Multilabel Classification and Ranking with Partial Feedback

Claudio Gentile
Universita' dell'Insubria

26th EURO-INFORMS
Rome, July 1st, 2013

Joint with: **F. Orabona** (TTI - Chicago)

Goals:

- Design and analysis of online prediction algorithms for multilabel and ranking problems under partial information
  - Motivation and model
  - Derived algorithms
  - Regret analysis
- Experimental investigation on real data
- General goal: Principled machine learning methods working well in pratice

Multilabel and Ranking with Partial Information

- Only feedback on suggested items
- Some suggestions more important than others
- No ranking info from feedback

Small U.S. Farms Find Profit in Tourism
By WILLIAM NEUMAN
To survive, small farmers in America are increasingly turning to nonfarm activities, like operating bed-and-breakfasts.

The New York Times
Business Day

U.S. Is Falling Behind in the Business of 'Green'
By ELISABETH ROSENTHAL
Strong incentives in European and Asian countries have given them the lead in clean energy technologies.

The New York Times
Business Day

Region in Revolt
Egypt's Economy Slows to a Crawl; Revolt Is Tested
By DAVID D. KIRKPATRICK and DINA SALAH AMER

The New York Times
Business Day

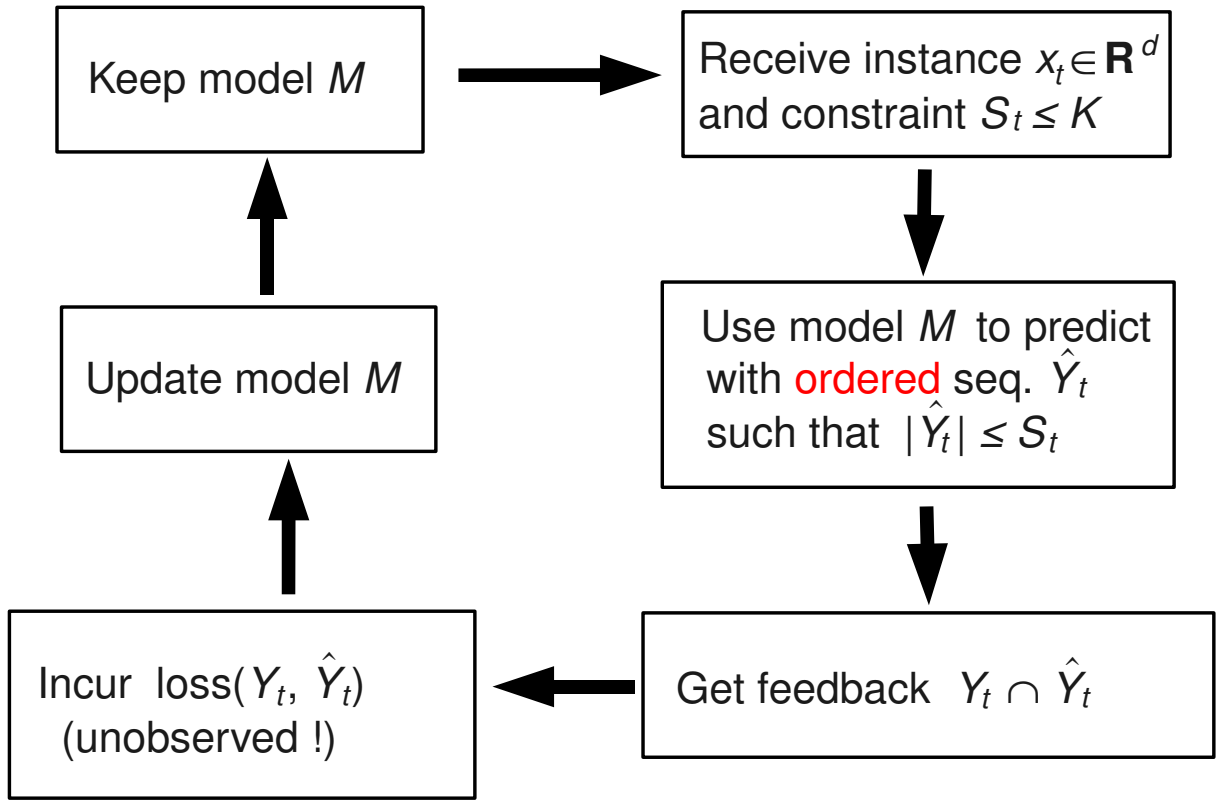
I.B.M. Researchers Create High-Speed Graphene Circuits
By JOHN MARKOFF
The advance, reported in the journal Science, may have applications that include future smartphone and telephone displays.

