Lecture 3
Humans as Data Sources: Label Aggregation

Instructor: Chien-Ju (CJ) Ho

Logistics

- Website: http://chienjuho.com/courses/cse518a/fa2019/
- Piazza: http://piazza.com/wustl/fall2019/cse518a/home

- You are responsible for following the announcements/discussion made on the website and Piazza.

Logistics: Assignment 1

- Amazon has been putting stronger restrictions on new accounts
  - For tax and data quality reasons

- If you have a hard time getting the account:
  - Borrow a MTurk account from others
  - Use Figure Eight (the number of available tasks to new workers might be low)
    - If you see less than 3 types of tasks, just earning $0.25 is ok
    - Please attach screenshots of available tasks
  - Use other crowdsourcing platforms
    - Clickworker
    - Be careful about potential scamming tasks (that ask you to give personal information or ask you to write fake reviews)

Logistics: Paper Reviews

- Reserve more time if you are not used to read research papers

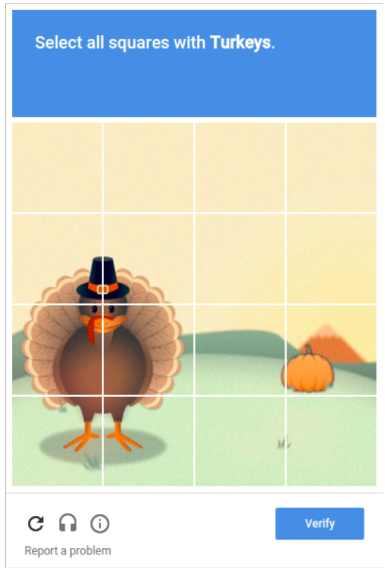
Sep 9	Label Aggregation: EM-based Algorithms	**Required** Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. Whitehill et al. NIPS 2009. **Optional** Learning from Crowds. Raykar et al. JMLR 2010. Maximum Likeihood Estimation of Observer Error-Rates Using the EM Algorithm. Dawid and Skene. Applied Statistics. 1979.	Submit Review (Due: Midnight, Sep 8)

- Review questions
  - Common questions
    - Summarize the paper in 2~3 sentences
    - List 1~3 points you like/dislike about the paper.
  - 1~2 paper-specific questions

Logistics

- Enrollment and waitlist
- Grades

Course Overview



Human as data sources: Label aggregation
Probabilistic reasoning to aggregate noisy human data

Humans are "Humans": Incentive design
Game theoretical modeling of humans and incentive design

Practical challenges: Real-time and complex tasks
Studies on workflow and team designs from HCI perspective

