Machine Learning Methods

Siger

August 19, 2023

Si Dieu est infini, alors je suis une partie de Dieu sinon je serai sa limite…

Contents

Chapter 1

Collect and Pre-process Data

1.1 Data cleaning

[1]

1.1.1 Data Quality

Validity

- **Data-Type Constraints**: for a given column a fixed data-type must be associated with.
- **Range Constraints**: only a range of values should be taken.
- **Mandatory Constraints**: some columns cannot be empty.
- **Unique Constraints**: across a given dataset a field or a combination of
- **Foreign-key constraints**: a foreign key column cannot have a value that does not exist in the primary key.
- **Regular expression patterns**: text fields that have to follow a given alphanumerical pattern.
- **Cross-field validation**: consistency of values, for example considering a given man, his birth date have to be older than his death date.

Accuracy The degree to which the data is close to the true value.

Completeness The degree to which the all the required data is known.

Consistency The degree to which the data is consistent, within the same data set or across multiple data sets.

Uniformity The degree to which the data is specified using the same unit of measure.

1.1.2 The workflow

Inspection Detect unexpected behavior in the data.

- **Data profiling**: summary statistics about the data, see ydata-profiling in Python.
- **Visualizations**: visualize the data using statistical metrics, see plotly
- **Software packages**: to note and check the constraints regarding the data see pydeequ

Cleaning Fix or remove anomalies discovered in the above phase.

- **Irrelevant Data**: ask to the expert what can be the unnecessary columns, check them and remove them if they are not useful.
- **Duplicates**
- **Type conversion**: make sure the appropriate data type is associated with a given column.
- **Syntax errors**: white spaces, pad strings ...
- **Standardize**: same unit across the dataset, same pattern for text.
- **Scaling/Transformation**: in order to compare different scores for example.
- **Normalization**: useful for some statistical methods.
- **Missing values**:
  - Drop: only if the missing values in a column rarely and randomly occur.
  - Impute: many methods, *mean* is relevant when data is not skewed otherwise we should use *median*. A linear regression or a hot-deck (copying of values) approach can be taken as well, and more interestingly a *k-nearest* method approach.
  - Flag: let the missing value as it is.
- Outliers: Remove outliers only if they are harmful for the chosen model.
- In-record & cross-datasets errors: fix non-consistent situations like married kids, quantity being different of the one when we compute using other columns.

Verifying Check correctness of the cleaning phase.

Reporting Report about changes made, using one of the software summarising the data quality for example.

Chapter 2

Statistics

2.1 Fundamental probability concepts

2.1.1 Distribution function

Probability Density Function (PDF)

Chapter 3

Conventional Statistical Learning

Chapter 4

Deep Learning

Bibliography

[1] Omar Elgabry. *The Ultimate Guide to Data Cleaning*. 2019. URL: https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4.