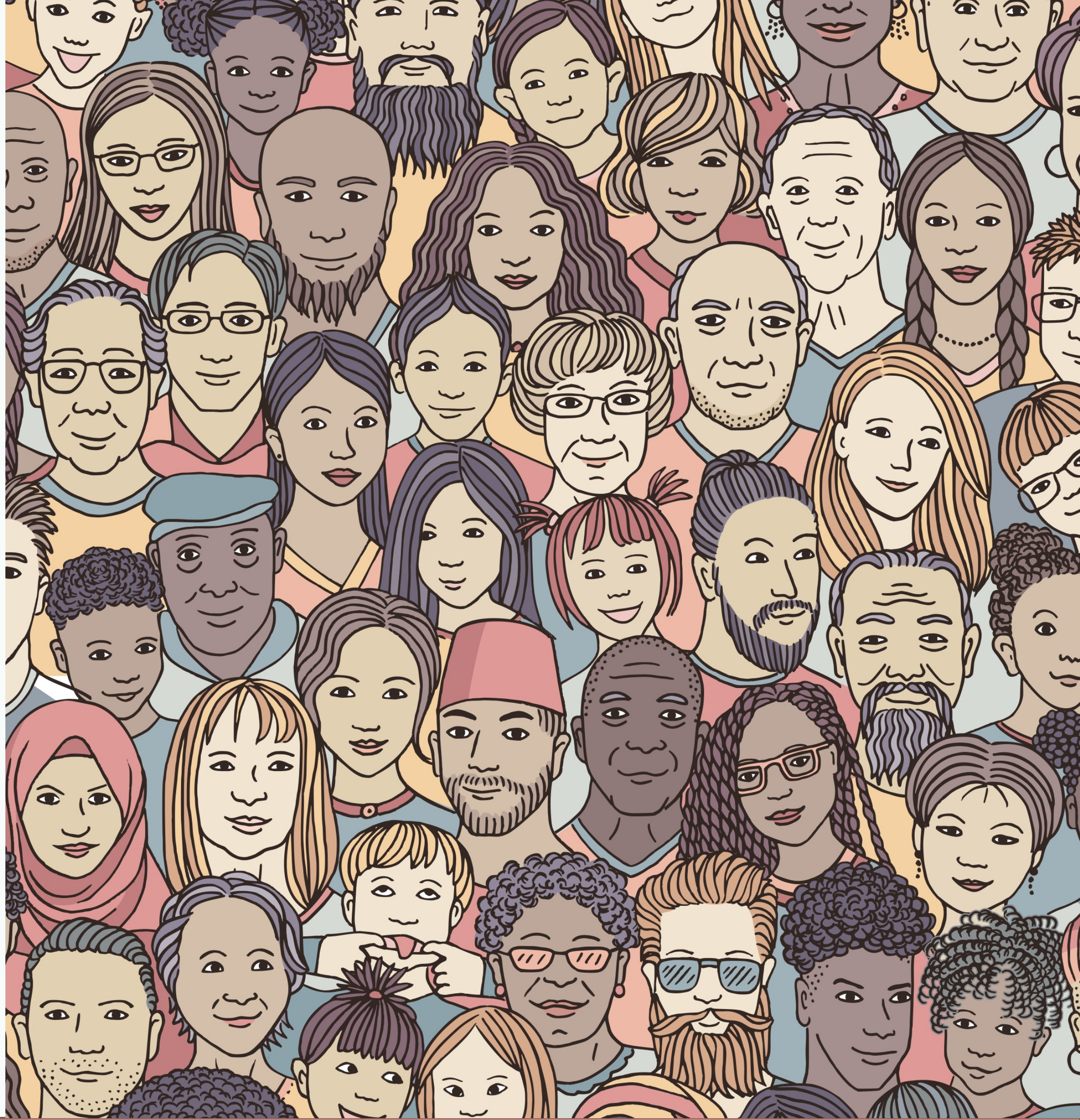
NETS 213: CROWDSOURCING AND HUMAN COMPUTATION

Quality Control part 3



Penn Engineering
UNIVERSITY of PENNSYLVANIA

Different Mechanisms for Quality Control

Aggregation and redundancy

Embedded gold standard data

Economic incentives

Reputation systems

Statistical models

Expectation Maximization algorithm

EM is an algorithm for finding the probabilities of unobserved variables

We will use it to estimate how accurate workers' labels are, and infer how good each worker is

This is more sophisticated than voting

Dawid and Skene (1977)

Maximum Likelihood Estimation of Observer Error-rates using the EM Algorithm

Examined application to medical diagnosis

Patients are sometimes treated by multiple physicians, who can give different diagnoses

Why? Doctors may has different questions. Patient may describe history differently. Doctors may classify symptoms differently

Observer Error

Given that different doctors have different opinions, they can’t all be right.

