

Lecture 2

Humans as Data Sources: Label Aggregation

Instructor: Chien-Ju (CJ) Ho

Logistics

- Website: <http://chienjuho.com/course/cse518a>
- Piazza: <http://piazza.com/wustl/spring2024/cse518a>
- Please follow the updates and announcements.
- You are responsible for following the announcements/discussion made on the website and Piazza.

Logistics: Paper Reviews

- Submit your review for the “required reading” of each lecture
 - Submit via Gradescope
 - Due on **11:59pm the day before the lecture**
 - There will be no reminders; make sure to do it before each lecture
- Review questions
 - 2 common questions (summary, critics)
 - 2 paper-specific questions
 - Expectation: An informative paragraph (a few sentences will do) for each question.
- I’ll use the # review submission for Wednesday to **estimate class size**
 - Let me know if you plan to enroll but can’t submit the review for some reason
- Reserve more time if you are not used to read research papers
 - Some papers are heavier (mathematically) than the others
 - Expect a very math-heavy reading next Monday.

Logistics: Assignment 1

- Due: Feb 9 (Friday)
 - You don't have all background information to do it yet
 - I post it early so you know what to expect
- Programming assignment
 - Implement and compare the performance of majority voting, EM, and SVD
 - You can use any programming language you like
 - You will be graded based on the report
- You need to submit your codes
 - Used for plagiarism tests
 - Might check the codes if we have confusions/doubts on the reported results

Logistics: Late-Day Policies

- Homework assignments
 - 4 late days in total
 - You can use up to 2 late days for each assignment
- Reviews
 - You can skip up to 2 reviews without penalty
 - No late submissions
 - You don't need to submit reviews for the reading you present
 - Presenters are responsible for designing the two paper-specific questions
- Presentation and Project-related reports
 - No late submissions and no skips

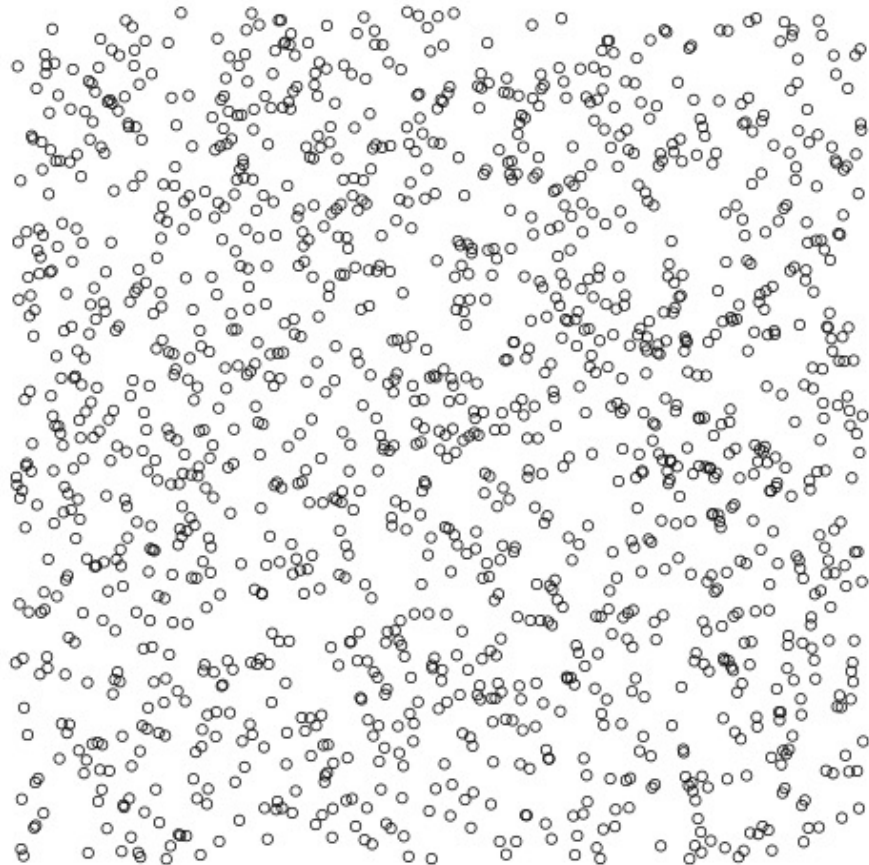
Lecture Today

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

243	875	1370	2000
250	1000	1400	2500
256	1000	1400	3025
400	1134	1555	3728
439	1170	1600	4000
445	1200	1800	8,900
855	1250	2000	300,000

Mean: 12349.82143

Median: 1310

True Answer: 721

How to aggregate the answers?

