNK #	1	2	3	4	5
H-SVM	0.702	0.727	0.727	0.741	0.682
B-SVM	**0.701**	**0.727**	**0.712**	**0.741**	**0.681**

OHSUMED

CHUNK #	1	2	3	4	5
H-SVM	1.176	1.163	1.171	1.170	1.171
B-SVM	**1.164**	**1.152**	**1.158**	**1.157**	**1.161**

Experimental results/4: Breakdown on levels

	SYNTH1		SYNTH2	
DEPTH #	H-SVM	B-SVM	H-SVM	B-SVM
0	**0.432**	0.445	0.123	**0.121**
1	0.572	**0.519**	0.184	**0.166**
2	0.757	**0.678**	0.187	**0.183**

	RCV1		OHSUMED	
DEPTH #	H-SVM	B-SVM	H-SVM	B-SVM
0	**0.190**	0.192	**1.010**	1.018
1	**0.471**	0.474	0.316	**0.297**
2	0.131	**0.126**	0.160	**0.152**
3	0.097	**0.093**	**0.099**	0.101
4			0.413	**0.389**

Experimental results/5: Comments

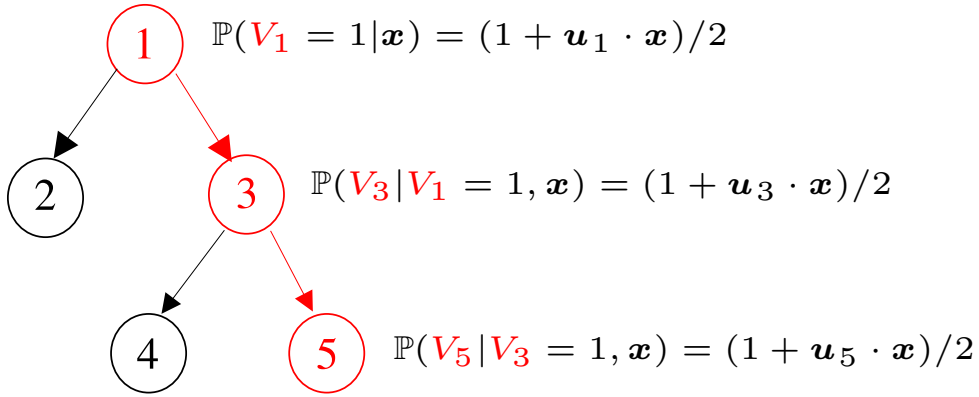
- B-SVM beats H-SVM in all cases:
  - clear on SYNTH1 and SYNTH2
  - significant on OHSUMED
  - marginal on RCV1
- Chunk-wise (RCV1, OHSUMED):
  B-SVM beats H-SVM on all chunks
- Across levels:
  B-SVM beats H-SVM at deeper levels

On-line hierarchical classification [CGZ06]/1

Parametric model

Instance $\boldsymbol{x} \in \mathbb{R}^d$: $||\boldsymbol{x}|| = 1$

$$\mathbb{P}(V_i = 1 \,|\, V_{parent(i)} = 1, \boldsymbol{x}) = (1 + \boldsymbol{u}_i \cdot \boldsymbol{x})/2 \in [0, 1]$$