How often do individual physicians suffer from “observer error”? Are their errors systematic?

Answers depend on the “true” diagnosis.

Observer Error

Observer error would be easy to calculate if we had ground truth

Simply count the misdiagnoses and divide by the total number of diagnoses

However, sometimes it is impossible to know what diagnosis is correct. Same set of symptoms can arise from multiple root causes.

“I know it when I see it”

I shall not today attempt further to define "hard-core pornography"; and perhaps I could never succeed in intelligibly doing so. But I know it when I see it.
—Justice Potter Stewart

url	worker 1	worker 2	worker 3	worker 4	worker 5
sunnyfun.com	porn	not	not	not	porn
sex-mission.com	porn	porn	porn	porn	porn
google.com	not	porn	not	not	porn
youporn.com	porn	porn	porn	porn	not
yahoo.com	porn	not	not	not	porn

Solution?

Can't have Justice Stewart rule on everything

Instead, we will apply Dawid and Skene's EM algorithm, which iteratively

1. Estimates the correct answers, using labels from multiple workers, and accounts for the quality of each worker
2. Estimates the quality of the workers by comparing the submitted answers to the inferred correct answers

Inputs

a set of ***N*** objects $o_1 \ldots o_N$
sunnyfun.com, sex-mission.com, google.com, youporn.com, yahoo.com

a set of ***L*** possible labels:
{porn, not porn}

Labels for each object by ***K*** workers
worker1, worker2, worker3, worker4, worker5

Goal 1

Recover the true class label $T(o_n)$ for each object o_n when “gold” truth is unknown

Since the true labels are not known / never directly observed, they are called ***latent*** variables

Goal 2

For each worker who contributed labels, calculate their accuracy or reliability

To calculate accuracy show how often they mistakenly choose one label when a different one is the actual truth

Chicken and egg problem

If we knew what the **true class labels** were for each object for each object, then we could compute each Turker's accuracy

If we had **accuracies for every Turker**, then we could infer what the true label for each object should be

Input: Labels $l[k][n]$ from worker (k) to object o_n,

Output: Confusion matrix $\pi_{ij}^{(k)}$ for each worker (k), Correct labels $T(o_n)$ for each object o_n, Class priors $Pr\{C\}$ for each class C

```
Initialize error rates π_ij^(k) for each worker (k) (e.g., assume each worker is perfect);
Initialize correct label for each object T(o_n) (e.g., using majority vote);
while not converged do
    Estimate the correct label T(o_n) for each object, using the labels l[·][n] assigned to o_n by workers, weighting the votes using the error rates π_ij^(k);
    Estimate the error rates π_ij^(k), for each worker (k), using the correct labels T(o_n) and the assigned labels l[k][n];
    Estimate the class priors Pr{C}, for each class C;
end
return Estimated error rates π_ij^(k), Estimated correct labels T(o_n), Estimated class priors Pr{C}
```

Algorithm 1: The EM algorithm for worker quality estimation.

	worker1	worker2	worker3	worker4	worker5
sunnyfun.com	porn	not	not	not	porn
sex-mission.com	porn	porn	porn	porn	porn
google.com	not	porn	not	not	porn
youporn.com	porn	porn	porn	porn	not
yahoo.com	porn	not	not	not	porn

Output: “True” Labels

url	True Labels
sunnyfun.com	not
sex-mission.com	porn
google.com	not
youporn.com	porn
yahoo.com	not

Repeat until convergence

You can continue to iterate until your values converge

For this example, we converge after the first iteration

	worker1	worker2	worker3	worker4	worker5
sunnyfun.com	porn	not	not	not	porn
sex-mission.com	porn	porn	porn	porn	porn
google.com	not	porn	not	not	porn
youporn.com	porn	porn	porn	porn	not
yahoo.com	porn	not	not	not	porn

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

	porn	not
sunnyfun	?	?
sex-mission	?	?
google	?	?
youporn	?	?
yahoo	?	?

worker1	porn	not
porn	?	?
not	?	?

worker2	porn	not
porn	?	?
not	?	?

worker3	porn	not
porn	?	?
not	?	?

worker4	porn	not
porn	?	?
not	?	?

worker5	porn	not
porn	?	?
not	?	?

