Uncertainty in Crowd Data Sourcing under Structural Constraints

Antoine Amarilli[1,2] Yael Amsterdamer[1] Tova Milo[1]

[1]Tel Aviv University, Tel Aviv, Israel

[2]Télécom ParisTech, Paris, France

April 21, 2014

Crowd data sourcing

- Crowdsourcing: reducing hard problems to elementary queries asked to an indiscriminate crowd of human users
- Crowd data sourcing: extracting knowledge from the crowd

⇒ *Would you recommend this restaurant for Indian food?*

⇒ *What is the topic of the following text?*

⇒ *Which of these designs seems neater to you?*

Answers are uncertain

- Crowd answers are noisy!
- *How would you rate the quality of this sound file?*
  - ⇒ 8/10
  - ⇒ 7/10
  - ⇒ 5/10 *(didn't actually listen)*
  - ⇒ 1/10 *(has lousy headphones)*
  - ⇒ 10/10 *(has poor taste)*
- Truth finding approaches but still different tastes

⇒ We are interested in the average answer

Problem statement

- We have a bunch of questions
  - ⇒ *What is the quality of file i?*
- We want to be efficient
  - ⇒ Don't ask too many questions
  - ⇒ Compute quickly the next question to ask
- We have an overall objective
  - ⇒ *Which file has average quality rating closest to 7/10?*

⇒ How to choose our next question?

Table of contents

Crowd model

- For each question i, a random variable X_i to model answers
- Asking a question means getting a observation
  - ⇒ *User gave grade 4/10 to file i*
- Our desired answer is the unknown mean of X_i
  - ⇒ *Average user grade for file i*
- Objective: minimize the loss of our current prediction
- Overall loss is a sum of each question's loss
  - ⇒ *How many files are misclassified w.r.t. the threshold 7/10*

Normal variables

So far, the questions are independent. Consider file i:

- We have already obtained answers S
- We assume the random variable X_i is Gaussian
- Unknown parameters of X_i
  - ⇒ Mean μ (desired answer)
  - ⇒ Variance σ^2

Maximum Likelihood Estimation

- Maximum likelihood estimator $(\widehat{\mu}, \widehat{\sigma}^2)$ for S:
  - $\Rightarrow$ $\widehat{\mu}$ is the sample mean
  - $\Rightarrow$ $\widehat{\sigma}^2$ is the sample variance
  - $\Rightarrow$ Those parameters give the highest probability to S
- Example: answers $S = \{7/10, 9/10\}$
  - $\Rightarrow$ $\widehat{\mu} = 8/10$

Error estimation

- Assume that our guess $(\widehat{\mu}, \widehat{\sigma}^2)$ is the truth
- Consider which answers we could have obtained:
  How often would we still believe $(\widehat{\mu}, \widehat{\sigma}^2)$?
  - ⇒ Say we see answers $S = \{1/10, 9/10\}$
  - ⇒ $\widehat{\mu} = 5/10$ and high $\widehat{\sigma}^2$
  - ⇒ Under $(\widehat{\mu}, \widehat{\sigma}^2)$ we could have seen $S' = \{2/10, 3/10\}$
  - ⇒ We would have guessed $(\widehat{\mu}, \widehat{\sigma}^2)$ differently then
- Formally: expected loss of the MLE for outcomes under the estimated distribution according to the computed MLE.

Best error decrease

- We can estimate our error…
- … but how much does one more answer help?
- Our predicted $(\widehat{\mu}, \widehat{\sigma}^2)$ tells us which answers to expect
- We can compute a new error estimation for each answer

⇒ Average error decrease, under the estimated distribution

Overall, we should ask the question with the highest decrease.

Table of contents

Order on numerical answers

- The previous approach assumes independent variables
- Sometimes, they are ordered
  - ⇒ Sound file quality with various compression levels
  - ⇒ Target price for various deals (flight, flight and hotel)
  - ⇒ Frequency of activity combinations (beach, beach and surfing)
- Order on true answers but not on our observations!
  - User A rates lossless with $6/10$
  - User B rates high compression with $8/10$
  - ⇒ Monotonicity only on the mean values!

