

Pipelines in Python (Finally)

Patrick Smith

Pipelines



Pipelines



Pipelines

OPENING

Pipelines

Data pipelines are a series of automated data transformations used to perform (and ensure the validity of) routine data maintenance and analysis tasks.

Pipelines

Data pipelines are a series of automated data transformations used to perform (and ensure the validity of) routine data maintenance and analysis tasks.

Many organizations rely on data engineering teams to encode common tasks into pipelines. It is likely that you will at some point be required to productionize a data pipeline.

Pipelines

Naive Bayes is a method that uses the probabilities of all of the attributes of each class in a dataset to make a prediction.

We use Naive Bayes to model a predictive problem
utilizing probability

Pipelines

The term pipeline is jargon for a series of concatenated data transformations. Each stage of a pipeline feeds from the previous stage, i.e. the output of a stage is plugged into the input of the next stage and data flows through the pipeline from beginning to end.

Pipelines



Pipelines

What are some examples of data pipelines?

Pipelines

What are some examples of data pipelines?

- Change in Units
- Change in Scale
- Missing Data Imputation
- Image and Sound Processing

Pipelines

Pipelines improve coding and model management in scikit-learn. These tie together all the steps that you may need to prepare your datasets and make your predictions.

Pipelines

Because you will need to perform all of the exact same transformations on your evaluation data, encoding the exact steps is important for reproducibility and consistency. This is especially important and convenient when sharing code with a team!

Pipelines

Creating a Pipeline

Pipelines

The preprocessing module comes loaded with many very useful pre-processing classes.

Data Manipulators

 Binarizer

 KernelCenterer

 MaxAbsScaler

 MinMaxScaler

 Normalizer

 OneHotEncoder

 PolynomialFeatures

 RobustScaler

 StandardScaler

Data Imputation

 Imputer

Function Transformer

FunctionTransformer

Label Manipulators

 LabelBinarizer

 LabelEncoder

MultiLabelBinarizer

Conclusion

Pipelines

Q & A