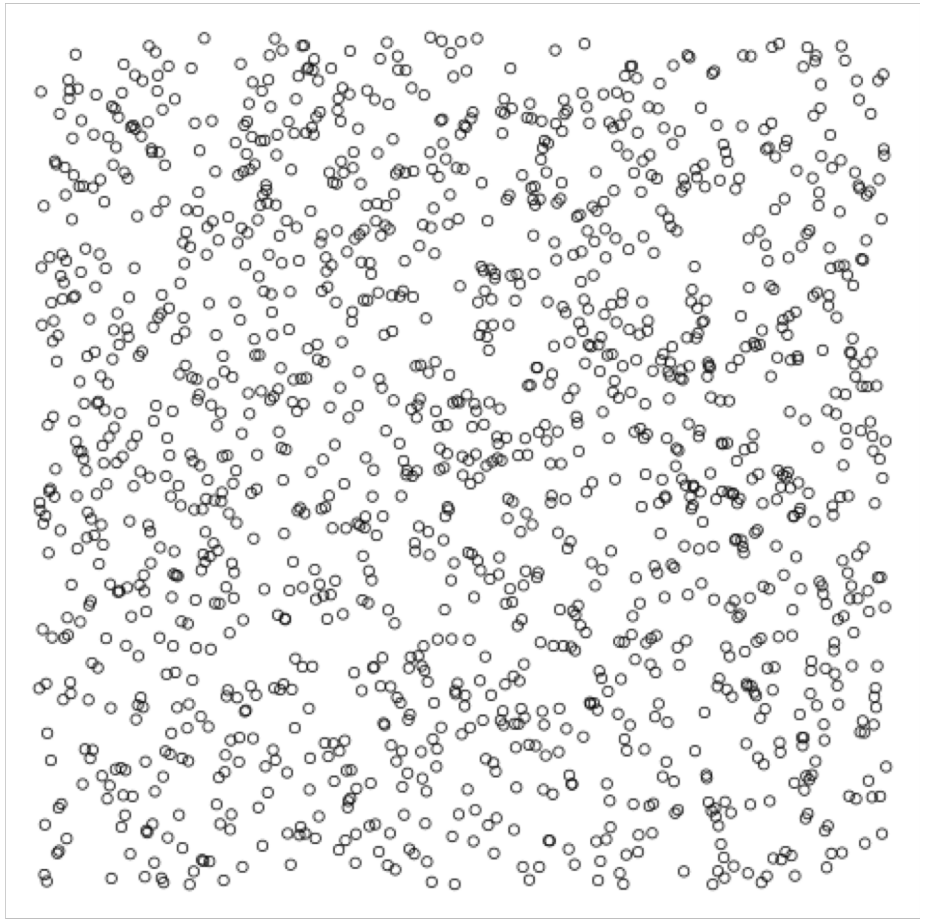
Selected recent topics:
Ethical issues of AI/ML, learning with strategic behavior, Human-AI collaborations.

Today's Lecture

- Probability background on label aggregation
  - (Weighted) Majority Voting
  - Maximum likelihood estimation
  - Concentration bounds

Remember this task?

- How many circles are in the image



These are the “labels” from you

11	494	853	1200
100	500	888	1280
163	500	888	1575
400	550	1000	1920
400	650	1000	2000
441	779	1000	2500
450	784	1086	3500
484	800	1200	4500

Mean: 1059.25
Median: 826.5
Answer: 721

- How to aggregate the answers?
  - Depend on how the labels are generated.

Example Model on Aggregation

- People have unbiased estimates of the true answer

  user guess = **true answer** + **Gaussian noise**

- With this model, we can estimate the number of users needed to achieve a certain level of accuracy (some form of law of large numbers).

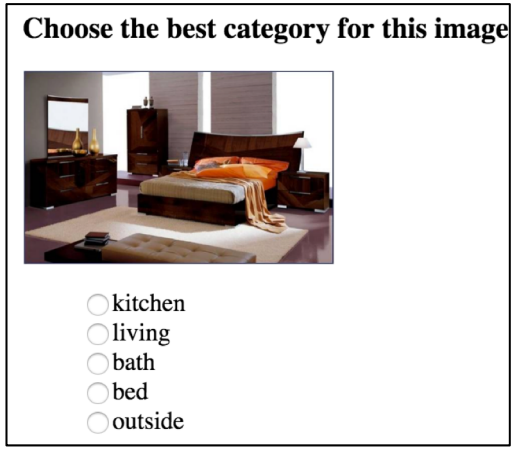
- Does this model capture the real human behavior?
- Is simple averaging the best aggregation method?

Focus on a common setting: Binary categorization/classification

- Binary classification

N.B.
- Guessing the Dots: **regression** problem
- Aggregation in general space is hard/non-trivial (e.g., aggregating multiple transcriptions)

- Most techniques/results can be extended to multi-label case, but the presentation could be more complicated.



Warm-Up Discussion

{1,0} or {+1, -1} are two common choices of binary labels

- Case 1: What's your prediction of the true label of task 1? Why?

	Worker 1	Worker 2	Worker 3	Worker 4	Worker 5
Task 1	+1	-1	+1	+1	-1

- Case 2: What's your prediction of the true label of task 2? Why?
  - What assumptions have you implicitly made in your arguments?

	True label	Worker 6	Worker 7	Worker 8	Worker 9
Task 2		+1	-1	+1	-1
Task 3	+1	+1	-1	+1	-1
Task 4	+1	-1	+1	-1	+1
Task 5	-1	-1	+1	+1	+1

Majority voting (MV)

Q1: ***Why*** MV might be a good idea?