Initialize confusion matrices to uniform

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

	porn	not
sunnyfun	?	?
sex-mission	?	?
google	?	?
youporn	?	?
yahoo	?	?

worker1	porn	not
porn	1	0
not	0	1

worker2	porn	not
porn	1	0
not	0	1

worker3	porn	not
porn	1	0
not	0	1

worker4	porn	not
porn	1	0
not	0	1

worker5	porn	not
porn	1	0
not	0	1

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Compute labels using majority vote

	porn	not
sunnyfun	0.4	?
sex-mission	?	?
google	?	?
youporn	?	?
yahoo	?	?

worker1	porn	not
porn	1	0
not	0	1

worker2	porn	not
porn	1	0
not	0	1

worker3	porn	not
porn	1	0
not	0	1

worker4	porn	not
porn	1	0
not	0	1

worker5	porn	not
porn	1	0
not	0	1

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Compute labels using majority vote

	porn	not
sunnyfun	0.4	0.6
sex-mission	?	?
google	?	?
youporn	?	?
yahoo	?	?

worker1	porn	not
porn	1	0
not	0	1

worker2	porn	not
porn	1	0
not	0	1

worker3	porn	not
porn	1	0
not	0	1

worker4	porn	not
porn	1	0
not	0	1

worker5	porn	not
porn	1	0
not	0	1

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Normalize label probabilities

	porn	not
sunnyfun	0.4	0.6
sex-mission	1	0
google	0.4	0.6
youporn	0.8	0.2
yahoo	0.4	0.6

worker1	porn	not
porn	1	0
not	0	1

worker2	porn	not
porn	1	0
not	0	1

worker3	porn	not
porn	1	0
not	0	1

worker4	porn	not
porn	1	0
not	0	1

worker5	porn	not
porn	1	0
not	0	1

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Recompute worker confusion matrices

	porn	not
sunnyfun	0.4	0.6
sex-mission	1	0
google	0.4	0.6
youporn	0.8	0.2
yahoo	0.4	0.6

worker1	porn	not
porn	0	0
not	0	0

worker2	porn	not
porn	0	0
not	0	0

worker3	porn	not
porn	0	0
not	0	0

worker4	porn	not
porn	0	0
not	0	0

worker5	porn	not
porn	0	0
not	0	0

	worker1	worker2	worker3	worker4	worker5
sunnyfun	**porn**	not	not	not	porn
sex-mission	**porn**	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	**porn**	porn	porn	porn	not
yahoo	**porn**	not	not	not	porn

Recompute worker confusion matrices

	porn	not
sunnyfun	**0.4**	0.6
sex-mission	**1**	0
google	0.4	0.6
youporn	**0.8**	0.2
yahoo	**0.4**	0.6

worker1	**porn**	not
porn	**2.6**	0
not	0	0

worker2	porn	not
porn	0	0
not	0	0

worker3	porn	not
porn	0	0
not	0	0

worker4	porn	not
porn	0	0
not	0	0

worker5	porn	not
porn	0	0
not	0	0

	worker1	worker2	worker3	worker4	worker5
sunnyfun	**porn**	not	not	not	porn
sex-mission	**porn**	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	**porn**	porn	porn	porn	not
yahoo	**porn**	not	not	not	porn

Recompute worker confusion matrices

	porn	not
sunnyfun	0.4	**0.6**
sex-mission	1	**0**
google	0.4	0.6
youporn	0.8	**0.2**
yahoo	0.4	**0.6**

worker1	**porn**	not
porn	2.6	0
not	**1.4**	0

worker2	porn	not
porn	0	0
not	0	0

worker3	porn	not
porn	0	0
not	0	0

worker4	porn	not
porn	0	0
not	0	0

worker5	porn	not
porn	0	0
not	0	0

Recompute worker confusion matrices

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	**not**	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

	porn	not
sunnyfun	0.4	0.6
sex-mission	1	0
google	**0.4**	0.6
youporn	0.8	0.2
yahoo	0.4	0.6

worker1	porn	**not**
porn	2.6	**0.4**
not	1.4	0

worker2	porn	not
porn	0	0
not	0	0

worker3	porn	not
porn	0	0
not	0	0

worker4	porn	not
porn	0	0
not	0	0

worker5	porn	not
porn	0	0
not	0	0

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	**not**	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Recompute worker confusion matrices

	porn	not
sunnyfun	0.4	0.6
sex-mission	1	0
google	0.4	**0.6**
youporn	0.8	0.2
yahoo	0.4	0.6