Parameters $\boldsymbol{u}_i$

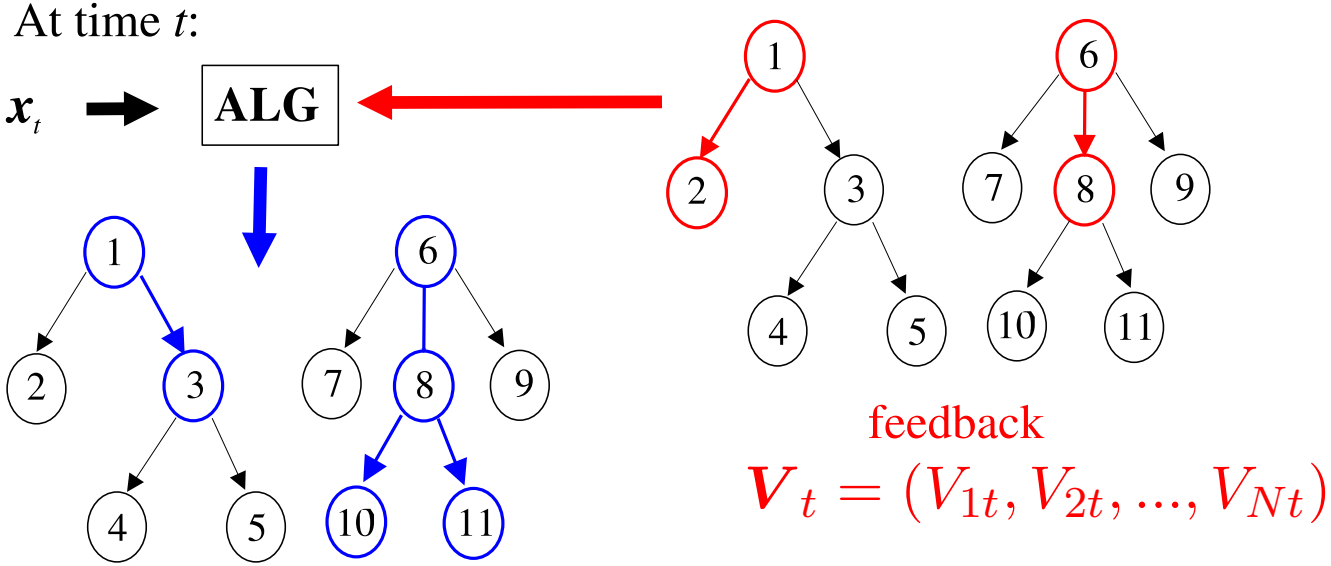
$||\boldsymbol{u}_i|| = 1, i = 1...N$

$$\mathbb{P}(V_i = 1 \,|\, V_{parent(i)} = 0, \boldsymbol{x}) = 0 \quad \forall \boldsymbol{x}$$

On-line hierarchical classification/2

On-line learning protocol

At time t:



prediction

$$\hat{\boldsymbol{Y}} = (\hat{y}_{1t}, \hat{y}_{2t}, ..., \hat{y}_{Nt}) \in \{0,1\}^N$$

feedback

$$\boldsymbol{V}_t = (V_{1t}, V_{2t}, ..., V_{Nt})$$

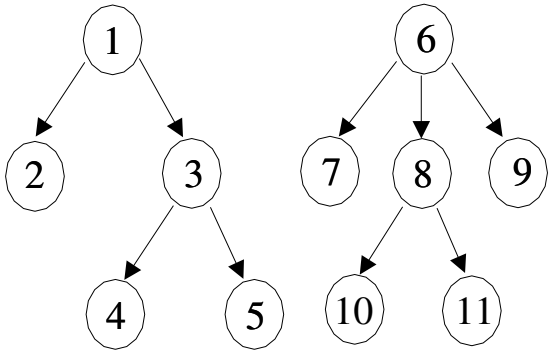
Accuracy on H-loss

On-line hierarchical classification/3

Comparison predictor and regret

Top-down comparator:

$$y_i = \begin{cases} \{\boldsymbol{u}_i \cdot \boldsymbol{x} \geq 0\} & \text{if } i \text{ is root} \\ \{\boldsymbol{u}_i \cdot \boldsymbol{x} \geq 0\} & \text{if } i \text{ is } \textbf{not} \text{ root } \& \\ & y_{parent(i)} = 1 \\ 0 & \text{otherwise} \end{cases}$$



Not Bayes optimal (independent of c_i!)

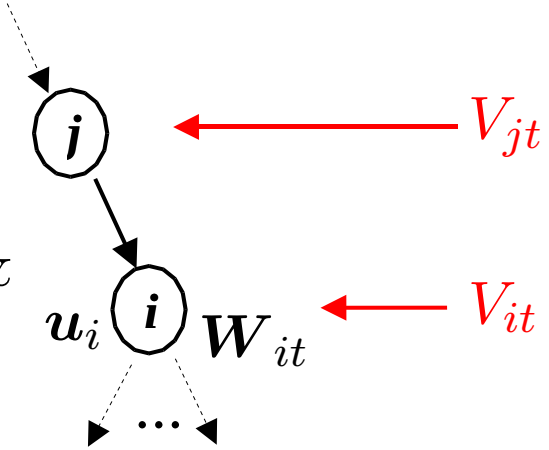
Cumulative regret:

$$\sum_t \left(\mathbb{E}[\ell_H(\hat{\boldsymbol{Y}}_t, \boldsymbol{V}_t)] - \mathbb{E}[\ell_H(\boldsymbol{Y}_t, \boldsymbol{V}_t)] \right)$$