- Depend on how the labels are generated.

A Naïve Model of Label Generation

- People have unbiased estimates of the true answer

$$\text{user guess} = \text{true answer} + \text{Gaussian noise}$$

Observations

Latent values we
want to know

Zero-Mean Noises

- If this model approximates the reality well, we can decide on **aggregation**
 - **Mean** of user guesses is an **unbiased** estimator for **true answer**

This Lecture Focuses on Binary Classification

- Binary classification

Is this the Golden Gate Bridge?



☐ Yes
☐ No

Note

- 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, though with more complicated details

What type of business is this ?

Bank of America

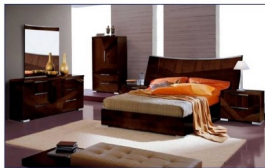
☒ Financial Institute

☐ Retailer

☐ Restaurant

☐ Other

Choose the best category for this image



☐ kitchen

☐ living

☐ bath

☐ bed

☐ outside

Defining Label Aggregation

- Input

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

$\{1,0\}$ or $\{+1, -1\}$ are two common choices of binary labels
We'll use $\{+1,-1\}$ for its mathematical convenience

- Output: Estimated task labels

- Label aggregation is sometimes also called truth discovery

Warm-Up Discussion

- 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 in probability/statistics if 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
 - Likelihood: $\Pr(D|\theta)$ [More discussion in CSE417T]
 - Posterior: $\Pr(\theta|D)$ [More discussion in CSE515T]
 - Connection: $\Pr(\theta|D) = \frac{\Pr(\theta)\Pr(D|\theta)}{\Pr(D)}$

Maximum likelihood estimation (MLE)
Find $\theta^* = \operatorname{argmax}_{\theta} \Pr(D|\theta)$

Maximum a posteriori (MAP)
Find $\theta^* = \operatorname{argmax}_{\theta} \Pr(\theta|D)$

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

Why Majority Voting?

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

Under some reasonable assumptions,

Majority voting leads to **maximum likelihood estimation**

Formulation

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

- Consider a task with true label $l^* \in \{-1, +1\}$
- We collect labels $L = \{l_1, l_2, \dots, l_n\}$ from n workers for this task.

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

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

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

Maximum likelihood estimation
Find $\theta^* = \operatorname{argmax}_{\theta} \Pr[D|\theta]$

It requires models/assumptions to calculate

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

Maximum likelihood estimation (MLE):

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

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

- Assumptions:
 - Each worker gives a label in a probabilistic manner
 - Each worker has the same ability of giving correct labels
 - Each worker gives a label independently without influence from others
 - Each worker is more likely to provide a correct label than a wrong label
- Model
 - 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

Derivation of MLE \Leftrightarrow MV

Maximum likelihood estimation (MLE):

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

Predict -1 otherwise

- Let (n_+, n_-) be the number of $(+1, -1)$ labels in L
 - $\Pr[L|l^* = +1] =$
 - $\Pr[L|l^* = -1] =$

Model: Each worker gives the correct label independently
with probability $p > 0.5$

Derivation of MLE \Leftrightarrow MV

Maximum likelihood estimation (MLE):

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

Predict -1 otherwise

- Let (n_+, n_-) be the number of $(+1, -1)$ labels in L

- $\Pr[L|l^* = +1] = \binom{n}{n_+} p^{n_+} (1-p)^{n_-}$

- $\Pr[L|l^* = -1] = \binom{n}{n_+} p^{n_-} (1-p)^{n_+}$

Model: Each worker gives the correct label independently
with probability $p > 0.5$

- MLE rule is equivalent to

Derivation of MLE \Leftrightarrow MV

Maximum likelihood estimation (MLE):

