Department of Computer Science
THE UNIVERSITY OF TEXAS AT AUSTIN

THE UNIVERSITY OF TEXAS
SCHOOL OF
information

Modeling Annotator Accuracies for Supervised Learning

Abhimanu Kumar
Department of Computer Science
University of Texas at Austin
abhimanu@cs.utexas.edu
http://abhimanukumar.com

Matthew Lease
School of Information
University of Texas at Austin
ml@ischool.utexas.edu

Supervised learning from noisy labels

- Labeling is inherently uncertain
  - Even experts disagree and make mistakes
  - Crowd tends to be noisier with higher variance
- Use *wisdom of crowds* to reduce uncertainty
  - Multi-label + aggregation = consensus labels
- How to maximize learning rate (labeling effort)?
  - Label a new example?
  - Get another label for an already-labeled example?
- See: Sheng, Provost & Ipeirotis, KDD'08

Task Setup

- Task: Binary classification
- Learner: C4.5 decision tree
- Given
  - An initial seed set of single-labeled examples (64)
  - An unlimited pool of unlabeled examples
- Cost model
  - Fixed unit cost for labeling any example
  - Unlabeled examples are freely obtained
- Goal: Maximize learning rate (for labeling effort)

Compare 3 methods: SL, MV, & NB

- **Single labeling (SL):** label a new example
- Multi-Labeling: get another label for pool
  - **Majority Vote (MV)**: consensus by simple vote
  - **Naïve Bayes (NB)**: weight vote by annotator accuracy

$$\begin{aligned}
\widehat{x} &= \operatorname*{argmax}_{x} P(X^j = x | Y^j_{1:w}) \\
&\propto P(Y^j_{1:w} | X^j) P(X^j) \\
&= \prod_{i=1}^{w} P(Y^j_i | X^j) P(X^j)
\end{aligned}$$

Assumptions

- Example selection: random
  - From pool for SL, from seed set for multi-labeling
  - No selection based on active learning
- Fixed commitment to a single method *a priori*
  - No switching between methods at run-time
- Balanced classes
  - model & measure simple accuracy (not P/R, ROC)
  - Assume uniform class prior for NB
- Annotator accuracies are known to system
  - In practice, must estimate these: from gold data (Snow et al. '08) or EM (Dawid & Skene'79)

Simulation

- Each annotator
  - Has parameter *p* (prob. of producing correct label)
  - Generates exactly one label
- Uniform distribution of accuracies U(min,max)
- Generative model for simulation
  - Pick an example *x* (with true label *y**) at random
  - Draw annotator accuracy *p* ~ U(min,max)
  - Generate label *y* ~ P(*y* | *p, y**)

Evaluation

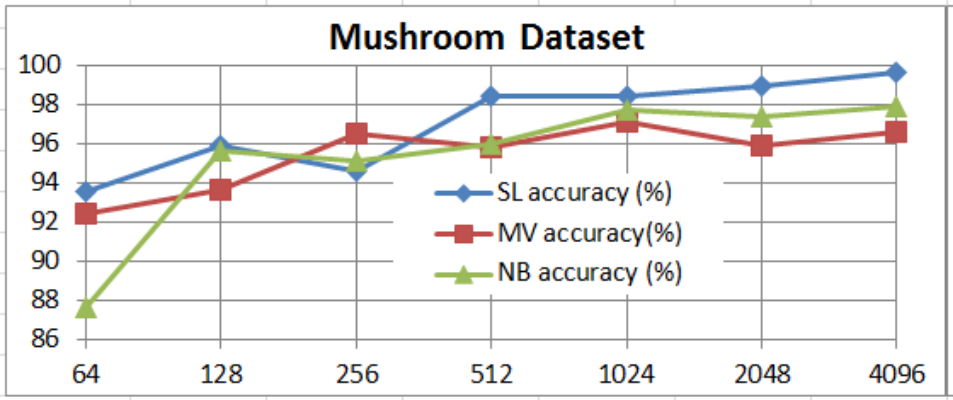
- Data: 4 datasets from UCI ML Repository
  - Mushroom
  - Spambase
  - Tic-Tac-Toe
  - Chess: King-Rook vs. King-Pawn