Q2: Can we obtain ***theoretical guarantees*** for majority voting?

Understanding this simple scenario helps us develop aggregation methods for more complicated scenarios.

Probabilistic Approach

- Foundations of modern machine learning
  - You should develop a strong background if you are interested in doing research in AI/ML.

- High-level ideas:
  - Let D be the set of observations (e.g., training dataset, the set of labels we got from workers)
  - Let θ be the set of latent parameters we care about (e.g., ML hypothesis, true labels)

  - Two important concepts
    - Posterior: $\Pr(\theta|D)$ [More discussion in CSE515T]
    - Likelihood: $\Pr(D|\theta)$ [More discussion in CSE417T]

  - Connection: $\Pr(\theta|D) = \frac{\Pr(\theta)\Pr(D|\theta)}{\Pr(D)}$

Maximum likelihood estimation:
Find $\theta^* = argmax_\theta \Pr(D|\theta)$

$\Pr(\theta)$: Prior
(Additional assumption)

Why Majority Voting:
Majority Voting Gives the Maximum-Likelihood Estimation

- Consider a task with true label l^*
- We collect labels $L = \{l_1, l_2, \ldots, l_n\}$ from n workers for this task.
- Each worker gives the correct label with probability $p > 0.5$.

- l^* is the latent variable and L is our observation.

- Maximum likelihood estimation (MLE):
  - Predict +1 if $\Pr[\, L|l^* = +1] \geq \Pr[\, L|l^* = -1]$
  - Predict -1 otherwise

Likelihood: $\Pr[D|\theta]$
D: Observations
θ: latent variables

MLE approach (roughly speaking):
Find $\theta^* = argmax_\theta \Pr[D|\theta]$

How should we model the label generation process?

A Simple Model for Case 1

	Worker 1	Worker 2	Worker 3	Worker 4	Worker 5
Task 1	+1	-1	+1	+1	-1

- Assumption:
  - Each worker has the same ability of giving correct labels
  - Each worker gives label in probabilistic way
  - Each worker gives the correct label independently ***with probability p > 0.5***
  - Given no additional information, this is close to the best you can model

Maximum likelihood estimation (MLE):

Predict +1 if $\Pr[L|l^* = +1] \geq \Pr[L|l^* = -1]$

Predict -1 otherwise

Derivation of MLE ⇔ MV (details on the board)

