Holoclean Documentation

Release 0.01

Holoclean_team

Feb 22, 2018

Contents

CHAPTER 1

Holoclean

Noisy and erroneous data is a major bottleneck in analytics. Data cleaning and repairing account for about 60% of the work of data scientists. To address this bottleneck, we recently introduced HoloClean, a semi-automated data repairing framework that relies on statistical learning and inference to repair errors in structured data. In HoloClean, we build upon the paradigm of weak supervision and demonstrate how to leverage diverse signals, including user-defined heuristic rules (such as generalized data integrity constraints) to repair erroneous data.

1.1 Holoclean object

```
class HoloClean()

class Session("Session", holo_obj)
```

1.2 Ingesting Input file

```
session.ingest_dataset(dataset)
```

CHAPTER 2

Error detection

In the Holoclean pipeline, the user can choose the way he wants to seperate the clean from the don’t know cells.See tutorials for a more in-depth explanation.

```
ErrorDetectors(session.Denial_constraints, holo_obj.dataengine,holo_obj.spark_
↪session, session.dataset)
```

CHAPTER 3

Featurization

In the Holoclean pipeline, the user can choose the signal that he wants to use in order to train the model.

3.1 Domain Pruning

Holoclean gives the option to the user to prune the active domain

```
session.ds_domain_pruning(pruning_threshold)
```

3.2 Signals

```
SignalInit(session.Denial_constraints, holo_obj.dataengine,session.dataset)

SignalCooccur(session.Denial_constraints, holo_obj.dataengine,session.dataset )

SignalDC(session.Denial_constraints, holo_obj.dataengine, session.dataset, holo_obj.
↪spark_session)
```

CHAPTER 4

Learning

Currently we provide one basic models in Pytorch: Logistic regression. See tutorials for a more in-depth explanation.

```
SoftMax(holo_obj.dataengine, session.dataset, holo_obj.spark_session,session.X_
↪training)
```