Multiclass Multilabel Classification with More Classes than Examples

Ohad Shamir

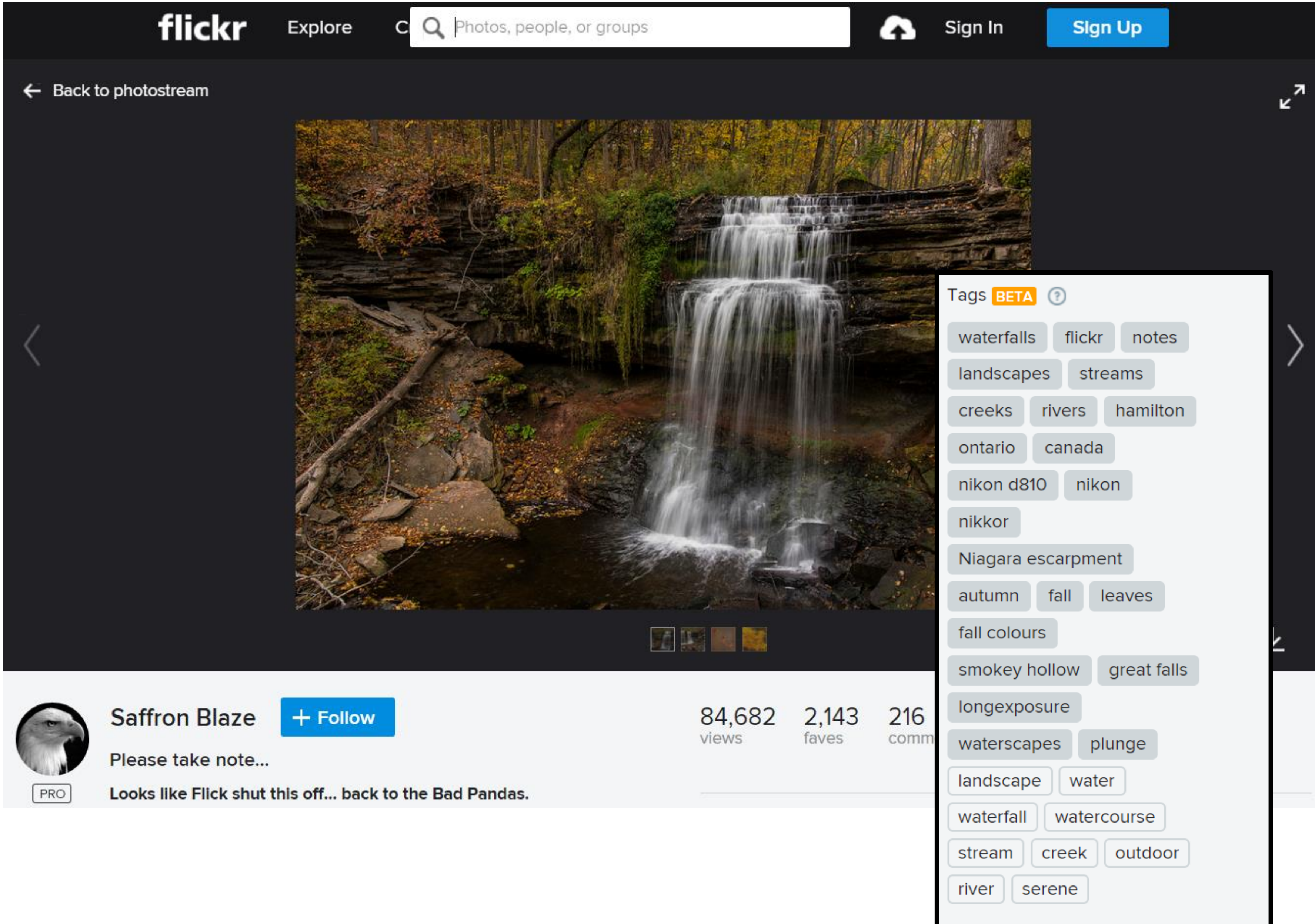
Weizmann Institute of Science

Joint work with Ofer Dekel, MSR

NIPS 2015 Extreme Classification Workshop

Extreme Multiclass Multilabel Problems

Label set is a folksonomy (a.k.a. collaborative tagging or social tagging)


flickr
Explore
Photos, people, or groups
Sign In
Sign Up
Back to photostream
Tags BETA
waterfalls
flickr
notes
landscapes
streams
creeks
rivers
hamilton
ontario
canada
nikon d810
nikon
nikkor
Niagara escarpment
autumn
fall
leaves
fall colours
smokey hollow
great falls
longexposure
waterscapes
plunge
landscape
water
waterfall
watercourse
stream
creek
outdoor
river
serene
Saffron Blaze
+ Follow
PRO
Please take note...
Looks like Flick shut this off... back to the Bad Pandas.
84,682
views
2,143
faves
216