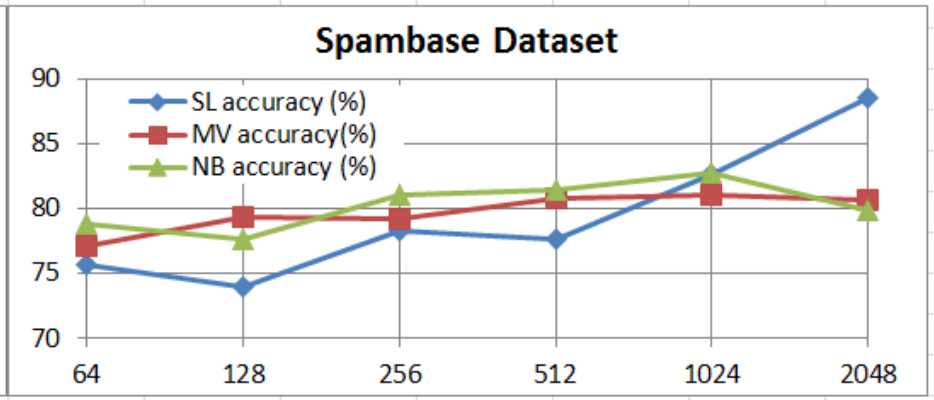
http://archive.ics.uci.edu/ml/datasets.html

- Same trends across all 4, so we report first 2
- Random 70 / 30 split of data for seed+pool / test
- Repeat each run 10 times and average results

p ~ U(0.6, 1.0)

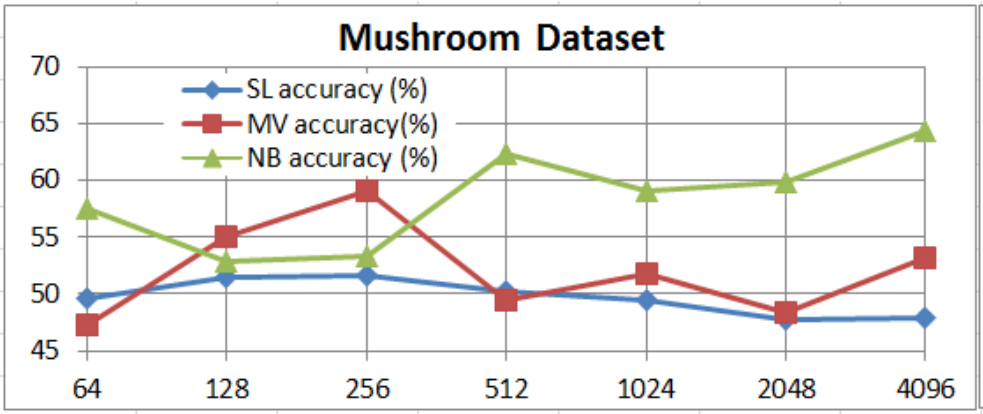
- Fairly accurate annotators (mean = 0.8)
- Little uncertainty -> little gain from multi-labeling

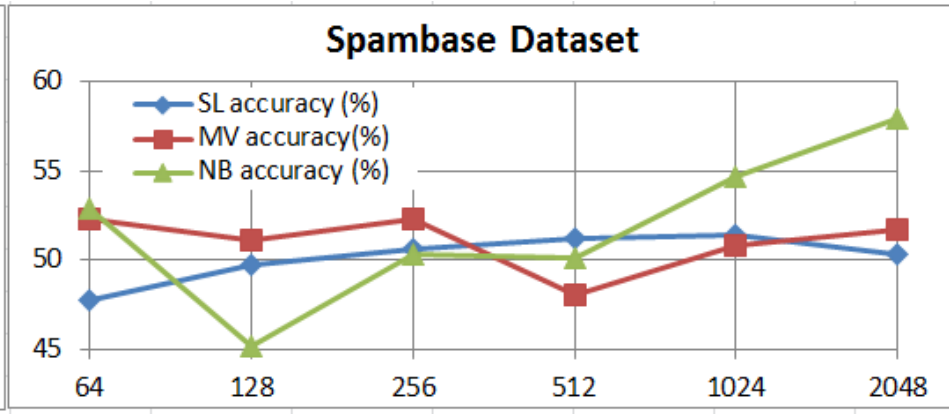




p ~ U(0.4, 0.6)