worker1	porn	**not**
porn	2.6	0.4
not	1.4	**0.6**

worker2	porn	not
porn	0	0
not	0	0

worker3	porn	not
porn	0	0
not	0	0

worker4	porn	not
porn	0	0
not	0	0

worker5	porn	not
porn	0	0
not	0	0

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Renormalize confusion matrices
(based on true labels)

	porn	not
sunnyfun	0.4	0.6
sex-mission	1	0
google	0.4	0.6
youporn	0.8	0.2
yahoo	0.4	0.6

worker1	porn	not
porn	0.87	0.13
not	0.70	0.30

worker2	porn	not
porn	0	0
not	0	0

worker3	porn	not
porn	0	0
not	0	0

worker4	porn	not
porn	0	0
not	0	0

worker5	porn	not
porn	0	0
not	0	0

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Renormalize confusion matrices
(based on true labels)

	porn	not
sunnyfun	0.4	0.6
sex-mission	1	0
google	0.4	0.6
youporn	0.8	0.2
yahoo	0.4	0.6

worker1	porn	not
porn	0.87	0.13
not	0.70	0.30

worker2	porn	not
porn	0.69	0.31
not	0.44	0.56

worker3	porn	not
porn	0.60	0.40
not	0.10	0.90

worker4	porn	not
porn	0.60	0.40
not	0.10	0.90

worker5	porn	not
porn	0.73	0.27
not	0.90	0.10

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Recompute labels using weighted majority vote

	porn	not
sunnyfun	**2.71**	
sex-mission		
google		
youporn		
yahoo		

worker1	porn	not
porn	**0.87**	0.13
not	0.70	0.30

worker2	porn	not
porn	0.69	**0.31**
not	0.44	0.56

worker3	porn	not
porn	0.60	**0.40**
not	0.10	0.90

worker4	porn	not
porn	0.60	**0.40**
not	0.10	0.90

worker5	porn	not
porn	**0.73**	0.27
not	0.90	0.10

Recompute labels using weighted majority vote

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

	porn	not
sunnyfun	2.71	**3.96**
sex-mission		
google		
youporn		
yahoo		

worker1	porn	not
porn	0.87	0.13
not	**0.70**	0.30

worker2	porn	not
porn	0.69	0.31
not	0.44	**0.56**

worker3	porn	not
porn	0.60	0.40
not	0.10	**0.90**

worker4	porn	not
porn	0.60	0.40
not	0.10	**0.90**

worker5	porn	not
porn	0.73	0.27
not	**0.90**	0.10

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Renormalize label probabilities

	porn	not
sunnyfun	0.41	0.59
sex-mission	0.61	0.39
google	0.41	0.59
youporn	0.68	0.32
yahoo	0.41	0.59

worker1	porn	not
porn	0.87	0.13
not	0.70	0.30

worker2	porn	not
porn	0.69	0.31
not	0.44	0.56

worker3	porn	not
porn	0.60	0.40
not	0.10	0.90

worker4	porn	not
porn	0.60	0.40
not	0.10	0.90

worker5	porn	not
porn	0.73	0.27
not	0.90	0.10

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Iterate until no more boxes on the questionnaire

	porn	not
sunnyfun	0.45	0.55
sex-mission	0.54	0.46
google	0.45	0.55
youporn	0.59	0.41
yahoo	0.45	0.55

worker1	porn	not
porn	0.84	0.16
not	0.76	0.24

worker2	porn	not
porn	0.68	0.32
not	0.52	0.48

worker3	porn	not
porn	0.51	0.49
not	0.29	0.71

worker4	porn	not
porn	0.51	0.49
not	0.29	0.71

worker5	porn	not
porn	0.73	0.27
not	0.87	0.13

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Iterate until...$&@!#!
Whatever, just turn it in...