Create account Not logged in Talk Contributions Log in

Article Talk Read View source View history Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Perm
Page

Leonardo da Vinci

From Wikipedia, the free encyclopedia

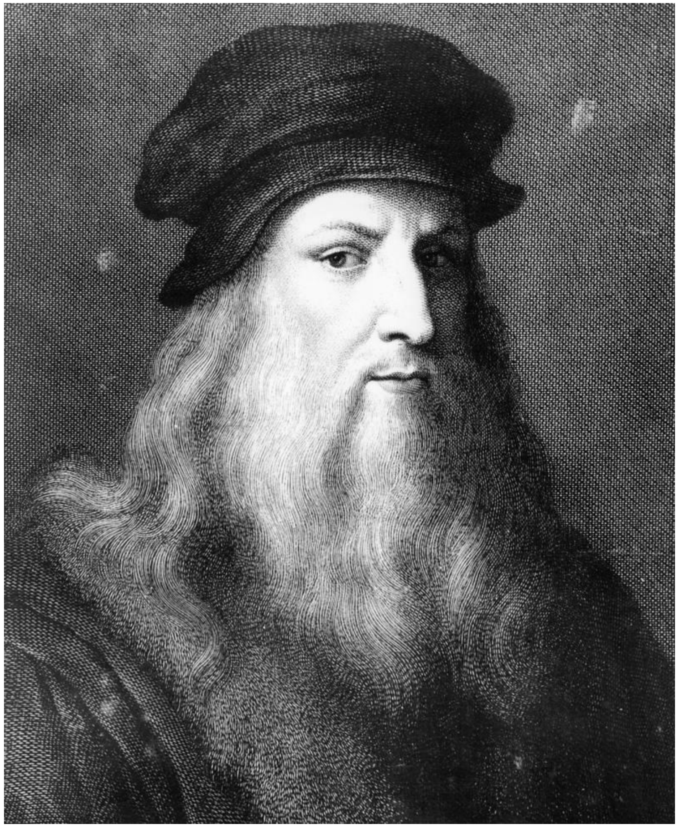
"Da Vinci" redirects here. For other uses, see Da Vinci (disambiguation).
This is a Renaissance Florentine name. The name daVinci is an indicator of birthplace, not a family name; this person is properly referred to by the given name Leonardo.

Leonardo di ser Piero da Vinci, more commonly **Leonardo da Vinci**, (Italian: [leoˈnardo da (v)ˈvintʃi] (listen); 15 April 1452 – 2 May 1519) was an Italian polymath whose areas of interest included invention, painting, sculpting, architecture, science, music, mathematics, engineering, literature, anatomy, geology, astronomy, botany, writing, history, and cartography. He has been variously called the father of paleontology, ichnology, and architecture, and is widely considered one of the greatest painters of all time.[1] Sometimes credited with the inventions of the parachute, helicopter and tank,[2][3][4] his genius epitomized the Renaissance humanist ideal.

Many historians and scholars regard Leonardo as the prime exemplar of the "Universal Genius" or "Renaissance Man", an individual of "unquenchable curiosity" and "feverishly inventive imagination".[5]

Leonardo da Vinci

Categories: Leonardo da Vinci | 1452 births | 1519 deaths | 15th century in science | 15th-century scientists | 16th century in science | 16th-century scientists | Age of Enlightenment | Ambassadors of the Republic of Florence | Ballistics experts | Fabulists | Giftedness | History of anatomy | Italian anatomists | Italian civil engineers | Italian inventors | Italian military engineers | Italian physiologists | Italian Renaissance humanists | Mathematical artists | Mathematics and culture | Members of the Guild of Saint Luke | People from the Province of Florence | People prosecuted under anti-homosexuality laws | Physiognomists | Renaissance architects | Renaissance artists | Renaissance painters | Renaissance scientists | Tuscan painters



Categories

1452 births / 1519 deaths / 15th century in science / ambassadors of the republic of Florence / Ballistic experts / Fabulists / giftedness / mathematics and culture / Italian inventors / Members of the Guild of Saint Luke / Tuscan painters / people persecuted under anti-homosexuality laws...