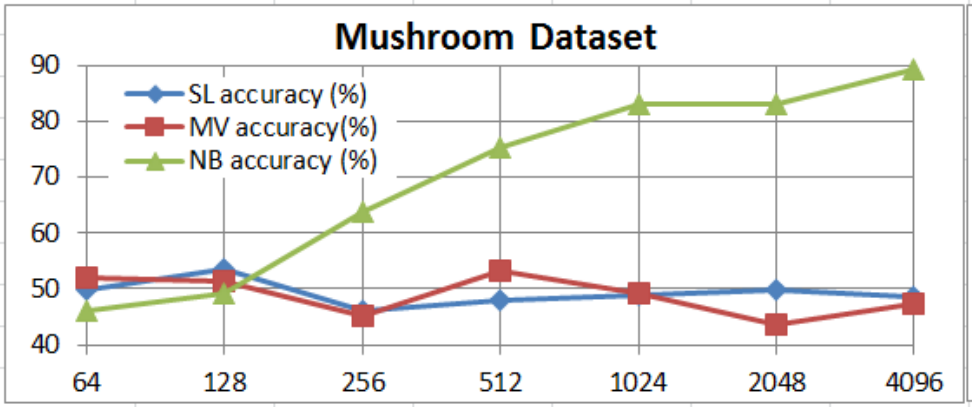
- Very noisy (mean = 0.5, random coin flip)
- SL and MV learning rates are flat
- NB wins by weighting more accurate workers

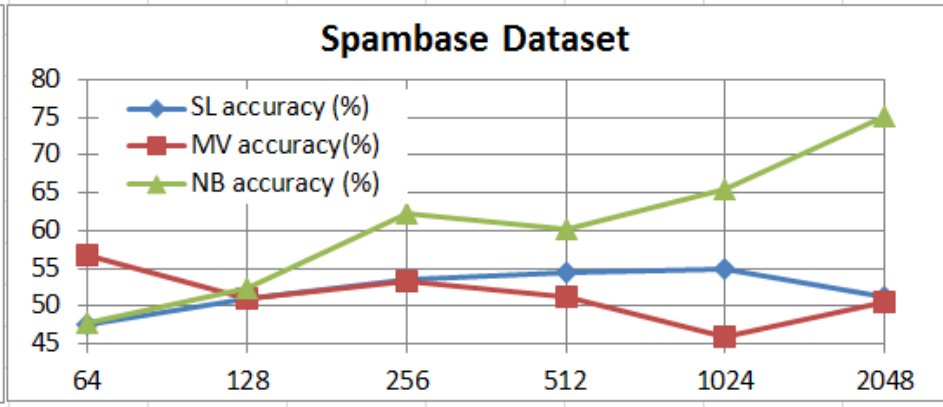




p ~ U(0.3, 0.7)

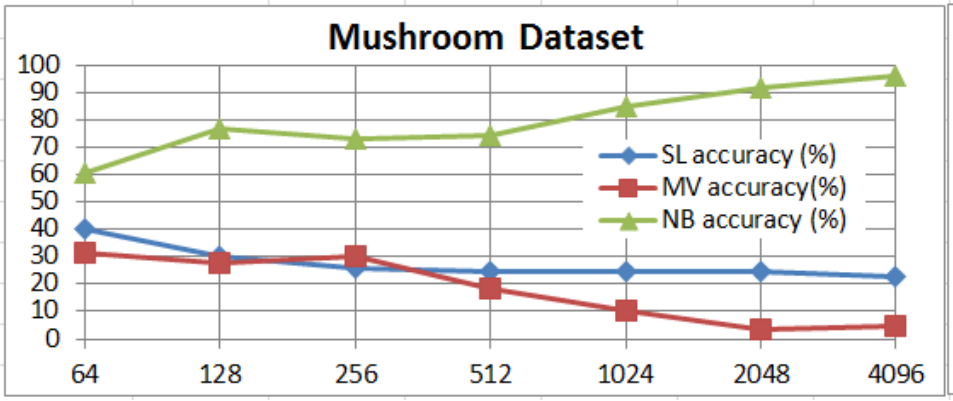
- Same noisy mean (0.5), but widen range
- SL and MV stay flat
- NB further outperforms