GLOBAL EDITION
Science

Learning Model/1 [SGK09,KRS10,SJ12,DWCH12,...]

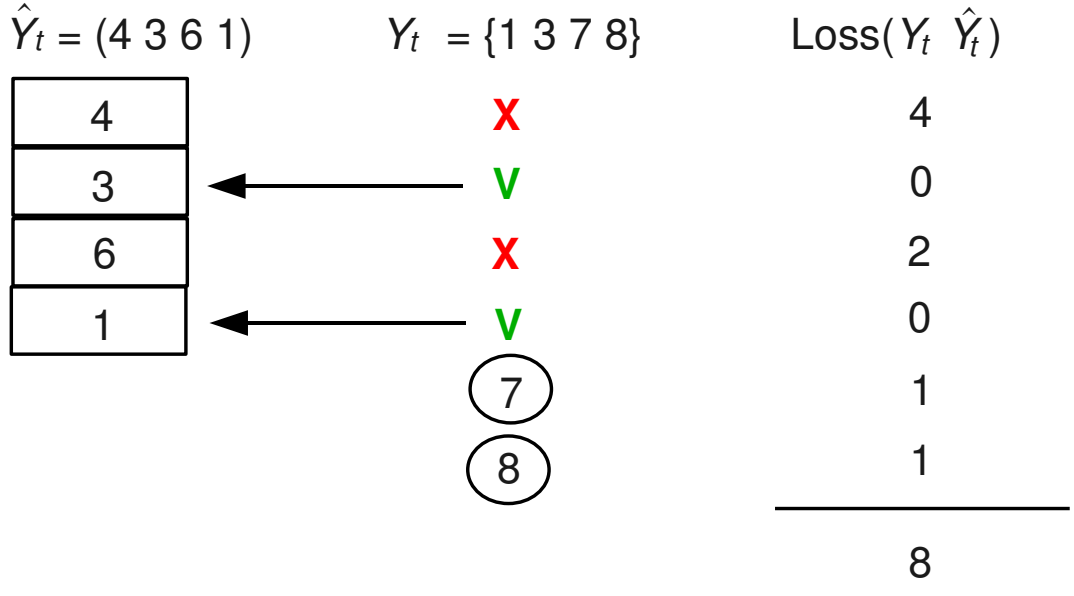
Observe training set $(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_t, Y_t) \ldots, \in \mathbf{R}^d \times 2^{[K]}$

in a sequential and partial manner (Y_t unordered subset of $[K]$)



Learning Model/2

K = 10 classes, S_t = 4



$$\text{Loss}(Y_t, \hat{Y}_t) = \sum_{i \in \hat{Y}_t \setminus Y_t} c(j_i) + |Y_t \setminus \hat{Y}_t|$$

Difference of DCGs restricted to $\hat{Y}_t$ — false negatives

Ranking loss(Y_t, $\hat{Y}_t$) = 4 + 2 = 6

[Y_t is unsorted] — # of flipped pairs in $\hat{Y}_t$ (1-4), (1-6), (3-4), (3-6) — false negatives

Learning Model/3

- Instance vectors $\boldsymbol{x}_t \in \mathbf{R}^d$ and constraints $S_t \leq K$ from adaptive adversary, $||\boldsymbol{x}_t|| = 1$
- Labels $Y_t = \{y_{1t}, \ldots, y_{Kt}\} \subseteq [K]$ from (generalized) linear model

$$p_{it} = \mathbf{P}(y_{it} = 1 \,|\, \boldsymbol{x}_t) = p(\boldsymbol{u}_i^\top \boldsymbol{x}_t) \qquad i = 1 \ldots K$$

- Bayes optimum $b(\boldsymbol{x}_t)$
  - Loss: sort classes $p_{i_1,t} \geq p_{i_2,t} \geq \ldots \geq p_{i_K,t}$ and decide on cutoff $s \leq S_t \implies b(\boldsymbol{x}_t) = (i_1, i_2, \ldots, i_s)$
  - Ranking loss: Same as above, but cutoff $s = S_t$
- Cumulative regret

$$R_T = \sum_{t=1}^{T} E_{Y_t}[\text{“}loss\text{”}(Y_t, \widehat{Y}_t) \,|\, \ldots] - E_{Y_t}[\text{“}loss\text{”}(Y_t, b(\boldsymbol{x}_t)) \,|\, \ldots]$$