	porn	not
sunnyfun	0.5	0.5
sex-mission	0.5	0.5
google	0.5	0.5
youporn	0.5	0.5
yahoo	0.5	0.5

worker1	porn	not
porn	0.80	0.20
not	0.80	0.20

worker2	porn	not
porn	0.60	0.40
not	0.60	0.40

worker3	porn	not
porn	0.40	0.60
not	0.40	0.60

worker4	porn	not
porn	0.40	0.60
not	0.40	0.60

worker5	porn	not
porn	0.80	0.20
not	0.80	0.20

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Recompute worker confusion matrices

	porn	not
sunnyfun	0.4	0.6
sex-mission	1	0
google	0.4	0.6
youporn	0.8	0.2
yahoo	0.4	0.6

worker1	porn	not
porn	0	0
not	0	0

worker2	porn	not
porn	0	0
not	0	0

worker3	porn	not
porn	0	0
not	0	0

worker4	porn	not
porn	0	0
not	0	0

worker5	porn	not
porn	0	0
not	0	0

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Recompute worker confusion matrices…as though your labels are 100% correct

	porn	not
sunnyfun	0	1
sex-mission	1	0
google	0	1
youporn	1	0
yahoo	0	1

worker1	porn	not
porn	0	0
not	0	0

worker2	porn	not
porn	0	0
not	0	0

worker3	porn	not
porn	0	0
not	0	0

worker4	porn	not
porn	0	0
not	0	0

worker5	porn	not
porn	0	0
not	0	0

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	**porn**	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Recompute worker confusion matrices...as though your labels are 100% correct

	porn	not
sunnyfun	0	1
sex-mission	**1**	0
google	0	1
youporn	**1**	0
yahoo	0	1

worker1	**porn**	not
porn	**2**	0
not	0	0

worker2	porn	not
porn	0	0
not	0	0

worker3	porn	not
porn	0	0
not	0	0

worker4	porn	not
porn	0	0
not	0	0

worker5	porn	not
porn	0	0
not	0	0

	worker1	worker2	worker3	worker4	worker5
sunnyfun	**porn**	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	**porn**	not	not	not	porn

Recompute worker confusion matrices…as though your labels are 100% correct

	porn	not
sunnyfun	0	**1**
sex-mission	1	0
google	0	1
youporn	1	0
yahoo	0	**1**

worker1	**porn**	not
porn	2	0
not	**2**	0

worker2	porn	not
porn	0	0
not	0	0

worker3	porn	not
porn	0	0
not	0	0

worker4	porn	not
porn	0	0
not	0	0

worker5	porn	not
porn	0	0
not	0	0

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Recompute worker confusion matrices...as though your labels are 100% correct

	porn	not
sunnyfun	0	1
sex-mission	1	0
google	0	1
youporn	1	0
yahoo	0	1

worker1	porn	**not**
porn	2	**0**
not	2	0

worker2	porn	not
porn	0	0
not	0	0

worker3	porn	not
porn	0	0
not	0	0

worker4	porn	not
porn	0	0
not	0	0

worker5	porn	not
porn	0	0
not	0	0

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Recompute worker confusion matrices…as though your labels are 100% correct

	porn	not
sunnyfun	0	1
sex-mission	1	0
google	0	**1**
youporn	1	0
yahoo	0	1

worker1	porn	**not**
porn	2	0
not	2	**1**

worker2	porn	not
porn	0	0
not	0	0

worker3	porn	not
porn	0	0
not	0	0

worker4	porn	not
porn	0	0
not	0	0

worker5	porn	not
porn	0	0
not	0	0

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Renormalize confusion matrices
(based on true labels)

	porn	not
sunnyfun	0	1
sex-mission	1	0
google	0	1
youporn	1	0
yahoo	0	1

worker1	porn	not
porn	1	0
not	0.67	0.33

worker2	porn	not
porn	0	0
not	0	0

worker3	porn	not
porn	0	0
not	0	0

worker4	porn	not
porn	0	0
not	0	0

worker5	porn	not
porn	0	0
not	0	0

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Renormalize confusion matrices
(based on true labels)

	porn	not
sunnyfun	0	1
sex-mission	1	0
google	0	1
youporn	1	0
yahoo	0	1

worker1	porn	not
porn	1	0
not	0.67	0.33

worker2	porn	not
porn	1	0
not	0.33	0.67

worker3	porn	not
porn	1	0
not	0	1

worker4	porn	not
porn	1	0
not	0	1

worker5	porn	not
porn	0.5	0.5
not	1	0

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Recompute labels using weighted majority vote

	porn	not
sunnyfun	**1.5**	
sex-mission		
google		
youporn		
yahoo		

worker1	porn	not
porn	**1**	0
not	0.67	0.33

worker2	porn	not
porn	1	**0**
not	0.33	0.67

worker3	porn	not
porn	1	**0**
not	0	1

worker4	porn	not
porn	1	**0**
not	0	1

worker5	porn	not
porn	**0.5**	0.5
not	1	0

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Recompute labels using weighted majority vote