Problem Definition

- Multiclass multilabel classification
- m training examples, k categories
- $m, k \to \infty$ together
  - Possibly even $k > m$
- **Goal:** Categorize unseen instances

Extreme Multiclass

- Supervised learning starts with binary classification ($k=2$) and extends to multiclass learning
  - Theory: VC dimension → Natarajan dimension
  - Algorithms: binary → multiclass
- Usually, assume $k = \mathcal{O}(1)$
- Some exceptions
  - Hierarchy with prior knowledge on relationships – not always available
  - Additional assumptions (e.g. talk by Marius earlier)

Application

- Classify the web based on Wikipedia categories
- Training set: All Wikipedia pages ($m = 4.2 \times 10^6$)
- Labels: All Wikipedia categories ($k = 1.1 \times 10^6$)

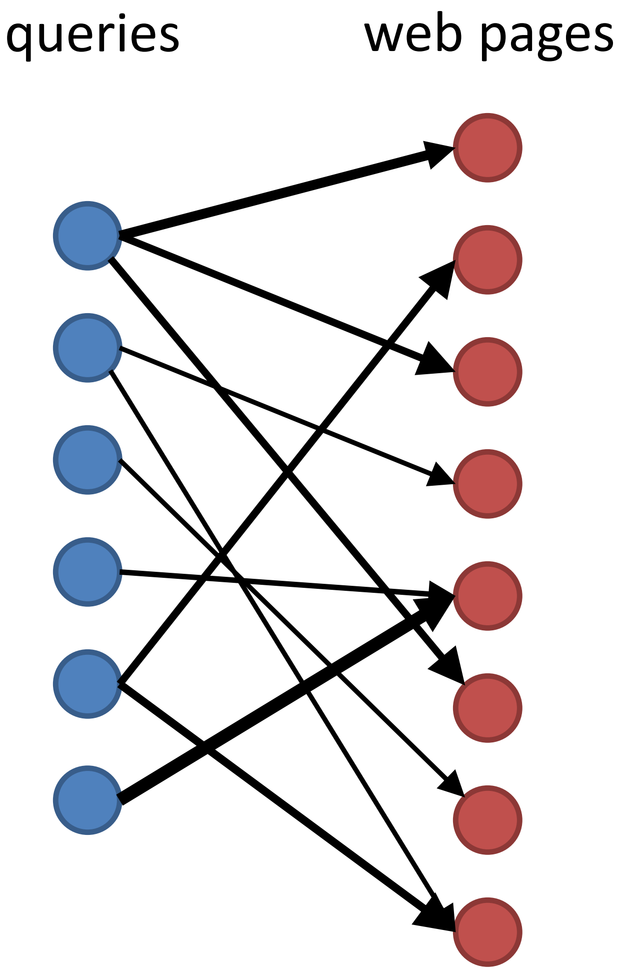
WIKIPEDIA
The Free Encyclopedia

Challenges

- Statistical problem: Can’t get a large (or even moderate) sample from each class.

- Computational problem: Many classification algorithms will choke on millions of labels

Propagating Labels on the Click-Graph



- A bipartite graph derived from search engine logs: clicks encoded as weighted edges
- Wikipedia pages are labeled web pages
- Labels propagate along edges to other pages

Example

- http://en.wikipedia.org/wiki/Leonardo da Vinci passes multiple labels to http://www.greatItalians.com

- Among them
  - “Renaissance artists” – good
  - “1452 births” – bad

- Observation: “1452 births” induces many false-positives (FP): best to remove it altogether from classifier output
  - (FP $\Rightarrow$ TN, TP $\Rightarrow$ FN)

Simple Label Pruning Approach

1. Split dataset to training and validation set
2. Use training set to build an initial classifier h_{pre} (e.g. by propagating labels over click-graph)
3. Apply h_{pre} to validation set, count FP and TP
4. $\forall j \in \{1, \dots, k\}$, remove label j if

$$\frac{FP_j}{TP_j} > \frac{1-\gamma}{\gamma}$$

- Defines a new "pruned" classifier h_{post}

Simple Label Pruning Approach

Explicitly minimizes empirical risk
with respect to the γ-weighted loss:

$$\ell(h(\boldsymbol{x}), \boldsymbol{y}) =$$

$$\sum_{j=1}^{k} \Big[\underbrace{\gamma\, \mathbb{I}\big(h_j(\boldsymbol{x}) = 1, y_j = 0\big)}_{\substack{\text{FP} \\ \text{(false positive)}}} + \underbrace{(1-\gamma)\, \mathbb{I}\big(h_j(\boldsymbol{x}) = 0, y_j = 1\big)}_{\substack{\text{FN} \\ \text{(false negative)}}}\Big]$$

Main Question

Would this actually reduce the risk?