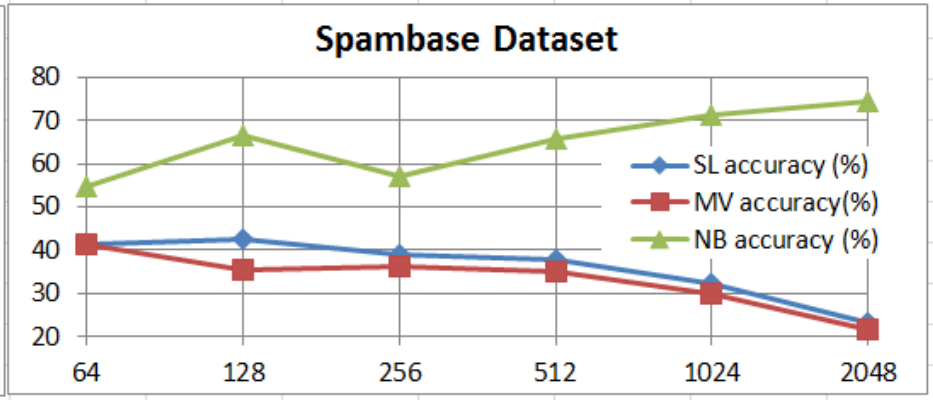




p ~ U(0.1, 0.7)

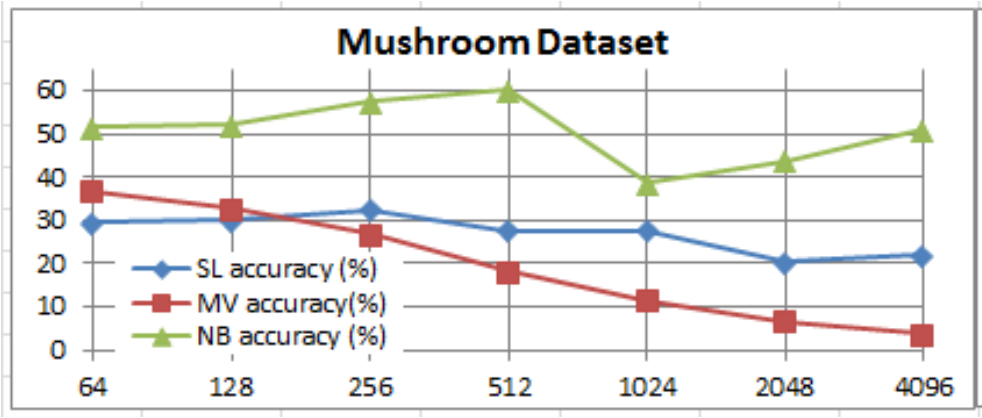
- Worsen accuracies further (mean = 0.4)
- NB virtually unchanged
- SL and MV predictions become anti-correlated
  - We should actually flip their predictions…

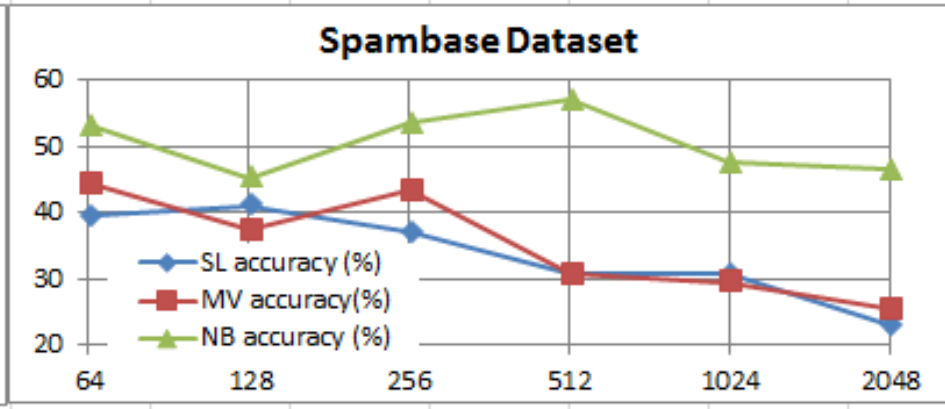




p ~ U(0.2, 0.6)