Algorithm/1

Keep proxy $W_t = (\boldsymbol{w}_{1t}, \ldots, \boldsymbol{w}_{Kt}) \in \mathbf{R}^{dK}$

Init: $W_0 = 0$

For $t = 1, 2, \ldots$:

- Sort classes i based on proxy

Upper confidence exploration/exploitation

$$\widehat{p}_{it} = p(\underbrace{\boldsymbol{w}_{it}^\top \boldsymbol{x}_t}_{\text{current approx.}} + \underbrace{\epsilon_{it}}_{\text{uncertainty on } i})$$

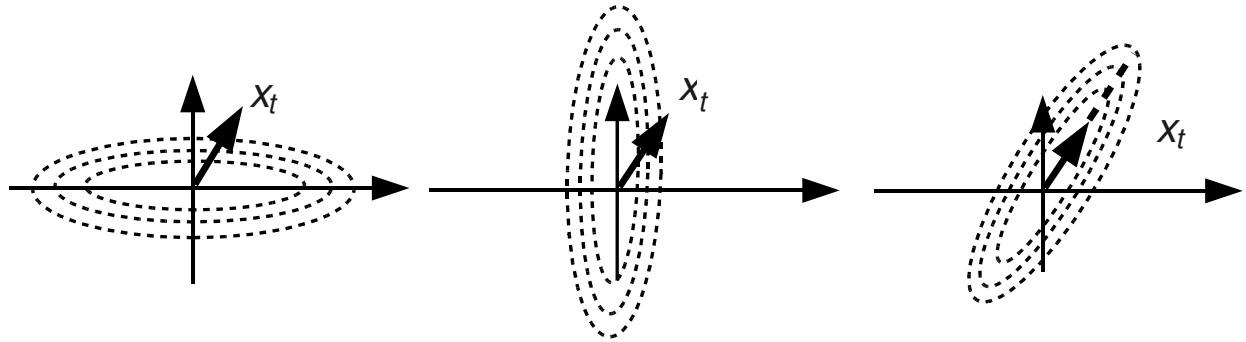
  and build $\widehat{Y}_t$: $|\widehat{Y}_t| \leq S_t$

- Get feedback $Y_t \cap \widehat{Y}_t$
- Promote classes $i \in Y_t \cap \widehat{Y}_t$ and demote classes $i \in \widehat{Y}_t \setminus Y_t$
- Update $\underbrace{W_{t-1} \to W_t}$

  Second-order descent (matrices $A_{i,t}$ + curvature of link $p(\cdot)$)

Algorithm/2: Remarks

- Ellipsoidal uncertainty $\epsilon_{i,t}$:
  Small if so far observed many $\boldsymbol{x}_s$, $s < t$, aligned with $\boldsymbol{x}_t$: $i \in \widehat{Y}_s$



- Running time per prediction: $O(d^2 + K \log K)$
- Also in dual variables: $O(t^2 + K \log K)$ per prediction at time t
- Variant with diagonal matrices $A_{i,t}$: quadratic $\rightarrow$ linear
- Feedback $Y_t \cap \widehat{Y}_t$ allows to estimate of gradient of surrogate loss associated with (generalized) linear model $p(\cdot)$

Regret Analysis

If $S_t \leq S \ \forall t$, with prob. $> 1 - \delta$

$$R_T = O\left(\sqrt{SK}\sqrt{T}\log(T/\delta)\right)$$

Remark:
In multilabel/ranking problems S usually much smaller than K

Existing literature: [SGK09,KRS10,SJ12,...]
Several results, even more general (e.g. sub-modular loss + arbitrary Y_t)
but:

- Either larger regret $O(T^{2/3})$
- or no side-info $\boldsymbol{x}_t$
- or purely theoretical algs.
- or different partial feedback models

