Aggregating outputs
Jérôme Waldspühler, McGill University

Computational problems
- Multiplication: two numbers ⇒ product
- Sorting: set of objects ⇒ set of objects sorted
- Medical diagnosis: x-ray, lab tests ⇒ diagnosis
- Object recognition: image ⇒ tag
- Translation: source sentence ⇒ target sentence
- Editing: text ⇒ corrected text
- Planning: goal, constraints ⇒ sequence of actions

What is an algorithm?
“A finite set of rules which gives a sequence of operations for solving a specific type of problem” such that:
- **Input**: One or more inputs
- **Output**: One or more outputs which have a specific relation to the input(s)
- **Finiteness**: It must terminate after a finite number of steps.
- **Effectiveness**: Each operation needs to be basic
- **Definiteness**: Each step must be well defined and unambiguous.

(Knuth, 1973)

Operations & Controls
- **Sequence**
  - Operation 1
  - Operation 2
  - ... Operation n
- **Selection**
  - Operation 1
  - Condition
  - Operation 2
  - ... Operation n
- **Repetition**
  - Parallel
  - Operation 1
  - Operation n

Programming Frameworks
- **Frameworks**
  - Turkit (for Amazon Mechanical Turk)
  - Boto (for AMT)
  - PyBossa
- **Functionality**
  - Post HITs (Human Intelligence Tasks)
  - Store and retrieve previous HIT results

Evaluation
- **Correctness**
  - HIT are noisy. Their robustness must be estimated.
- **Efficiency**
  - Time (execution + task discovery)
  - Query Complexity
  - Cost
Designing HIT

**Task routing**

**Task design**

**Task aggregation**

Today

Aggregating output

Challenges:
- Output are noisy (lack of expertise)
- Humans are not always reliable (cheating)
- Cultural context may bias the answers

Goal: Automatic procedure to merge HIT results

Assumptions:
- It exists a “True” answer
- Redundancy helps

What is truth?

**Objective truth**
- exists freely or independently from a mind (E.g. ideas, feelings)
- Medical diagnosis, protein structure, number of birds...

**Cultural truth**
- shared beliefs of a group of people, often involving perceptual judgments.
- Is the music sad? Is this image pornographic? Is this text offending?...

Latent class models

**Observed**: HIT outputs

**Latent (hidden)**: Truth, user experience, task difficulty.

Workers: Tasks

\[
\begin{bmatrix}
O_{11} & O_{12} & \cdots & O_{1N} \\
O_{21} & O_{22} & \cdots & O_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
O_{M1} & O_{M2} & \cdots & O_{MN}
\end{bmatrix}
\]

Solution:

\[
\begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_M
\end{bmatrix}
\]

- Often, the matrix is incomplete
- Ground truth may never been known

Majority vote

**Observed output**

\[
Y_n \propto \arg \max_j P(Y_n = j | O)
\]

**Solution**

\[
Y_n \propto \arg \max_j \frac{P(Y_n = j)}{P(O)}
\]

Majority vote (2)

\[
Y_n = \arg \max_j \frac{\prod^M_{m=1} P(O_{m,n} = a_{m,n} | Y_n = j) P(Y_n = j)}{P(O)}
\]

\[
Y_n = \arg \max_j \prod^M_{m=1} P(O_{m,n} = a_{m,n} | Y_n = j)
\]

\[
Y_n = \arg \max_j (1 - \epsilon) \sum^M_{m=1} h_{a_{m,n},j} - \epsilon \sum^M_{m=1} h_{a_{m,n},j}
\]

Agree

Disagree
Hidden factors

Task Characteristics
- Quality (e.g., blurry pictures)
- Difficulty (e.g., transcription of non-native speech)

Worker Characteristics
- Expertise (e.g., bird identification)
- Bias (e.g., mother vs college students)
- Physical Conditions (e.g., fatigue)

Incorporating worker quality

Observed output

Worker Characteristics

True answer

Example: Medical classification

Objective: Medical diagnosis by doctors

Model: Doctors have different rates and types of errors.
- $\pi_{jl}^{(k)}$ defines the probability of doctor $k$ to record a value $l$ when the true state is $j$.
- $\eta_{il}^{(k)}$ is the number of times the clinician $k$ gets responses $I$ from patient $i$.