On-line hierarchical classification/4: Algorithm/1

- **Storage:** Keeps at node i weight vector $\boldsymbol{W}_{it} \simeq \boldsymbol{u}_i$
- Prediction:

$$\hat{y}_{it} = \begin{cases} \{\boldsymbol{W}_{it} \cdot \boldsymbol{x}_t \geq 0\} & \text{if } i \text{ is root} \\ \{\boldsymbol{W}_{it} \cdot \boldsymbol{x}_t \geq 0\} & \text{if } i \text{ is } \textbf{not} \text{ root } \& \\ & \hat{y}_{jt} = 1 \\ 0 & \text{otherwise} \end{cases}$$



- Update: On feedback $V_{1t}, ..., V_{Nt}$, $\boldsymbol{x}_t$ is "passed" to node i ("$\boldsymbol{x}_t \rightarrow i$") iff $V_{jt} = 1$

Top-down and (once again) independent of c_i !

On-line hierarchical classification/4: Algorithm/2

$\boldsymbol{W}_{it}$ is regularized least squares-like [HK,AW,CBCG,RYT, ...]:

$$\boldsymbol{W}_{it} = \underbrace{\left(I + \sum_{s:\boldsymbol{x}_s \to i} \boldsymbol{x}_s \boldsymbol{x}_s^\top + \boldsymbol{x}_t \boldsymbol{x}_t^\top\right)^{-1}}_{\text{2nd-order structure}} \underbrace{\left(\sum_{s:\boldsymbol{x}_s \to i} V_{is} \boldsymbol{x}_s\right)}_{\text{(1st-order) Perc.-like}}$$

Identity

Variance control

- Naturally arises from parametric model:
  $\boldsymbol{W}_{it} \cdot \boldsymbol{x}_t$ (almost) conditionally unbiased estimator of $\boldsymbol{u} \cdot \boldsymbol{x}_t$
  accuracy depends on $\#\{s : \boldsymbol{x}_s \to i\}$
- Dual form (kernels): running time quadratic in
  $\#\{s \leq t-1 : \boldsymbol{x}_s \to i\}$

On-line hierarchical classification/5

Regret bound (pointwise)

Example sequence $(\boldsymbol{x}_1, \boldsymbol{V}_1) \ldots (\boldsymbol{x}_T, \boldsymbol{V}_T) \in \mathbb{R}^d \times \{0,1\}$

$$\sum_{\text{examples } t} \Big(\mathbb{E}\big[\ell_H(\hat{\boldsymbol{Y}}_t, \boldsymbol{V}_t)\big] - \mathbb{E}\big[\ell_H(\boldsymbol{Y}_t, \boldsymbol{V}_t)\big] \Big)$$

$$\leq \frac{1}{\Delta^2} \sum_{\text{nodes } i} C_i \, \mathbb{E}\left[\sum_{\text{dim. } j} \log(1 + \lambda_{ij}) \right]$$

$\min_{i,t} (\boldsymbol{u}_i \cdot \boldsymbol{x}_t)^2$ (margin)

$\sum_{k \in subtree(i)} c_k$ ("value" of subtree)

j-th eigenval of $\sum_{s : \boldsymbol{x}_s \to i} \boldsymbol{x}_s \boldsymbol{x}_s^\top$

Conclusions

- Framework for hierarchical classification
- Improving baseline top-down evaluation (label assignment) scheme (H-SVM) by bottom-up Bayes-optimal-like
- Modular approach
- Preliminary experiments
- On-line (top-down) regret analysis of regularized least-squares alg on linear parametric model

Open questions

- Experimental:
  - Clear advantage of bottom-up on synthetic,
    less clear on real-world:
    "human noise" ? Naive independence assumption ?
  - Replace SVM by better algs for probability estimation
    (e.g. regularized logistic regression)
- Theoretical:
  - Regret analysis w.r.t. Bayes optimal
    logistic model ?
  - Remove independence assumption