Joint distribution and MLE

- We assume normal distributions
- Parameters (μ_i, σ_i^2) for each variable
- Assumption $\mu_1 < \mu_2 < \cdots < \mu_n$
- What are the most likely parameters in this space?

$\Rightarrow$ No obvious closed form for the MLE

Approximating the MLE

- Approximation: first determine the mean values
  - ⇒ Enforce the monotonicity constraint
  - ⇒ Remain close to the sample mean of each variable...
  - ⇒ ... depending on the sample variances

⇒ Least squares under linear inequalities: quadratic programming

⇒ Then readjust the variances based on those means

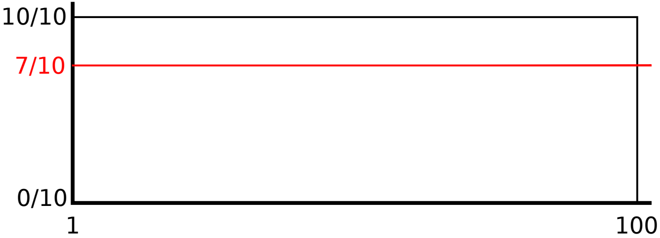
Estimated error and error decrease like before (but for all variables).

Table of contents

Interpolating variables

- We have a large collection of totally ordered variables
  - ⇒ e.g., 100 possible bitrate levels
- We want to find a threshold value
  - ⇒ Which is the strongest compression with quality $\geq 7/10$?
- We cannot ask questions about all variables

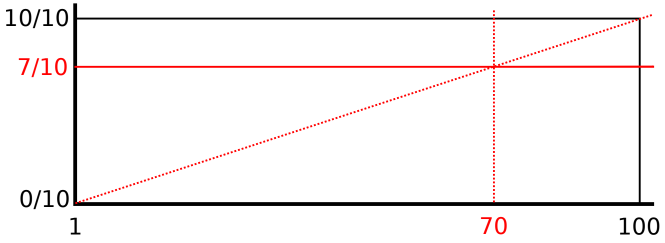
⇒ Under exact answers: interpolation search



Interpolating variables

- We have a large collection of totally ordered variables
  - ⇒ e.g., 100 possible bitrate levels
- We want to find a threshold value
  - ⇒ Which is the strongest compression with quality $\geq 7/10$?
- We cannot ask questions about all variables

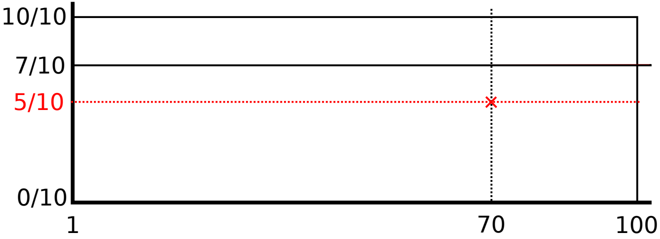
⇒ Under exact answers: interpolation search



Interpolating variables

- We have a large collection of totally ordered variables
  - ⇒ e.g., 100 possible bitrate levels
- We want to find a threshold value
  - ⇒ Which is the strongest compression with quality $\geq 7/10$?
- We cannot ask questions about all variables

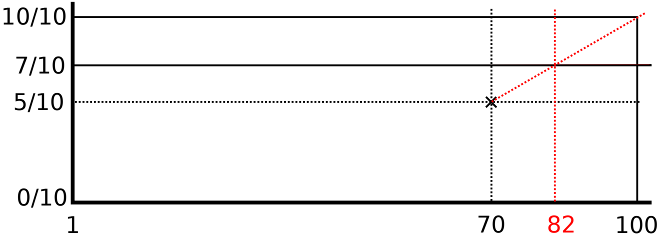
⇒ Under exact answers: interpolation search



