

tabular data

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland

*Corresponding author: alex.jung@aalto.fi

Abstract

Tabular data consist of data points that share a common, fixed set of attributes. The attributes serve as the features or as the label of a data point. The data points constitute the rows, or records, of a table, and the attributes are its columns. Every row has one cell per column, and the cells within one column hold values of the same attribute. The fixed attribute set separates tabular data from other types of data such as text or networks.

Definition

Tabular data consist of data points that share a common, fixed set of attributes (Everitt, 2002). As the name suggests, tabular data can be stored as a table: the rows of the table are the data points, and the columns of the table are the attributes (Fig. 1). In database theory, the relational model formalizes such a table as a relation. A relation is a set of tuples of attribute values. The relational model also provides the operations, such as selecting rows or joining tables, that database systems use to manipulate tabular data (Silberschatz et al., 2019).

The weather measurements at a Finnish Meteorological Institute (FMI) weather station are typically represented as a table: one row for each day, with columns for the morning minimum temperature, the precipitation, and the maximum daytime temperature (Finnish Meteorological Institute, 2026). The attributes serve as the features or as the label of a data point. In Fig. 1, a machine learning (ML) method forecasts the maximum daytime temperature of a day from its morning minimum temperature: it reads the minimum-temperature column as the feature and the maximum-temperature column as the label, while the precipitation column is left unused.

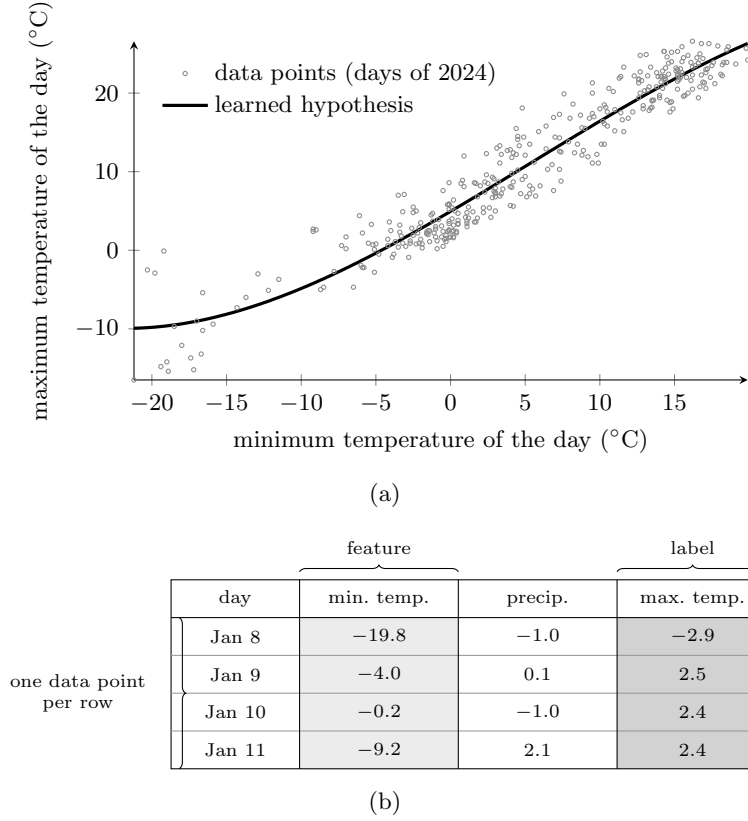


Figure 1: (a) Scatterplot of the 366 days of 2024 at the FMI station Helsinki Kaisaniemi, with the minimum temperature of a day as its feature and the maximum temperature as its label; the curve depicts a hypothesis map learned from these data points. (b) Four rows of the underlying table (January 8–11, 2024). Every row has one cell per column; the value range of the precipitation column includes the special value -1.0 , which indicates that no precipitation was observed on that day. Which columns serve as feature and label of a learning task is a design choice. Data generated by `pythondemos/tabulardata.py`

The distinct characteristic of tabular data is the rigid shape of a table. Every row contains the same number of cells, one per column, and the cells within one column hold values of the same attribute of the data points. The value range for each cell must be clearly defined. Often the value range includes a special value (such as “N/A”) that indicates an empty cell (Tennison et al., 2015). An empty cell is how missing data arises in tabular data: an attribute that was not recorded for a data point leaves its cell empty, while the table keeps its shape.

The fixed attribute set is what separates tabular data from other data types. A text is a sequence of tokens whose length varies from data point to data point,

and a graph varies in its node and edge sets. For such data points there is often no natural choice for a fixed list of attributes. Instead, feature learning methods are developed that map such data points to feature vectors (or embeddings) (Kipf and Welling, 2017; Mikolov et al., 2013).

Cross References

data, data point, dataset, feature, label, feature matrix, relational model, missing data, GBDT, feature learning

References

1. Everitt (2002). The Cambridge Dictionary of Statistics. 2nd ed., Cambridge Univ. Press, Cambridge, U.K.
2. Finnish Meteorological Institute (2026). The Finnish Meteorological Institute’s open data. Open data service, daily weather observations. URL: <https://opendata.fmi.fi>.
3. Mikolov et al. (2013). Efficient Estimation of Word Representations in Vector Space. In: ICLR (Workshop Poster).
4. Kipf and Welling (2017). Semi-Supervised Classification with Graph Convolutional Networks. In: 5th Int. Conf. Learn. Representations.
5. Tennison et al. (2015). Model for Tabular Data and Metadata on the Web. W3C Recommendation.
6. Silberschatz et al. (2019). Database System Concepts. 7th ed., McGraw-Hill Education, New York, NY, USA.