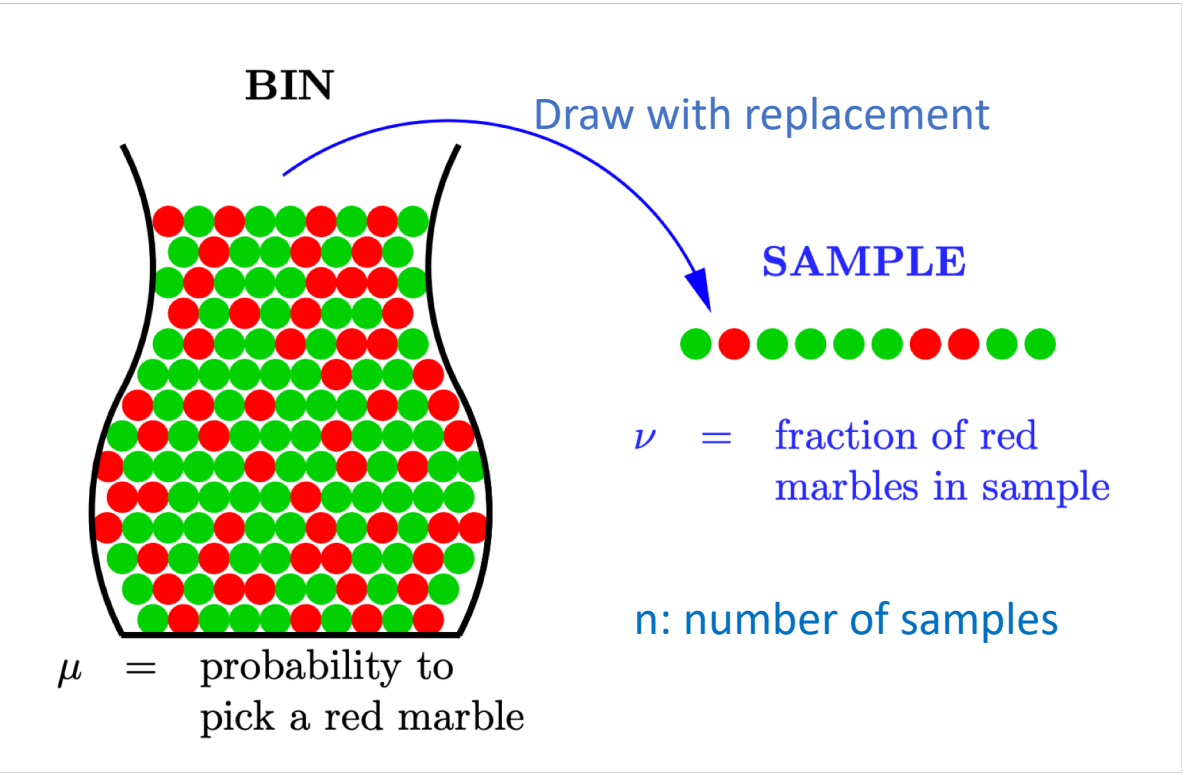
- Key Assumption: independent worker labels
  - Let (n_+, n_-) be the number of (+1, -1) labels in L
  - $\Pr[\, L | l^* = +1] = p^{n_+}(1-p)^{n_-}$
  - $\Pr[\, L | l^* = -1] = p^{n_-}(1-p)^{n_+}$

- MLE rule is equivalent to
  - Predict +1 if $\ln \frac{p^{n_+}(1-p)^{n_-}}{p^{n_-}(1-p)^{n_+}} \geq 0$

  - Predict +1 if $(n_+ - n_-)(\ln p - \ln(1-p)) \geq 0$
  - Predict +1 if $n_+ \geq n_-$
  - This is majority voting

What theoretical guarantee can MV achieve?

- Consider a thought experiment



What can we say about μ from ν?

Law of large numbers

- When $n \to \infty$, $\nu \to \mu$

Hoeffding's Inequality

- $\Pr[|\mu - \nu| > \epsilon] \leq 2e^{-2\epsilon^2 n}$ for any $\epsilon > 0$

Interpretations

$$\Pr[|\mu - \nu| > \epsilon] \leq 2e^{-2\epsilon^2 n}$$

Define $\delta = \Pr[|\mu - \nu| > \epsilon]$: Probably of "bad events"

- Fix $\epsilon, \delta = O(e^{-n})$; Fix $n, \delta = O(e^{-\epsilon^2})$; Fix $\delta, \epsilon = O(\sqrt{\frac{1}{n}})$
- n=1000
  - $\mu - 0.05 \leq \nu \leq \mu + 0.05$ with 99% chance
  - $\mu - 0.10 \leq \nu \leq \mu - 0.10$ with 99.9999996% chance

- ν is approximately close to μ with high probability
- ν as an estimate of μ is **p**robably **a**pproximately **c**orrect (P.A.C.)

PAC learning is proposed by Leslie Valiant, who wins the Turing award in 2010.

More general form of Hoeffding's inequality

- Let $X_1, \ldots, X_n$ be independent random variables
  - X_i is bounded in the range $[a_i, b_i]$

- Let $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$

- (One-sided) Hoeffding's inequality

$$\Pr[\mathbb{E}[\bar{X}] - \bar{X} \geq \epsilon] \leq \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

We get our previous bound by setting $b_i = 1$ and $a_i = 0$

Connection to Our Problem

$$\Pr[\mathbb{E}[\bar{X}] - \bar{X} \geq \epsilon] \leq \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

- Without loss of generality, assume $l^* = +1$
- X_i is the random variable of the label provided by worker i

