DataSheet For

WindmillOutputSmallDataset (from Pytorch-Geometric-Temporal library)

Introduction

This data sheet template is taken directly from Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Motivation

- **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
    - The dataset was created to be used for testing Pytorch-Geometric-Temporal models on node-level regression tasks.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
    - A research group (Rozemberczki et al.), who developed the Pytorch Geometric Temporal library, created the dataset.
- **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.
    - Could not find information about funding for this research group.
- **Any other comments?**
    - None.

Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

  - The instances represent hourly energy output of windmills from a European country for more than 2 years. Vertices, i.e. nodes, represent 11 windmills and weighted edges describe the strength of relationships between the windmills.

- **How many instances are there in total (of each type, if appropriate)?**
  - The number of data points is 17464.

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
  - The dataset is a subset of a larger dataset called "WindmillOutputLarge" which contains 319 nodes of windmills. The smaller dataset is more localized to a single region.

- **What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
  - Each instance consists of 8 features regarding hourly windmill energy.

- **Is there a label or target associated with each instance?** If so, please provide a description.
  - Each instance contains an edge index, i.e. id.

- **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
  - N/A (unknown)

- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.
    - The relationship between windmills (nodes) is made explicit through edge weight, which describes the distance (strength) between them.
- **Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.
    - On the paper, the example model used data splits of 90/10 for training and testing.
- **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.
    - N/A (unknown)
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
    - N/A (unknown)
- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor/patient confidentiality, data that includes the content of individuals' non-public communications)?** If so, please provide a description.
    - Hourly energy output does not include information about individuals. This is non-human data.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.
    - N/A
- **Does the dataset relate to people?** If not, you may skip the remaining questions in this section.
    - **Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
    - **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** If so, please describe how.
    - **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.
        - The dataset does not relate to people.
- **Any other comments?**

Collection Process

- **How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
    - Data collection method/process was not reported by the paper.

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** How were these mechanisms or procedures validated?
    - N/A (unknown)
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
    - The smaller dataset is more localized to a single region; therefore, deterministic.
- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**
    - N/A (unknown)
- **Over what timeframe was the data collected?** Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
    - The data was collected for 2 years.
- **Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
    - This dataset does not relate to people and hence does not require an institutional review board.
- **Does the dataset relate to people?** If not, you may skip the remainder of the questions in this section.
    - **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
    - **Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

    - **Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
    - **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
    - **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
        - The dataset does not relate to people.
- **Any other comments?**
    - None

Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remainder of the questions in this section.
    - N/A (unknow)
- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.
    - N/A (unknow)
- **Is the software used to preprocess/clean/label the instances available?** If so, please provide a link or other access point.

    - N/A (unknow)

- **Any other comments?**
    - None

Uses

- **Has the dataset been used for any tasks already?** If so, please provide a description.
    - The dataset has been used for testing the models of Pytorch Geometric Temporal library.
- **Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.
    - https://github.com/benedekrozemberczki/pytorch_geometric_temporal/blob/master/docs/source/modules/dataset.rst

- **What (other) tasks could the dataset be used for?**
    - The dataset could be used for regression tasks.
- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
    - N/A (unknown)
- **Are there tasks for which the dataset should not be used?** If so, please provide a description.
    - N/A (unknown)

- **Any other comments?**
    - None

Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.
    - The dataset is available through Pytorch Geometric Temporal library.
- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?
    - The dataset is distributed through API.
- **When will the dataset be distributed?**
    - The dataset is already available.
- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
    - © Copyright 2022, Benedek Rozemberczki. Revision e43e95d7.
- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
    - N/A (unknown)
- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

  - N/A (unknown)

- **Any other comments?**
  - None

Maintenance

- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
  - https://github.com/benedekrozemberczki/pytorch_geometric_temporal/blob/master/docs/source/modules/dataset.rst
  - Can be contacted by updating issues to the github repository.
- **Is there an erratum?** If so, please provide a link or other access point.
  - N/A (unknown)
- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
  - N/A (unknown)
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.
  - Not related to people.
- **Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
  - N/A (unknown)

- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
    - I don't believe it's possible since the raw dataset isn't available through API.
- **Any other comments?**
    - None.