Interpolating variables

- We have a large collection of totally ordered variables
  - ⇒ e.g., 100 possible bitrate levels
- We want to find a threshold value
  - ⇒ Which is the strongest compression with quality $\geq 7/10$?
- We cannot ask questions about all variables

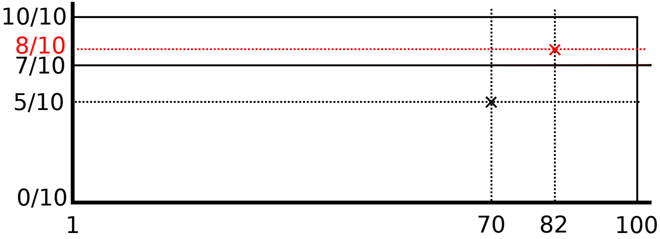
⇒ Under exact answers: interpolation search



Interpolating variables

- We have a large collection of totally ordered variables
  - ⇒ e.g., 100 possible bitrate levels
- We want to find a threshold value
  - ⇒ Which is the strongest compression with quality $\geq 7/10$?
- We cannot ask questions about all variables

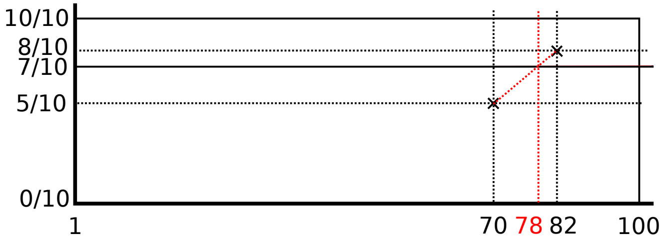
⇒ Under exact answers: interpolation search



Interpolating variables

- We have a large collection of totally ordered variables
  - ⇒ e.g., 100 possible bitrate levels
- We want to find a threshold value
  - ⇒ Which is the strongest compression with quality $\geq 7/10$?
- We cannot ask questions about all variables

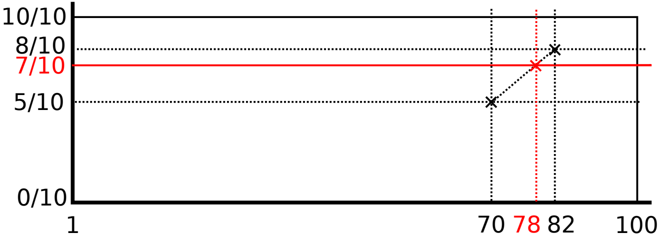
⇒ Under exact answers: interpolation search



Interpolating variables

- We have a large collection of totally ordered variables
  - ⇒ e.g., 100 possible bitrate levels
- We want to find a threshold value
  - ⇒ Which is the strongest compression with quality $\geq 7/10$?
- We cannot ask questions about all variables

⇒ Under exact answers: interpolation search



Interpolation issues

- Linear interpolation for the means
- Which interpolation for the variances?
  - ⇒ Variance from the neighboring points
  - ⇒ Variance from the interpolation uncertainty
- Computing expected decrease for each point may be too slow!

Table of contents

Conclusion

- A general scheme to choose questions in crowd data sourcing
- A method to incorporate order constraints on the variables
- Ways to perform interpolation for questions with no answers
- Ongoing work:
  - ⇒ A general interpolation scheme for arbitrary partial orders
  - ⇒ Support for complex queries
  - ⇒ Other criteria to choose next question
  - ⇒ Experiments for activity recommendations

Conclusion

- A general scheme to choose questions in crowd data sourcing
- A method to incorporate order constraints on the variables
- Ways to perform interpolation for questions with no answers
- Ongoing work:
  - ⇒ A general interpolation scheme for arbitrary partial orders
  - ⇒ Support for complex queries
  - ⇒ Other criteria to choose next question
  - ⇒ Experiments for activity recommendations

Thanks for your attention!