- $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$: $\mathbb{E}[\bar{X}] = 2p - 1 > 0$
- Majority voting => Predict $sign(\bar{X})$

- Probability of making a wrong prediction

$$\begin{aligned}\Pr[\bar{X} \leq 0] &= \Pr\left[\mathbb{E}[\bar{X}] - \bar{X} \geq \mathbb{E}[\bar{X}]\right] \\ &\leq \exp\left(-\frac{1}{2}n\,(\mathbb{E}[\bar{X}])^2\right) \\ &= \exp\left(-\frac{1}{2}n\,(2p-1)^2\right)\end{aligned}$$

Looks like we solved the problem?

if we assume all workers are the same….

	True label	Worker 6	Worker 7	Worker 8	Worker 9
Task 2		+1	-1	+1	-1
Task 3	+1	+1	-1	+1	-1
Task 4	+1	-1	+1	-1	+1
Task 5	-1	-1	+1	+1	+1

What happens if workers are different

- Assume we obtain n labels from n workers.
- Worker $i \in \{1, \dots, n\}$
  - provides label $l_i \in \{-1, +1\}$
  - assumption: each label is correct with probability p_i
  - assume we know p_i

- How should we aggregate?
  - Weighted majority voting?

Predict $sign(\sum_{i=1}^{n} w_i l_i)$

Weighted Majority Voting

- Weighted majority voting **Predict $sign(\sum_{i=1}^{n} w_i l_i)$**

- Turns out weighted majority voting leads to MLE
  - With weight $w_i = \ln \frac{p_i}{1-p_i}$ for label l_i
  - Proof on the blackboard

- The weights to minimize the Hoeffding error are different
  - To minimize Hoeffding error, set weights $w_i = 2p_i - 1$ for label l_i
  - Proof on the blackboard (Lemma 1 in Ho et al. ICML 2013)

For the next three lectures

	True label	Worker 6	Worker 7	Worker 8	Worker 9
Task 2		+1	-1	+1	-1
Task 3		+1	-1	+1	-1
Task 4		-1	+1	-1	+1
Task 5		-1	+1	+1	+1

- Unknown worker skills
- Different task difficulties
- More factors to consider (some structures of tasks/workers?)
- …

Typical label aggregation approach

- Propose a model to describe the label generation process
- True labels are the “latent variables” of the process
- Using inference algorithms (e.g., EM) to learn the latent variables

Sep 9	Label Aggregation: EM-based Algorithms	**Required** Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. Whitehill et al. NIPS 2009. **Optional** Learning from Crowds. Raykar et al. JMLR 2010. Maximum Likeihood Estimation of Observer Error-Rates Using the EM Algorithm. Dawid and Skene. Applied Statistics. 1979.
Sep 11	Label Aggregation: Matrix-based Methods	**Required** Who Moderates the Moderators? Crowdsourcing Abuse Detection in User-Generated Content. Ghosh, Kale, and McAfee. EC 2011. - If you want to refresh your memory on matrix algebra, Matrix Cookbook is a good resource. Section 5 contains the matrix decomposition part. **Optional** Budget-Optimal Crowdsourcing using Low-rank Matrix Approximations. Karger, Oh, and Shah. Allerton 2011. Spectral Methods Meet EM: A Provably Optimal Algorithm for Crowdsourcing. Zhang et al. JMLR 2016.
Sep 16	Label Aggregation: Others and Discussion	**Required** Iterative Learning for Reliable Crowdsourcing Systems. Karger, Oh, and Shah. NIPS 2011. **Optional** Variational Inference for Crowdsourcing. Liu, Peng, and Ihler. NIPS 2012. Learning from the Wisdom of Crowds by Minimax Entropy. Zhou et al. NIPS 2012.

Write down likelihood/posterior function
Using EM algorithms to find the parameters that maximize likelihood/posterior

Write labels as a matrix (worker by task)
Using low rank matrix approximation

A bunch of other methods

Discussion

- Do you think the models we made so far make sense? Why?
- Can you think of other important aspects (at least in some particular applications) that should be modeled?
- Take this time to find your potential teammates!