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

Predict -1 otherwise

- Let (n_+, n_-) be the number of (+1, -1) labels in L

- $\Pr[L|l^* = +1] = \binom{n}{n_+} p^{n_+} (1-p)^{n_-}$

- $\Pr[L|l^* = -1] = \binom{n}{n_+} p^{n_-} (1-p)^{n_+}$

Model: Each worker gives the correct label independently
with probability $p > 0.5$

- MLE rule is equivalent to

- Predict +1 if $\ln \frac{p^{n_+} (1-p)^{n_-}}{p^{n_-} (1-p)^{n_+}} \geq 0$

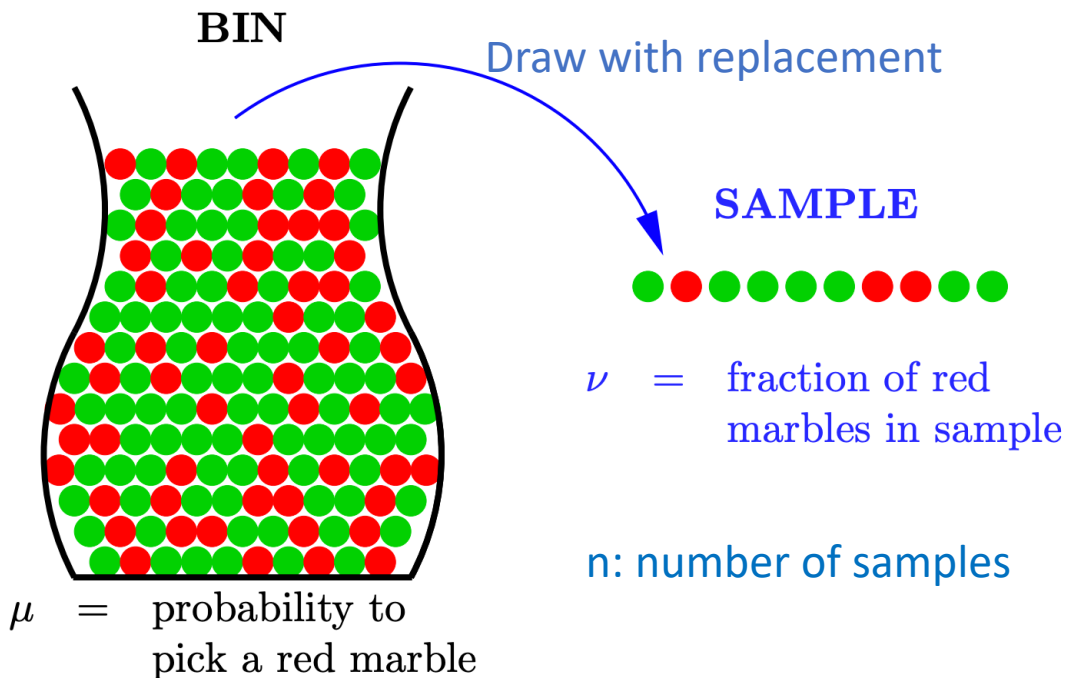
- Predict +1 if $(n_+ - n_-)(\ln p - \ln(1-p)) \geq 0$

- 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 $n \rightarrow \infty$, $\nu \rightarrow \mu$

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 **probably approximately correct** (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, \dots, X_n be independent random variables
 - X_i is bounded in the range $[a_i, b_i]$

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

- (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

- $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$
- $\mathbb{E}[\bar{X}] = 2p - 1 > 0$

- Majority voting => Predict $\text{sign}(\bar{X})$
- Probability of making a wrong prediction

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

Looks like we solved the problem?

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

Weighted Majority Voting

- Weighted majority voting

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

- Turns out weighted majority voting leads to MLE

- With weight $w_i = \ln \frac{p_i}{1-p_i}$ for label l_i

- The weights to minimize the Hoeffding error are different

- To minimize Hoeffding error, set weights $w_i = 2p_i - 1$ for label l_i
 - (Lemma 1 in [Ho et al. ICML 2013](#))

For the next two 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

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 Likelihood Estimation of Observer Error-Rates Using the EM Algorithm](#). Dawid and Skene. Applied Statistics. 1979.

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.

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

Discussion

- Do you think the models we made so far make sense? Why? Under what conditions can our model break? What can we do to address those conditions?
- Can you think of other important aspects (at least in some applications) that should be modeled?
- Take this time to find your potential teammates!

