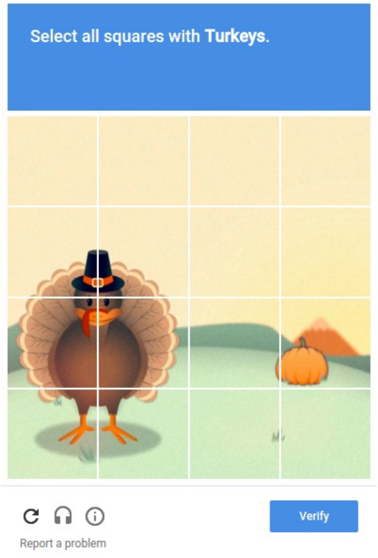
Lecture 3
Humans as Data Sources: Label Aggregation

Instructor: Chien-Ju (CJ) Ho

Lecture Today

Course Overview



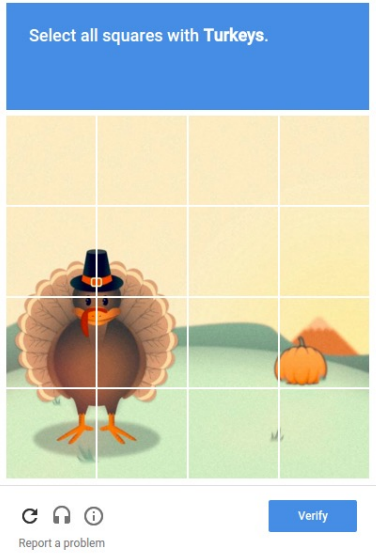
Human as data sources:
Label aggregation
Probabilistic reasoning to aggregate noisy human data

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

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

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

Course Overview



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

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

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

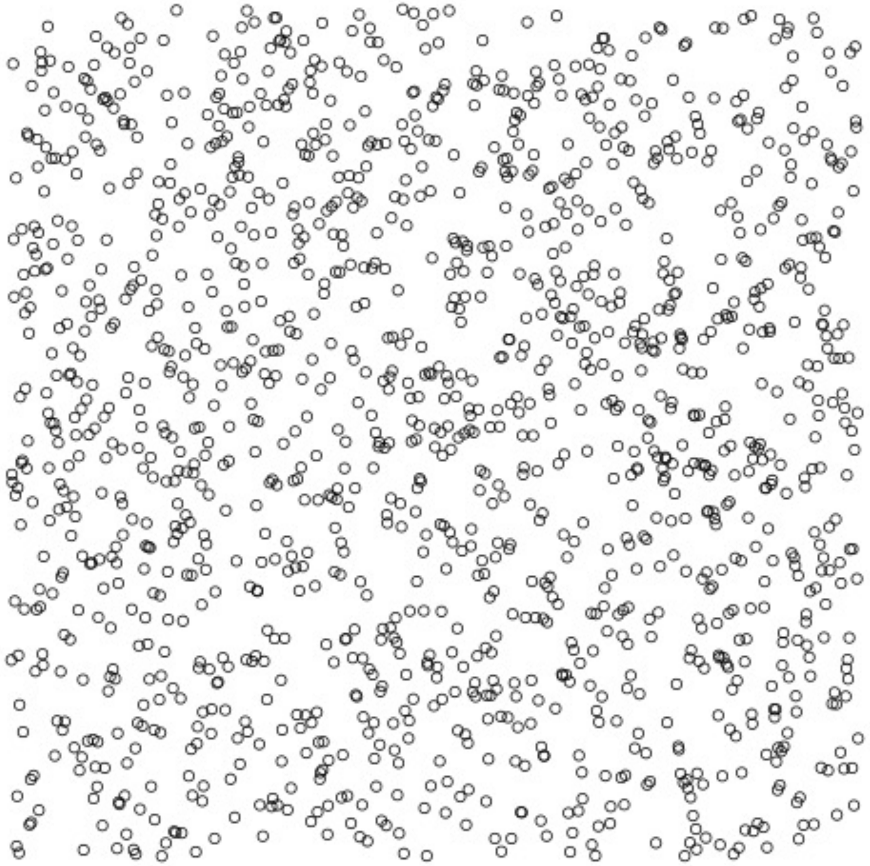
Selected recent topics:
Ethical issues of AI/ML, learning with strategic behavior, Human-AI collaborations.