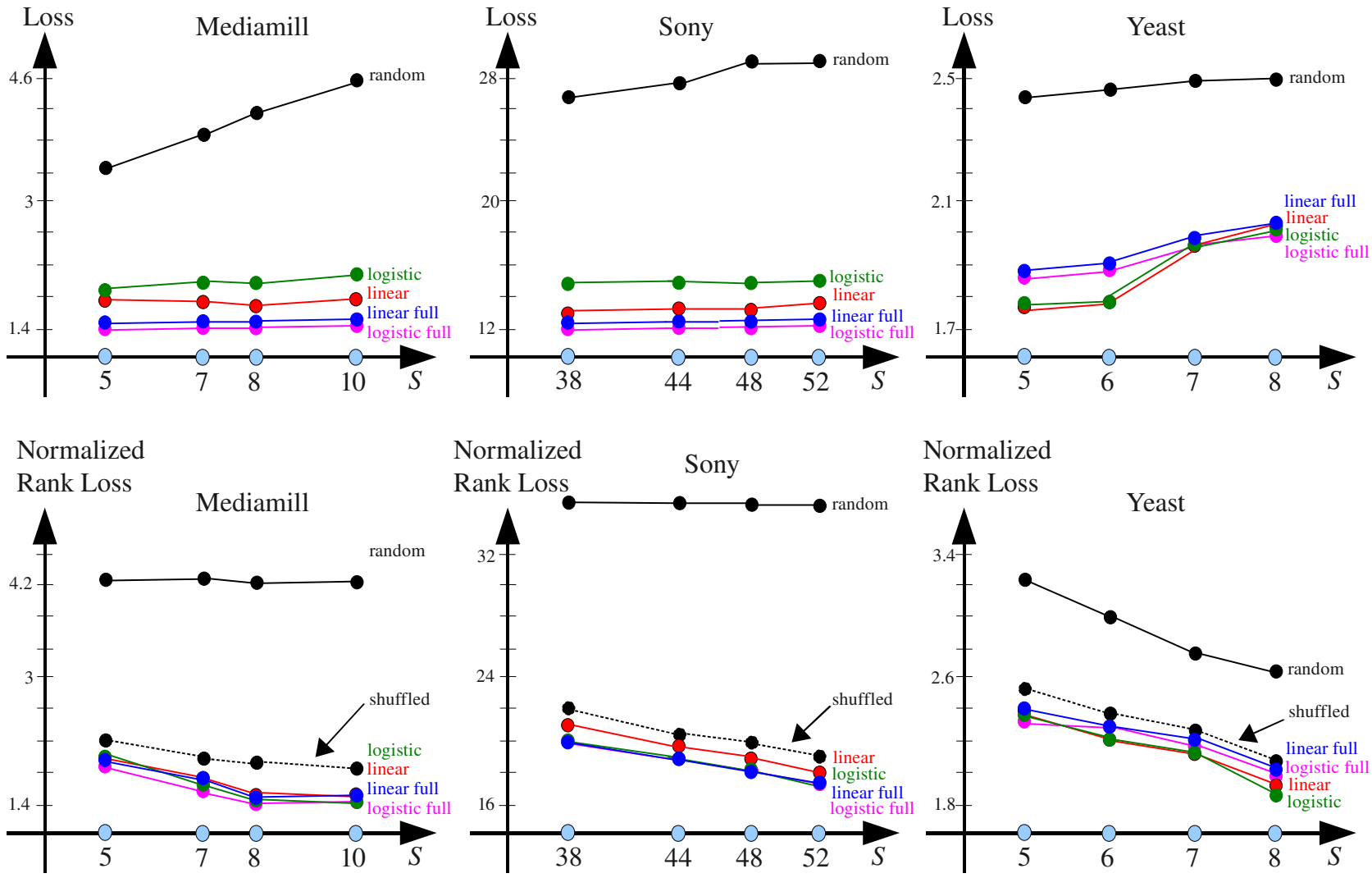
Esperiments/1: Datasets and Setting

- Three multilabel datasets (diverse no. of labels and label density):

Task	T	d	K	Avg	Avg+stdv	95%	99%
Mediamill	43907	120	101	5	7	8	10
Sony	32971	98	632	38	44	48	52
Yeast	2417	103	14	5	6	7	8

- Online and batch results
- Train on single epoch
- Different values of S in "$S_t \leq S$"
- Two $p(\cdot)$ models: linear and logistic
- Baselines: Online Binary Relevance (OBR) with full info
- Both Loss and Ranking loss

Experiments/2: Results



Experiments/3: Conclusions

- Partial info algs work well, especially when S is set in informed way ...
- ... almost as good as full info counterparts (sometimes even slightly better...)
- Linear vs. logistic not clear
- Ranking based on $p_{i,t}$ seems effective (check "shuffled" results)