Assignment 1

Assignment 1

- Goal: Perform Label Aggregation on the given dataset
 - Recognizing Textual Entailment (RTE) task

Text:

- Many experts think that there is likely to be another terrorist attack on American soil within the next five years.

Hypothesis:

- There will be another terrorist attack on American soil within the next five years.

Answer: NO

Task: Whether the first sentence implies the second hypothesis

- 800 tasks; 164 workers; 10 labels per task

You might want to convert the labels/gold to {+1,-1}

Assignment 1

Ignore this column

Worker ID

Task ID

Worker Label

Ground Truth
(only used to evaluate the aggregation algorithm)

!amt_annotation_ids	!amt_worker_ids	orig_id	response	gold
89KZPYXSTGTJ0CZY2Y1ZB28YQ9GBT88Z2W1KDYZT	A19IBSKBTABMR3	266	1	1
89KZPYXSTGTJ0CZY2Y1ZYAJC56Z6FBPGXJYVPXM0	AEX5NCH03LWSG	266	1	1
89KZPYXSTGTJ0CZY2Y1ZFWHATWX49Y3ZTPX4FYH0	A17RPF5ZMO75GW	266	1	1
89KZPYXSTGTJ0CZY2Y1ZV89Z3WRZ6R8ZM4ZZZ070	A15L6WGK3VU7N	266	0	1
89KZPYXSTGTJ0CZY2Y1ZWZHYZCCYYVYPDZVNRAZ	A3U7T47F498T1P	266	1	1
89KZPYXSTGTJ0CZY2Y1Z09PZYS137RPZT6SY4A20	AXBQF8RALCIGV	266	1	1
89KZPYXSTGTJ0CZY2Y1ZQ30CJXY2EB96XJS543YZ	A1DCEOFAUIDY58	266	1	1
89KZPYXSTGTJ0CZY2Y1ZXZ3ZNY7VZKZSCY0B94Z	A1Q4VUJBM78YR	266	0	1
89KZPYXSTGTJ0CZY2Y1ZDZGGWVY8XDZTKYC9XKZ	A18941IO2ZZWW6	266	1	1
89KZPYXSTGTJ0CZY2Y1Z3Z7ZWY9J4WFMX60VRVXZ	A11GX90QFWDLMM	266	1	1
89KZPYXSTGTJ0CZY2Y1ZB28YQ9GBT88Z2W1KDYZT	A19IBSKBTABMR3	934	0	0
89KZPYXSTGTJ0CZY2Y1ZYAJC56Z6FBPGXJYVPXM0	AEX5NCH03LWSG	934	0	0
89KZPYXSTGTJ0CZY2Y1ZFWHATWX49Y3ZTPX4FYH0	A17RPF5ZMO75GW	934	0	0
89KZPYXSTGTJ0CZY2Y1ZV89Z3WRZ6R8ZM4ZZZ070	A15L6WGK3VU7N	934	0	0
89KZPYXSTGTJ0CZY2Y1ZWZHYZCCYYVYPDZVNRAZ	A3U7T47F498T1P	934	0	0
89KZPYXSTGTJ0CZY2Y1Z09PZYS137RPZT6SY4A20	AXBQF8RALCIGV	934	1	0
89KZPYXSTGTJ0CZY2Y1ZQ30CJXY2EB96XJS543YZ	A1DCEOFAUIDY58	934	0	0
89KZPYXSTGTJ0CZY2Y1ZXZ3ZNY7VZKZSCY0B94Z	A1Q4VUJBM78YR	934	0	0

Assignment 1

- Requirements
 - Create random subsampled datasets:
 - Original: 800 tasks; 164 workers; **10** labels per task
 - Randomly sub-sample the labels, such that each task has **k** labels
 - **k=1, 2, 3, ..., 10**
 - Implement **majority voting, EM, SVD**
 - Calculate the error (ratio of tasks the algorithms make wrong predictions)
 - Compare the performance of algorithms
 - Generate a figure with x-axis being k, y-axis being error
 - Plot 3 curves, each corresponding to an algorithm
 - Offer brief discussion
 - **"Expand"** the dataset to see the performance of SVD with larger **k**