- Keep noisy mean 0.4, tighten range
- NB best of the worst, but only 50%
- Again, seems we should be flipping labels…



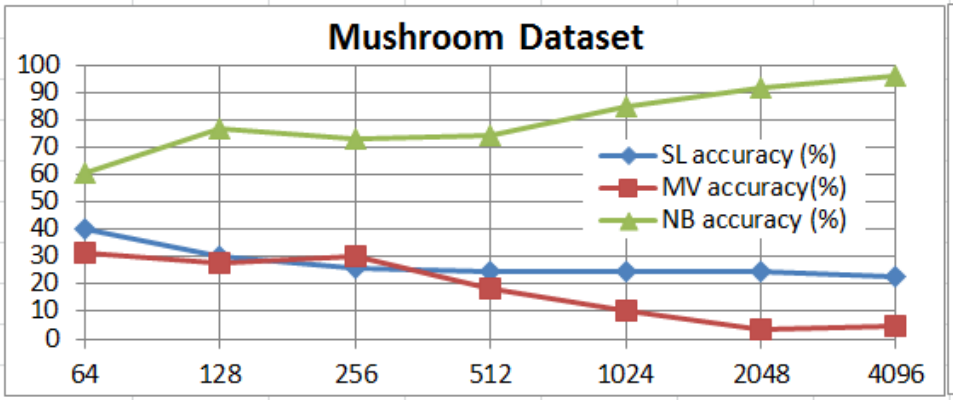


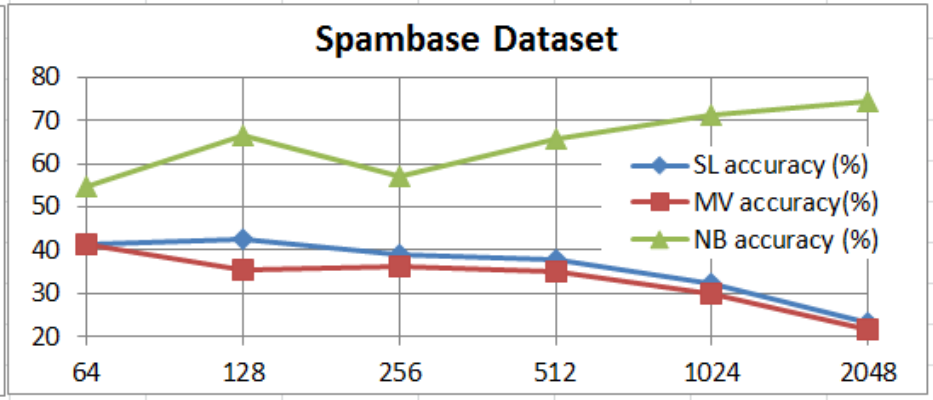
Label flipping

- Is NB doing better due to *how* it uses accuracy, or simply because it's using more information?
- If a worker's average accuracy is below 50%, we know he tends to be wrong (we've ignored this)
  - whatever he says, we should guess the opposite
- Flipping: put all methods on even-footing
  - Assume a given $p < 0.5$ produces label = y
  - Use label = (1-y) instead; for NB, use 1-p accuracy
  - Same as changing distribution so p always > 0.5

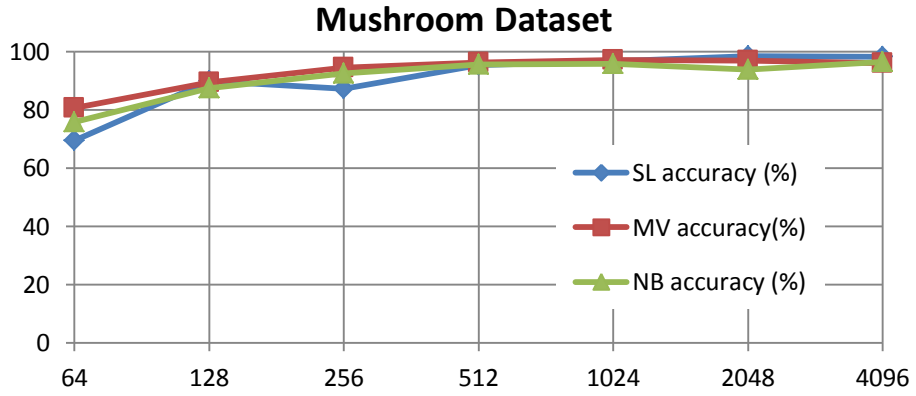
p ~ U(0.1, 0.7)