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})}\left[\ell\left(h_{post}(\boldsymbol{x}),\boldsymbol{y}\right)\right] \ < \ \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})}\left[\ell\left(h_{pre}(\boldsymbol{x}),\boldsymbol{y}\right)\right]$$ - positive

Baseline Approach

- Prove that uniformly for all labels j

$$\frac{\widehat{FP}_j}{\widehat{TP}_j} \longrightarrow \frac{FP_j}{TP_j}$$

Pr(label j and not predicted)

Pr(label j and predicted)

Problem: $m, k \to \infty$ together. Many classes only have a handful of examples

Uniform Convergence Approach

- Algorithm implicitly chooses a hypothesis from a certain hypothesis class
  - Pruning rules on top of fixed predictor h_{pre}
- Prove uniform convergence by bounding VC dimension / Rademacher complexity
- Conclude that if empirical risk decreases, the risk decreases as well

Uniform Convergence Fails

- Unfortunately, no uniform convergence...
- ... and even no algorithm/data-dependent convergence!

$$\mathbb{E}\left[R\left(h_{post}\right) - \hat{R}\left(h_{post}\right)\right] \geq$$

$$\sum_{j=1}^{k} Pr(j \text{ pruned}) \left(TP_j - FP_j\right)$$

$$= \sum_{j=1}^{k} Pr\left(\widehat{FP}_j > \widehat{TP}_j\right) \left(TP_j - FP_j\right)$$

Weak correlation in $m \approx k$ regime

A Less Obvious Approach

- Prove directly that risk decreases
- Important (but mild) assumption: Each example labeled by $\leq s$ labels
- Step 1: Risk of h_{post} is concentrated. For all ϵ,

$$\Pr\left(\left| R\left(h_{post}\right) - \mathbb{E}R\left(h_{post}\right) \right| \right.$$

A Less Obvious Approach

- Part 2: Enough to prove $R(h_{pre}) - \mathbb{E}R(h_{post}) > 0$
- Assuming for $\gamma = \frac{1}{2}$ for simplicity, can be shown that

$$R(h_{pre}) - \mathbb{E}R(h_{post})$$

$$> \text{pos} - \mathcal{O}\left(\sqrt{\frac{\left\|\left(FP_j + TP_j\right)_j\right\|_{1/2}}{m}}\right)$$

where $\|\boldsymbol{w}\|_{1/2} = \left(\sum_j \sqrt{w_j}\right)^2$

A Less Obvious Approach

$$\sum_{j:FP_j \geq TP_j} \left(FP_j - TP_j\right)$$

- For probability vector, **always at most k**
- **Smaller** the more non-uniform is the distribution

