



Human-centered Artificial Intelligence

2024/2025

Worksheet #4: Data Collection & Annotation

Hugo Oliveira, Luís Macedo

1.1 Topics

- The Data Mining Process
- Data Collection & Annotation
- Crowdsourcing

1.2 Pre-class Materials

- Course slides on Data Collection & Annotation.
- Video on typical Data Mining subproblems:
<https://www.youtube.com/watch?v=EH3bp5335IU> (11 min)
- Video on Data Annotation:
<https://www.youtube.com/watch?v=YJnnxitraac> (2 min)
- Video on “The Wisdom of the Crowd”:
<https://www.youtube.com/watch?v=i0ucwX7Z1HU> (5 min)
- Video on Human Annotation and ChatGPT:
https://www.youtube.com/watch?v=ug_p2wHh1a0 (3 min)
- Executive summary of [Schmidt \[2019\]](#)

1.3 Complementary Materials

- An Introduction to Data Mining [[Aggarwal, 2015](#)] (Chapter 1)
- Condensed report on Crowdsourcing for self-driving cars [[Schmidt, 2019](#)]
- News on ChatGPT and Kenyan workers: <https://time.com/6247678/openai-chatgpt-kenya-workers/https://time.com/6247678/openai-chatgpt-kenya-workers/>
- Book “The Wisdom of Crowds” [[Surowiecki, 2005](#)]
- Videos on Data Annotation and Crowdsourcing
 - <https://www.youtube.com/watch?v=hhzhamJUbmG>
 - <https://www.youtube.com/watch?v=EH3bp5335IU>
 - <https://www.youtube.com/watch?v=-38uPkyH9vI>

1.4 Theoretical-Practical Exercises

Question 1.1 Give and discuss an example where Data Mining is crucial to the success of a business. What Data Mining functionalities does this business need (e.g., think of the kinds of patterns that could be mined)? Can such patterns be generated alternatively by data query processing or simple statistical analysis?

Solution A department store, for example, can use DM to assist with its target marketing mail campaign. Using DM functions such as association, the store can use the mined strong association rules to determine which products bought by one group of customers are likely to lead to the buying of certain other products. With this information, the store can then mail marketing materials only to those kinds of customers who exhibit a high likelihood of purchasing additional products.

Data query processing is used for data or information retrieval and does not have the means for finding association rules, making predictions, etc. Similarly, simple statistical analysis cannot handle large amounts of data such as those of customer records in a department store.

Slides: 11–15 (Analytical Processing)

Question 1.2 To better understand their likes and dislikes, an analyst collects surveys from a sample of customers. Subsequently, the analyst uploads all the data to a database, corrects erroneous or missing entries, and designs a recommendation algorithm.

- a. Associate each of the following actions to the respective steps in the Data Mining process:
 - Conducting surveys
 - Uploading to database
 - Correcting missing entries
 - Designing a recommendation algorithm.

Solution Data Collection, Data Collection, Data Preprocessing, Analysis

Slides: (Data Mining Process, Data Collection, Preprocessing)

- b. Now suppose that the surveys are answered in free text, which is then read by a team of workers that manually extract positive and negative aspects, also to be included in the database. What step would this be?

Solution Data Annotation, possibly included in the pre-processing step.

Slides: (Data Annotation)

Question 1.3 Imagine that you are developing an application for optimising student academic performance and discuss:

- a. What kind of data would be useful for this purpose? For each kind, indicate its type.

Solution

- Age (numeric), gender (categorical), friends in social networks and information about them (numeric, ...), grades (numerical, for training and validation)
- Activity (?), social network posts, SMS, emails, voice, face, ...
- Current mood (time series)
- Time spent studying, sleeping, having fun, using smartphone / specific apps... (numeric, categorical, consecutive discrete)
- Light, noise, ... of the environment (time series, consecutive discrete)
- ...

Slides: 6–8 (Data Types)

- b. How could you extract such data? What would your data sources be?