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

256	649	900	1296
350	650	978	1500
375	700	1000	1700
387	720	1008	2000
455	730	1024	2100
494	739	1028	2500
519	800	1200	2500
550	800	1200	3000
625	847	1232	10,000
625	899	1250	102000

Mean: 3789.65
Median: 899.5

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

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

Observations | Latent values we want to know | Zero-Mean Noises

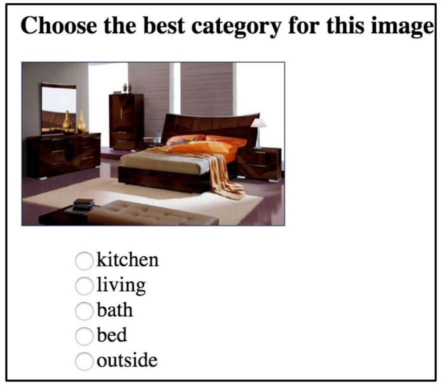
- If this model approximates the reality well, we can decide on **aggregation**
  - **Mean** of user guesses is an **unbiased** estimator for **true answer**

This Lecture Focuses on Binary Classification

- Binary classification

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



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		
…					

- Output: Estimated task labels
  - Label aggregation is sometimes also called truth discovery

Warm-Up Discussion

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

- 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 $\theta^* = argmax_\theta \Pr(D|\theta)$

Maximum a posteriori (MAP)
Find $\theta^* = argmax_\theta \Pr(\theta|D)$

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

Why Majority Voting?

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

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 $L = \{l_1, l_2, \ldots, l_n\}$ 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 $\boxed{\Pr[\, L|l^* = +1]} \geq \Pr[\, L|l^* = -1]$
  - Predict -1 otherwise

Maximum likelihood estimation
Find $\theta^* = argmax_\theta \Pr[D|\theta]$

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

- Assumption:
  - Each worker gives a label in a probabilistic manner
  - Each worker has the same ability of giving correct labels
  - Each worker gives a label on his/her own
  - 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 ⇔ MV

Maximum likelihood estimation (MLE):

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

Predict -1 otherwise

- Key assumption: independent worker labels

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

Derivation of MLE ⇔ MV

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

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

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

Derivation of MLE ⇔ MV

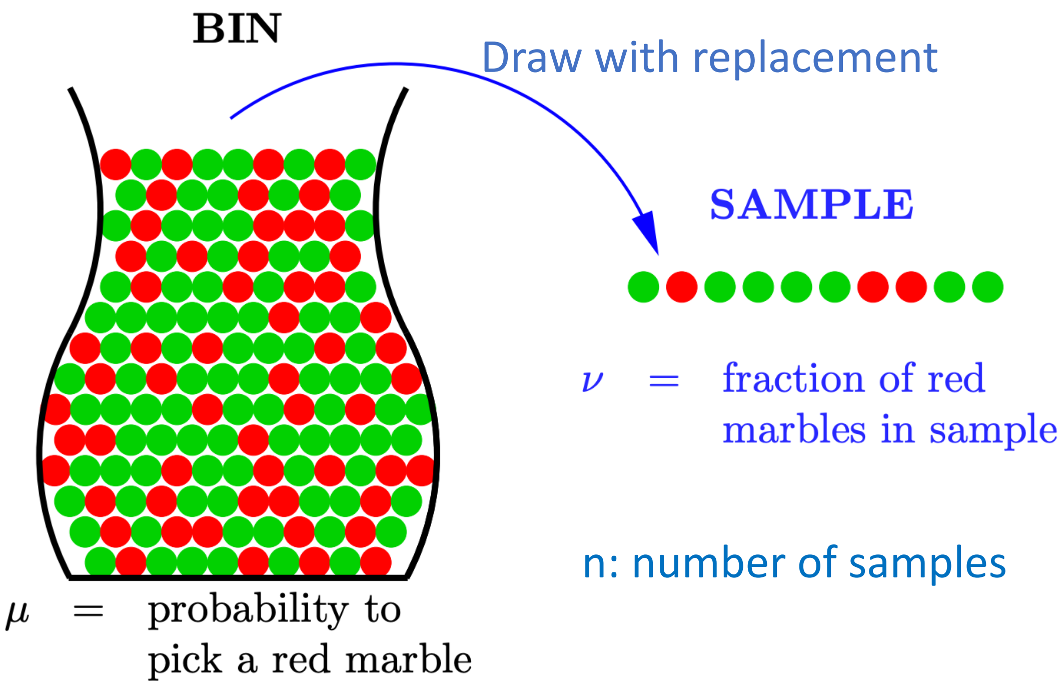
Maximum likelihood estimation (MLE):
Predict +1 if $\Pr[\,L|l^* = +1] \geq \Pr[\,L|l^* = -1]$
Predict -1 otherwise

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

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

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

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

- Fix $\epsilon, \delta = O(e^{-n})$; Fix $n, \delta = O\left(e^{-\epsilon^2}\right)$; Fix $\delta, \epsilon = O(\sqrt{\frac{1}{n}})$
- n=1000
  - $\mu - 0.05 \le \nu \le \mu + 0.05$ with 99% chance
  - $\mu - 0.10 \le \nu \le \mu - 0.10$ 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 $X_1, \dots, X_n$ be independent random variables
  - X_i is bounded in the range $[a_i, b_i]$

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

- (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

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

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

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 $\mathrm{w_i} = \ln \frac{p_i}{1-p_i}$ 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	[redacted]	+1	-1	+1	-1
Task 3	[redacted]	+1	-1	+1	-1
Task 4	[redacted]	-1	+1	-1	+1
Task 5	[redacted]	-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 Likeihood 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!