rove R

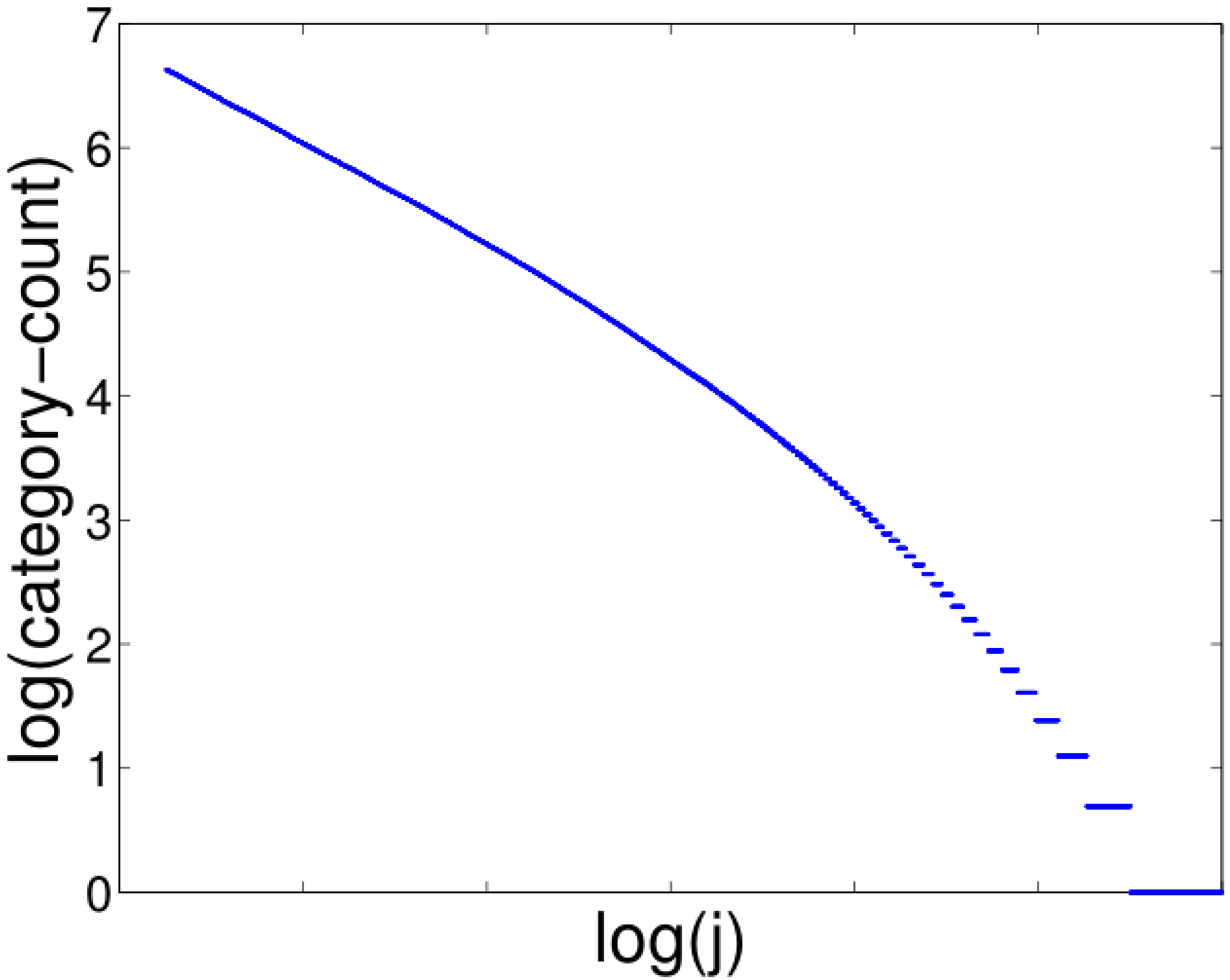
$\frac{1}{2}$ for sim

$$R\left(h_{pre}\right) - \mathbb{E}R\left(h_{post}\right)$$