	porn	not
sunnyfun	1.5	**4.34**
sex-mission		
google		
youporn		
yahoo		

worker1	porn	not
porn	1	0
not	**0.67**	0.33

worker2	porn	not
porn	1	0
not	0.33	**0.67**

worker3	porn	not
porn	1	0
not	0	**1**

worker4	porn	not
porn	1	0
not	0	**1**

worker5	porn	not
porn	0.5	0.5
not	**1**	0

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Recompute labels using weighted majority vote

	porn	not
sunnyfun	0.26	0.74
sex-mission	0.69	0.31
google	0.29	0.71
youporn	0.82	0.18
yahoo	0.26	0.74

worker1	porn	not
porn	1	0
not	0.67	0.33

worker2	porn	not
porn	1	0
not	0.33	0.67

worker3	porn	not
porn	1	0
not	0	1

worker4	porn	not
porn	1	0
not	0	1

worker5	porn	not
porn	0.5	0.5
not	1	0

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Recompute worker confusion matrices...as though your labels are 100% correct

	porn	not
sunnyfun	0	1
sex-mission	1	0
google	0	1
youporn	1	0
yahoo	0	1

worker1	porn	not
porn	1	0
not	0.67	0.33

worker2	porn	not
porn	1	0
not	0.33	0.67

worker3	porn	not
porn	1	0
not	0	1

worker4	porn	not
porn	1	0
not	0	1

worker5	porn	not
porn	0.5	0.5
not	1	0

	worker1	worker2	worker3	worker4	worker5
sunnyfun	porn	not	not	not	porn
sex-mission	porn	porn	porn	porn	porn
google	not	porn	not	not	porn
youporn	porn	porn	porn	porn	not
yahoo	porn	not	not	not	porn

Iterate forever...because now all is well and good in the world.

	porn	not
sunnyfun	0.26	0.74
sex-mission	0.69	0.31
google	0.29	0.71
youporn	0.82	0.18
yahoo	0.26	0.74

worker1	porn	not
porn	1	0
not	0.67	0.33

worker2	porn	not
porn	1	0
not	0.33	0.67

worker3	porn	not
porn	1	0
not	0	1

worker4	porn	not
porn	1	0
not	0	1

worker5	porn	not
porn	0.5	0.5
not	1	0

Question

How would you use gold standard data in the EM process?

EM Algorithm

Re-Calculate Worker Scores over two steps:

1. Estimate the probability that each answer is correct, using labels from multiple workers weighted by the probability that they are correct
2. Estimate the quality of the workers by comparing their submitted answers to the inferred correct answers

Confusion Matrix gives us worker error

From the confusion matrix we can measure the overall error rate for each worker

Sum of the non-diagonal elements of the confusion matrix (weighted by the priors)

This results in a single, scalar value as the quality score for each worker

Worker Error

w1	porn	not porn
porn	100%	0%
not porn	100%	0%

100

w3	porn	not porn
porn	100%	0%
not porn	0%	100%

0

w2	porn	not porn
porn	100%	0%
not porn	33%	67%

33

w4	porn	not porn
porn	0%	100%
not porn	100%	0%

200

Is worker5 the worst?

url	worker1	worker2	worker3	worker4	worker5
google.com	porn	not porn	not porn	not porn	porn
panda-cam.gov	porn	porn	not porn	not porn	porn
sex-mission.com	porn	porn	porn	porn	not porn
sunnyfun.com	porn	not porn	not porn	not porn	porn
youporn.com	porn	porn	porn	porn	not porn

Advanced Topics

Bias versus error

How noisy can the workers be and still allow us to still converge to a correct solution?

Bias versus Error

Error rate alone is not sufficient to measure the inherent value of a worker.

For example, workers may be careful but biased

In a non-binary case, this is more apparent

What if instead of asking our workers to label sites porn or not porn, we asked them to label the G, PG, R, X?

Bias versus Error

Parents with young children tend to be more conservative

They tend to classify PG-rated sites as R-rated sites, and R-rated sites as X-rated.

Such workers give consistently and predictably incorrect answers

It is possible to automatically correct for bias

Implications

Unlike with spammers, with biased workers it is possible to “reverse” the errors

We can recover a label assignment of much higher quality

In the presence of systematic bias, the naive measurement of error rate results in underestimates of the true quality of the worker

This potentially leads to incorrect rejections and blocks of legitimate workers

For more details

Check out two papers by Panos Ipeirotis and his collaborators

Managing Crowdsourcing Workers discusses separating error and bias

Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers discusses how noisy judgements can be, with us still getting good quality results