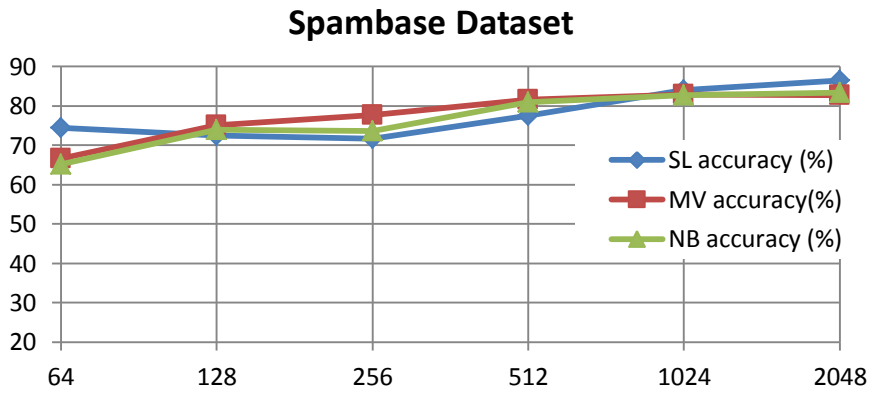
No flipping





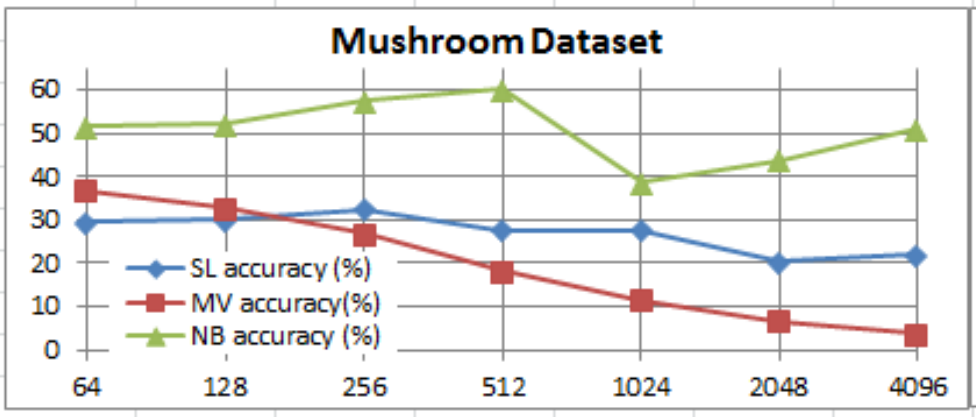
With flipping

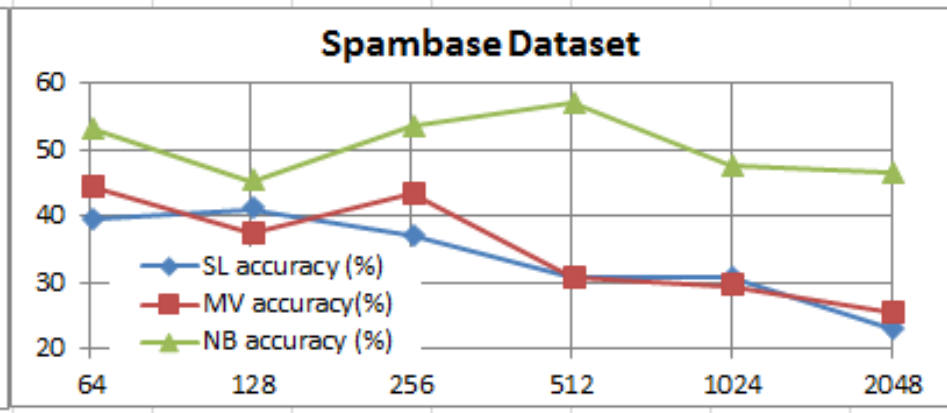




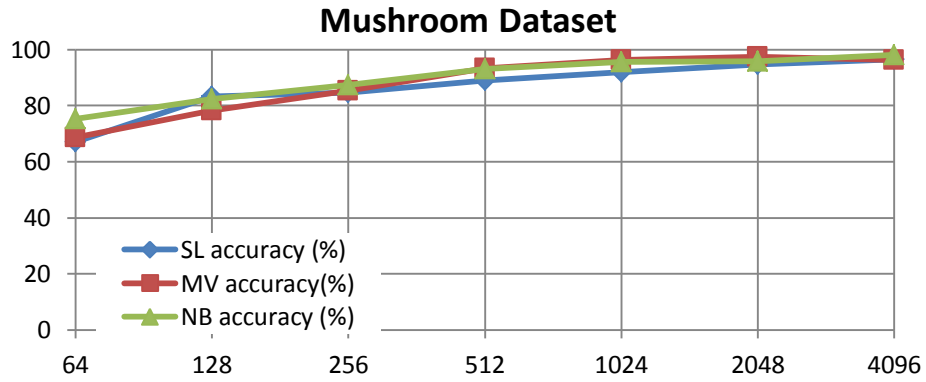
p ~ U(0.2, 0.6)

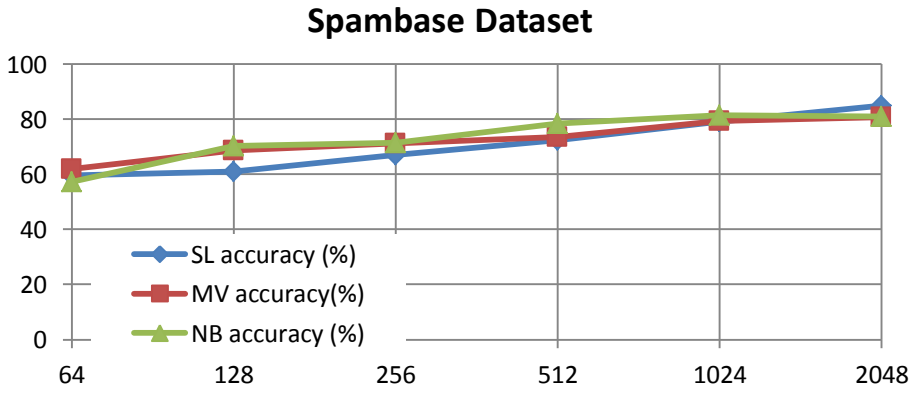
No flipping





With flipping





Conclusion

- Take-home: modeling accuracies matters, even if single labeling and majority vote
- But what about…
  - When accuracies are estimated (noisy)?
  - With real annotation errors (real distribution)?
  - With different learners or tasks (e.g. ranking)?
  - With dynamic choice of new example or re-label?
  - With active learning example selection?
  - With imbalanced classes?
  - …

Recent Events (2010 was big!)

http://ir.ischool.utexas.edu/crowd

- Human Computation: HCOMP 2009 & HCOMP 2010 at KDD
- IR: Crowdsourcing for Search Evaluation at SIGIR 2010
- NLP
  - The People's Web Meets NLP: Collaboratively Constructed Semantic Resources: 2009 at ACL-IJCNLP & 2010 at COLING
  - Creating Speech and Language Data With Mechanical Turk. NAACL 2010
  - Maryland Workshop on Crowdsourcing and Translation. June, 2010
- ML: Computational Social Science and Wisdom of Crowds. NIPS 2010
- Advancing Computer Vision with Humans in the Loop at CVPR 2010

- Conference: CrowdConf 2010 (organized by CrowdFlower)

Upcoming Crowdsourcing Events

http://ir.ischool.utexas.edu/crowd

Special issue of Information Retrieval journal on Crowdsourcing (papers due May 6, 2011)

Upcoming Conferences & Workshops

- CHI 2011 workshop (May 8)
- HCOMP 2011 workshop at AAAI (papers due April 22)
- CrowdConf 2011 (TBA)
- SIGIR 2011 workshop? (in review)
- TREC 2011 Crowdsourcing Track

Thanks!

Special thanks to our diligent crowd annotators and their relentless dedication to science…