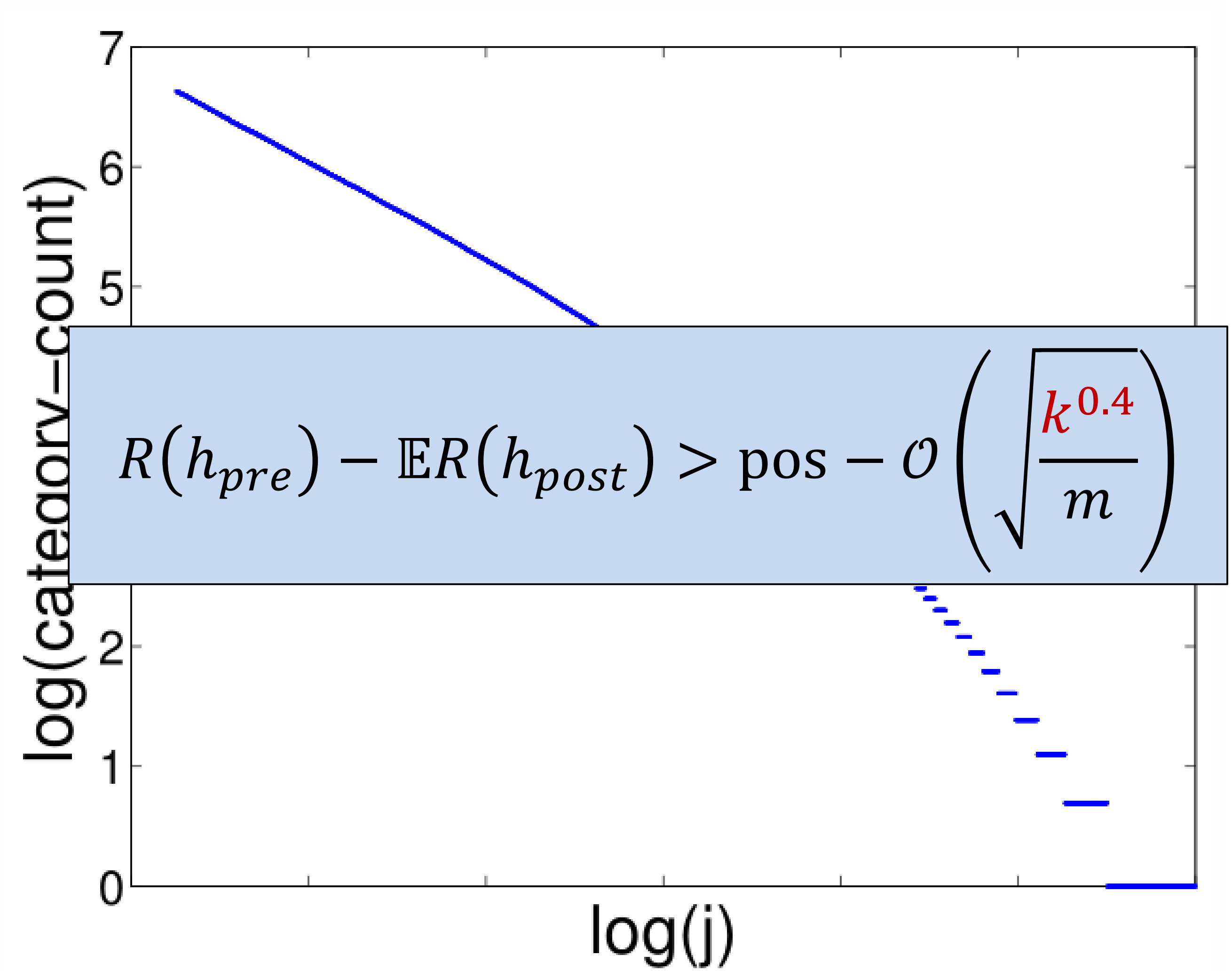
$$> \mathrm{pos} - \mathcal{O}\left(\sqrt{\frac{\left\|\left(FP_j + TP_j\right)_j\right\|_{1/2}}{m}}\right)$$