Solution Surveys (daily?), school Information System, smartphone, other IoT sensors, ... text mining, including sentiment analysis from messages written

- c. What Data Mining problems would be involved?

Solution

- If there is historic data, classification (types of students, predicted performance, ...);
- Outlier detection;
- Clustering, e.g., for grouping users according to certain features;
- Pattern mining could be helpful for identifying “good” or “bad” patterns, i.e., associated with high or low grades or even moods.

Slides: 11–15 (Analytical Processing)

d. Regarding the data, what difficulties do you anticipate?

Solution Data protection and privacy laws, not enough training data, missing values (users do not answer the survey or only answer part, unavailable readings, ...), errors of Machine Learning models, dealing with many data sources, ...

Question 1.4 Discuss the pros and cons of Crowdsourcing and:

- a. Give examples of concrete tasks that it suits well and others that it does not. You may think of real situations where crowdsourcing is or has been used.

Solution When the solution is more or less obvious for a human but there is much data to annotate, or more subjective annotations that cannot be based on the opinion of a single person, ...

Challenge: ask everyone to turn their laptops off, then ask some question for which no one should know the answer (e.g., a year), compute the average, check how close it is from the real answer.

- * How many beans in the jar? Would have to count them in advance...
- * How many monthly listeners do the Beatles have on Spotify?
- * How many seconds does the song Bohemian Rhapsody have in total?
- * How many followers does Elon Musk have on X?

Slides: (Crowdsourcing)

- b. Enumerate reasons that should be weighted when opting for crowdsourcing.

Solution Available funds, how much it would cost to pay an expert, how easy it is to explain and to perform the task, how critical the results are, ... When firms receive too many ideas, they tend to focus on ideas that are already familiar to them, defeating the entire purpose of crowdsourcing, which is to surface new thinking.

- c. Refer important aspects that should be taken care of when opting for microwork.

Solution Task should be simple, direct, transmitted easily in words; even if the answer is subjective, the task should be transmitted as objectively as possible!

Users that take it seriously *vs* users that just want to do it quickly and get the reward, mechanisms to filter the latter, detect outliers, ...

Slide: 24 (Microwork)

- d. How could crowdsourcing be exploited in *The Price is Right* tv show? Could you think of limitations of answering with the linear average of audience guesses?

Solution Guesses above the right price are not accepted!

1.6

Question Concerning the development of self-driving cars and the Data Mining Process...

- a. What tasks do you think fit in the steps of Data Collection and Data Annotation?

Solution

- Collection: driving around, collecting all kinds of data, especially images (millions of images!)
- Annotation: images, object recognition (roads, signs, obstacles, other vehicles, pedestrians, ...): bounding boxes (roughly mark the position of individual objects) and semantic segmentation maps (maps every pixel in an image has to be covered by a descriptive label.).

- b. How well does Crowdsourcing suit this goal?

Solution

- It has been used intensively, but ... apparently, car manufacturers do not want to be associated with crowdsourcing!
- Annotations must be as detailed as possible (recognise objects but also predict how vehicles or people will behave in traffic), may take hours to complete: high degree of accuracy required: more training, higher payments and expertise, dedicated platforms (nomenclature: "AI training data production" instead of Crowdsourcing!)

- c. What is the global perception of human involvement in the development of LLM-based chatbots, including ChatGPT?

Solution Not sure... some people might think that an army of humans created a large set of rules; other will think it is just a 100% Machine Learning?

Bibliography

Charu C Aggarwal. *Data mining: the textbook*. Springer, 2015.

Florian Alexander Schmidt. Crowdsourced production of AI training data: How human workers teach self-driving cars how to see. Technical report, Working Paper Forschungsförderung, 2019. <https://www.econstor.eu/bitstream/10419/216075/1/hbs-fofoe-wp-155-2019.pdf>.

James Surowiecki. *The Wisdom of Crowds*. Anchor Books, 2005.