Types of Variables

Data Mining

Edgar Roman-Rangel.
edgar.roman@itam.mx

Computer Science Department.
Instituto Tecnológico Autónomo de México, ITAM.

Outline

Reminder

We already talked about different types of data, e.g., tabular, images, text, etc.

Now, let's see that regardless of the type of data, databases can be made out of different types of variables: binary, numeric, ordinal, categorical.

Data organization

Data is often represented by a set of attributes, which might have different names across disciplines,

- Variables (statistics).
- Attributes (data mining).
- Features (machine learning).
- Dimensions (data warehousing).
- Descriptors.

Often, data can be represented as a 2-D array (table): where each row is an object (a.k.a., record, datum, sample, individual, entity, point), and each column is a variable.

Varying natures

Depending on the type of data that is being described, variables might exist in different data types: categorical (nominal), binary (boolean), ordinal, numeric.

Part of the data mining job is to put them into a numeric representations that can be exploited (consistently and efficiently) by numerical (data mining, machine learning) methods.

Data can be univariate or multivariate.

It is important to get familiar with our data.

Numeric

- Quantitative attributes.
- Can be positive or negative.
- Can be discrete (integers) or continuous (real valued).
- We can exploit them in numeric processes.
- $x \in \mathbb{R}$ or $x \in \mathbb{Z}$.

Often, these type of variable can be exploited in computations without transforming it.

Nominal

a.k.a., categorical.

- x belongs to an unordered set of discrete labels.
- Names, labels, enumerations.
- Might be numbers as well, but with no numeric meaning.
- Useless to make numeric computations on them.
- Mode can be calculated.

Transformation might be needed for some computations. For instance, one-hot encoding.

Binary

- Nominal with only two possible values, either 0 or 1.
- Absent vs Present; True vs False; On vs Off.
- Numerical meaning is sometimes associated.

Depending on the computation required, transformations might, or not, be needed.

Ordinal

- Nominal ordered.
- x belongs to an ordered set of discrete labels.
- Meaningful order with unknown magnitude.
- e.g., small, medium, large.
- Ranking: satisfaction or grading.
- Mode and median make sense, but mean might not.

Transformation might be needed for some computations. For instance, one-hot encoding.

References

Data Science Project Scoping Guide.

http://www.datasciencepublicpolicy.org/home/resources/data-science-project-scoping-guide/

Jiawei Han, Micheline Kamber, Jian Pei.

"Data Mining Concepts and Techniques". Ch 2.

Elsevier. 2012.

Q&A

Thank you!

edgar.roman@itam.mx