where $\|\boldsymbol{w}\|_{1/2} = \left(\sum_j \sqrt{w_j}\right)^2$

Wikipedia Power-Law: $r = 1.6$



Wikipedia Power-Law: $r = 1.6$



$$R\left(h_{pre}\right) - \mathbb{E}R\left(h_{post}\right) > \text{pos} - \mathcal{O}\left(\sqrt{\frac{k^{0.4}}{m}}\right)$$

Experiment

Click graph on the entire web
(based on search engine logs)

Experiment

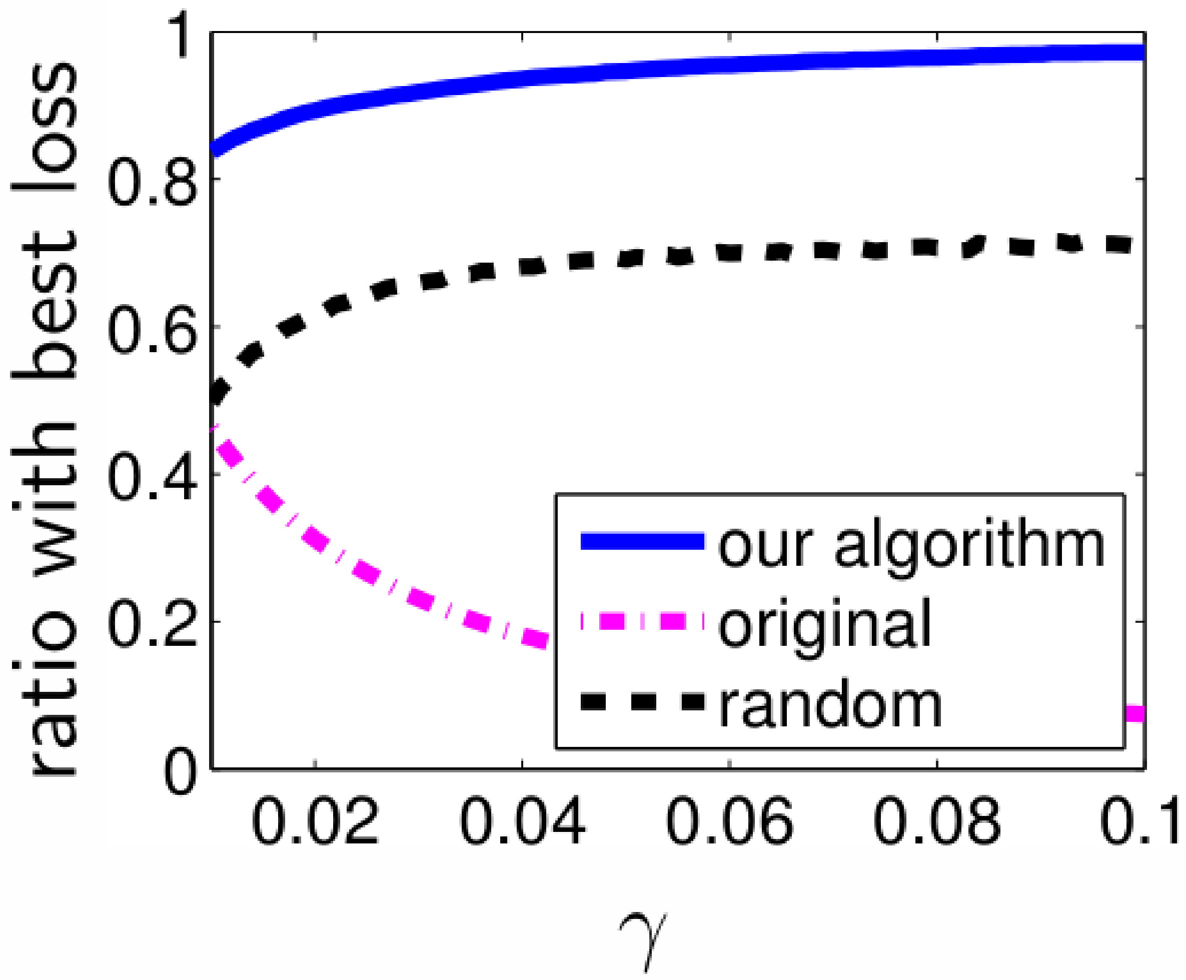
Categories from Wikipedia pages
propagated twice through graph