Solution: Expectation-Maximization (EM) Algorithm.
(Dawid & Stone, 1979)

EM Algorithm in a nutshell

Goal: Maximize the likelihood

$$p(data \ on \ patient \ i) \propto \sum_{j=1}^{J} \prod_{k=1}^{K} \pi_{jl}^{(k)} \eta_{il}^{(k)}$$

- $\pi_{jl}^{(k)}$: probability of doctor $k$ to record a value $l$ when the true state is $j$.
- $\eta_{il}^{(k)}$: number of times the clinician $k$ gets responses $I$ from patient $i$.
- $p_j$: probability that a patient drawn at random has true response $j$.

We can calculate the maximum likelihood estimates:

$$\hat{\pi}_{jl}^{(k)} = \frac{\sum_{T_{ij} = 1} T_{ij} \eta_{il}^{(k)}}{\sum_{I} \sum_{j} T_{ij} \eta_{il}^{(k)}}$$

And estimate the probabilities:

$$\hat{p}_j = \frac{\sum_{T_{ij} = 1} I_j}{I}$$

Where $T_i$ is the set of indicators ($T_{ij} = 1$ if $j$ is the true response and 0 otherwise).
**EM in a nutshell**

1. Take initial estimates of the T's.
2. Compute \( n \)'s and \( p \)'s using previous equations.
3. Use estimates of \( n \)'s and \( p \)'s to compute new T's s.t.
   \[
   p(T_j = 1 \mid data) = \prod_{k=1}^{K} \left( \sum_{l=1}^{J} \prod_{q=1}^{Q} \left( \pi_{ql}^{(k)} \right) \eta_{il}^{(k)} p_j^{(l)} \right) \]
4. Iterate 2. and 3. until convergence.

**Incorporating task difficulty**

<table>
<thead>
<tr>
<th>Task difficulty</th>
<th>Observed output</th>
<th>Worker Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_n )</td>
<td>( O_{nm} )</td>
<td>( \pi_m )</td>
</tr>
<tr>
<td>( Y_n )</td>
<td>( \pi_m )</td>
<td>( O_{nm} )</td>
</tr>
</tbody>
</table>

**Example**

**HIT:** Select images containing at least one "duck"

- Competence varies with bird image,
- Worker's bias toward various mistakes,
- Difficulty of the image.

(Weinler et al., 2010)

**Learning from imperfect data**

So far we assumed that:
- The system can distribute HIT to many unique workers
- The worker performs enough task to estimate their performance rates.

Problem: Not always true…

Dekel & Shamir (2009):
1. Train classifier on unfiltered data
2. Use learned hypothesis to "guess" the truth and use it to remove bad workers.

**Learning from Imperfect data**

Other source of errors:
- Inaccurate labels
  - Modeling reliability of multiple workers
- Redundant labels
  - Synonym resolution techniques
Beyond labels

More sophisticated outputs:
- Ranking
- Clustering
- Beliefs

Challenges:
- How to decompose the problem?
- How to aggregate the results?

Ranking & voting

1. Pairwise comparison & creation of a ranking.

2. (When possible) Rank all objects & compute a consensus ranking.

Clustering

Objective: minimizing disagreement.

Techniques:
- HIT link or separate object & cluster are computed from these properties.
- HIT link object to predefined sets and consensus is performed.

Prediction markets

Hypothesis: Knowledge is distributed and can be accessed by aggregating the belief of many individuals.

Technique: Workers report their belief and we estimate the probability that an event happen based on mean & median.

References

Human Computation
Edith Law, Luis von Ahn
Morgan & Claypool Publishers