Tools for data science

Week 1

Tasks:

Data management

Integration and transformation

Data visualization

Model building

Model deployment

Model monitoring and assessment

Code asset management

Data asset management

Development environments

Execution environment

Fully integrated visual tools

Open Source Tools:

Data management: Mysql mongoDB ceph, couchdb, handoop, elasticsearch

Integration and transformation: ETL or ELT (extract. Load, transform). Apache airflow, Kubeflow, apache kafka, apache nifi, spark sql, node red.

Data visualization: hue, kibana, superset

Model deployment: predictionio, seldon, mleap, tensorflow.

Model monitoring: modelDB, Prometheus, ai fairness, ibm tobustness, ai explainability.

Code asset management (version control): git, github, gitlab, bitbucket

Data asset management: apacheatlas, Egeria, kylo

Development environments: jupyter, jupyter lab, RStudio, spyder,

Execution environment: apache spark, flink, ray

Fully integrated visual tools: knime, orange

Commercial tools:

Data management: Oracle, SQL, IBM DB2

Integration and transformation (ETL): Informatice, IBM Infosphere , talend

Data visualization: (BI tools) Tableau, Microsoft Power BI, IBM Cognos Analytics

Model building” SPSS modeler, SAS

Model deployment: SPSS collaboration and deployment

Model monitoring and assessment: no commercial tools

Code asset management: no commercial tools

Data asset management: informatica, IBM infosphere

Development environments: IBM Watson studio

Execution environment

Fully integrated visual tools: Watson studio, H2O Ai

Cloud Based Tools

Data management: software as a service. Amazon DynamoDB, Cloudant, CouchDB

Integration and transformation: ELT. Informatice, IBM data refinary

Data visualization: Datameer, IBM Cognos Analytics, Watson Studio

Model building: IBM Machine learning, Google Cloud

Model deployment: SPSS collaboration, SPSS modeler, IBM Watson Machine Learning, Amazon SageMaker Model Monitor

Fully integrated visual tools and platfors: Watson stuio, Microsoft azure machine learning, h2o driveless ai.

Packages

Scientific computing: pandas (data structures and tools, dataframe), NumPy (arrays & matrices)

Visualization: matplotlib (plots & graphs), seaborn (plots: heat maps, time series)

High Level-Machine Learning and deep learning: scikit-learn, keras (deep learning)

Deep learning libraries: TensorFlow, PyTorch

Apache Spark:

Scala-libraries:

Vegas, BigDL (deep learning)

R-libraries:

Ggplot2, Keras, TensorFlow

APIs

Lets two pieces of software talk to each other

REST APIs (representational test transfer)

Data Asset Exchange (DAX): curated collection of data sets

Model:

Supervised learning, labeled by a human: regression models, classification models

Unsupervised learning, not labeled by a human: clustering and anomaly detection

Reinforcement learning, conceptually similar to human learning:

Deep learning: emulates how the human brain works, requires very large datasets. Tensorflow, pytorch, keras.

Model Asset Exchange (MAX)

Git and Github

Other git repositories: gitlab, bitbucket, and beanstalk