Experiment

Train/test split of Wikipedia pages

How good are propagated categories from training set in predicting categories at test set pages?

Experiment


ratio with best loss
1
0.8
0.6
0.4
0.2
0
0.02
0.04
0.06
0.08
0.1
γ
our algorithm
original
random

Another less obvious approach

$$R\left(h_{pre}\right) - \mathbb{E}R\left(h_{post}\right)$$

$$= \sum_{j=1}^{k} Pr(j \text{ pruned}) \left(FP_j - TP_j\right)$$

$$= \sum_{j=1}^{k} Pr\left(\widehat{FP}_j > \widehat{TP}_j\right) \left(FP_j - TP_j\right)$$

Weak but positive correlation,
even if only few examples per label

For **large k**, sum will tend to be positive

Different Application: Crowdsourcing

(Dekel and S., 2009)



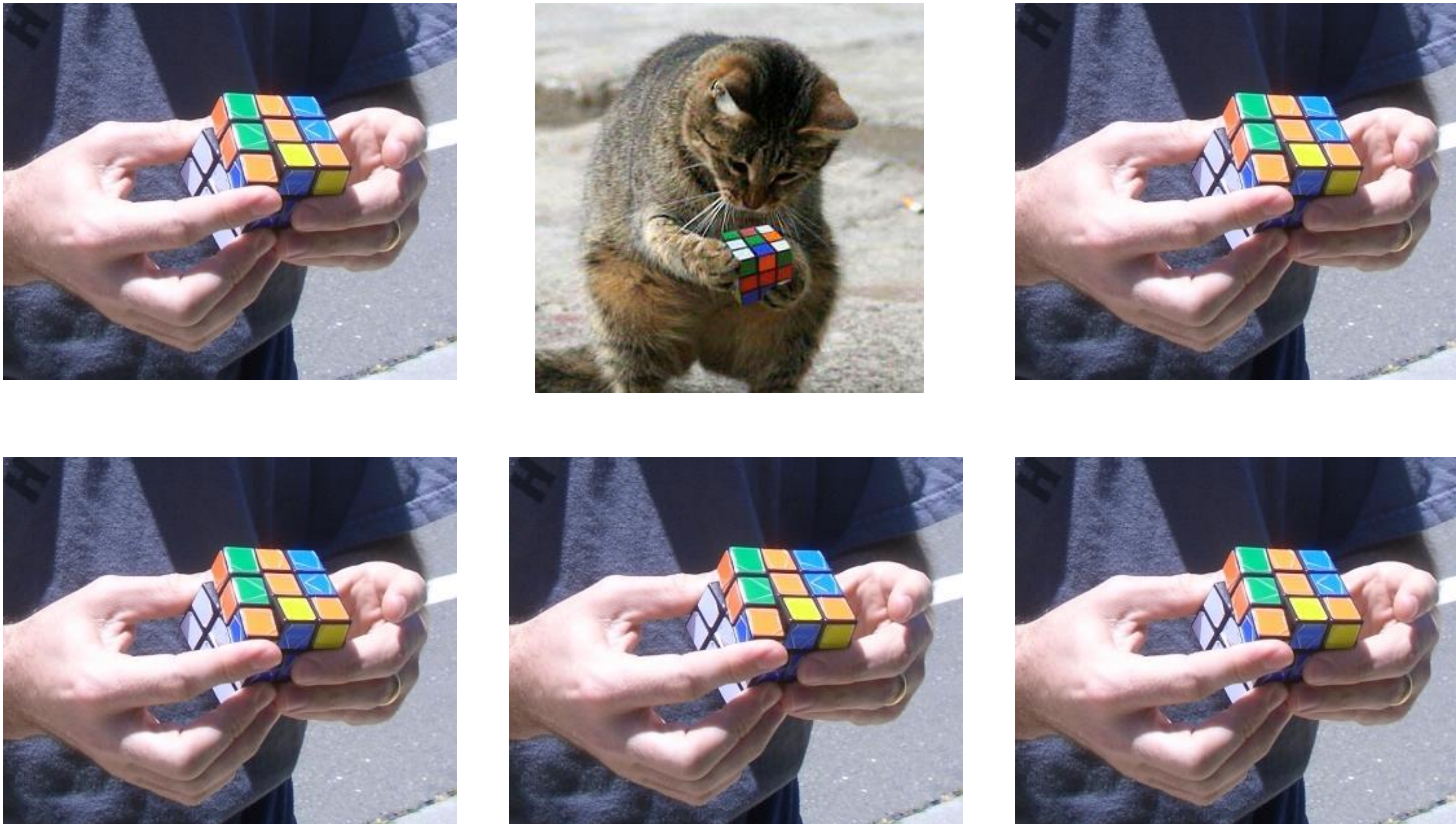



Different Application: Crowdsourcing

(Dekel and S., 2009)

Different Application: Crowdsourcing

(Dekel and S., 2009)



Different Application: Crowdsourcing

(Dekel and S., 2009)

Different Application: Crowdsourcing

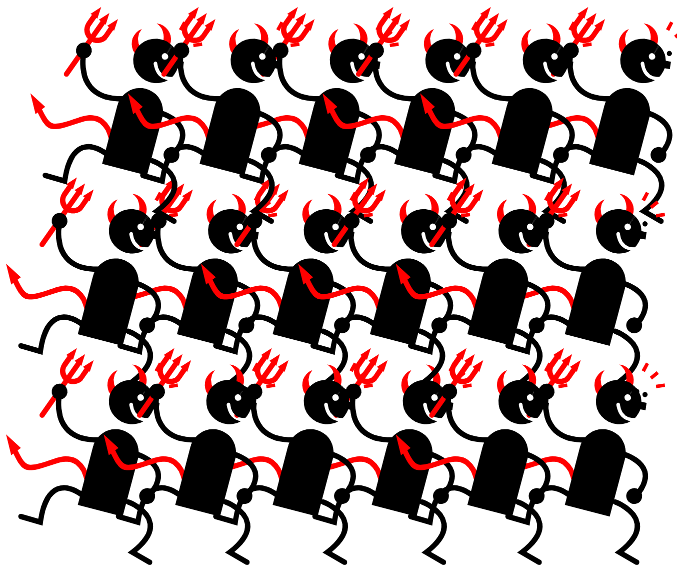
- How can we improve crowdsourced data?
- Standard approach: Repeated labeling, but expensive
- A bootstrap approach:
  - Learn predictor from data of all workers
  - Throw away examples labeled by workers disagreeing a lot with the predictor
  - Re-train on remaining examples
- Works! (Under certain assumptions)
- Challenge: Workers often labels only a handful of examples

Different Application: Crowdsourcing

examples/worker might be small, but many workers...

Different Application: Crowdsourcing

examples/worker might be small, but many workers...



Different Application: Crowdsourcing

examples/worker might be small, but many workers...

Conclusions

- $\#\ \text{classes} \rightarrow \infty$ violates assumptions of most multiclass analyses
  - Often based on generalizations of binary classification
- Possible approach
  - Avoid standard analysis
  - "Extreme X" can be a blessing rather than a curse
- Other applications? More complex learning algorithms (e.g. substitution)?

Thanks!


blue
leave
pole
delirious
moment
sing
hine
languid
would
whisper
the
worship
time
size
may
spring
stare
through
like
those
see
behind
picture
your
an