MATLAB EXPO 2017
KOREA
4월 27일, 서울
등록 하기 matlabexpo.co.kr
빅데이터 처리 및 머신 러닝 기법

Application Engineer
엄준상 과장
Data Analytics

Turn large volumes of complex data into actionable information
source: Gartner
Data Analytics Workflow

Access and Explore Data
- Files
- Databases
- Sensors

Preprocess Data
- Working with Messy Data
- Data Reduction/Transformation
- Feature Extraction

Develop Predictive Models
- Model Creation e.g. Machine Learning
- Parameter Optimization
- Model Validation

Integrate Analytics with Systems
- Desktop Apps
- Enterprise Scale Systems
- Embedded Devices and Hardware
Example: Working with Big Data in MATLAB

- **Objective:** Create a model to predict the cost of a taxi ride in New York City

- **Inputs:**
  - Monthly taxi ride log files
  - The local data set is **small** (~20 MB)
  - The full data set is **big** (~25 GB)

- **Approach:**
  - Access Data
  - Preprocess and explore data
  - Develop and validate predictive model (linear fit)
    - Work with subset of data for prototyping
    - Scale to full data set on a cluster
Example: Working with Big Data in MATLAB

**tall Arrays for Big Data in MATLAB**

Predict Cost of Taxi Ride in New York City

Analyze data from .csv files containing taxi trip information, separated by month. The data set is available from the City of New York.

Set up execution environment

```
numWorkers = 16;
setenv('HADOOP_HOME', '/mathworks/test/hadoop');
setenv('SPARK_HOME', '/mathworks/test/spark');
cluster = parallel.cluster.Hadoop;
cluster.SparkProperties('spark.executor.instances') = num2string(numWorkers);
```
Data Access and Pre-processing – Challenges

**Challenges**

- **Data aggregation**
  - Different sources (files, web, etc.)
  - Different types (images, text, audio, etc.)

- **Data clean up**
  - Poorly formatted files
  - Irregularly sampled data
  - Redundant data, outliers, missing data etc.

- **Data specific processing**
  - Signals: Smoothing, resampling, denoising, Wavelet transforms, etc.
  - Images: Image registration, morphological filtering, deblurring, etc.

- **Dealing with out of memory data (big data)**

---

Data preparation accounts for about **80%** of the work of data scientists - Forbes
Data Analytics Workflow: Data Access

Business and Transactional Data
- Repositories – SQL, NoSQL, etc.
- File I/O – Text, Spreadsheet, etc.
- Web Sources – RESTful, JSON, etc.

Engineering, Scientific and Field Data
- Real-Time Sources – Sensors, GPS, etc.
- File I/O – Image, Audio, etc.
- Communication Protocols – OPC (OLE for Process Control), CAN (Controller Area Network), etc.
Data Analytics Workflow: Big Data Access and Pre-processing

Download 2015 Taxi Data from Web using `websave` in parallel

```matlab
parfor i=1:12
    fileName = ['taxiData2015_', num2str(i)]
    url = ['https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2016-0',num2str(i), '.csv']
    websave(fileName, url)
end
```
Big Data in Recent Releases

- **datastore**
  - Tabular text files
  - Images
  - Excel spreadsheets
  - (SQL) Databases
  - HDFS (Hadoop)
  - S3 (Amazon Web Services)

- **MATLAB MapReduce**
  - Scales from Desktop to Hadoop

```matlab
airdata = datastore('*csv');
airdata.SelectedVariables = {'Distance', 'ArrDelay'};
data = read(airdata);
```
Data Analytics Workflow: Big Data Access and Pre-processing

Create a datastore to represent the data

A datastore is a repository for data and allows you to read part of the data memory.

```matlab
fileLoc = fullfile('taxiData','*.csv');
ds = datastore(fileLoc);
preview(ds)
```

Select variables of interest and give them more intuitive labels.

```matlab
vars = [2:3,5,12:13,16,19];
ds.VariableNames(vars) = {'Pickup','Dropoff','TripDistance','PaymentType','Fare','Tip','Total'};
ds.SelectedVariableNames = ds.VariableNames(vars);
```

Connect to the database application

```matlab
conn = database('taxiDemo','root','matlab',...  ...
    'Vendor','MYSQL',...  ...
    'Server','localhost',...  ...
    'PortNumber', 3306);
```

Create a database datastore and import data of interest

```matlab
sqlquery = ['select pickuptime, dropofftime, trip_distance,...  ...
    'payment_type, fare_amount from taxiData'];
ds = databaseDatastore(conn,sqlquery, 'ReadSize',100000);
```
tall arrays in **R2016b**

- New data type designed for data that doesn’t fit into memory
- Lots of observations (hence “tall”)
- Looks like a normal MATLAB array
  - Supports numeric types, tables, datetimes, strings, etc…
  - Supports several hundred functions for basic math, stats, indexing, etc.
  - **Statistics and Machine Learning Toolbox** support
    (clustering, classification, etc.)
tall arrays R2016b

- Automatically breaks data up into small “chunks” that fit in memory
- Tall arrays scan through the dataset one “chunk” at a time
- Processing code for tall arrays is the same as ordinary arrays
**tall arrays** R2016b

- With Parallel Computing Toolbox, process several “chunks” at once

- Can scale up to clusters with MATLAB Distributed Computing Server
Demo: Working with Tall Arrays

Create datastore to represent the data

```matlab
ds = datastore('smallerTaxiData/*2015.csv');
```

Identify data of interest and customize options.

```matlab
ds.VariableNames(2:3) = {'pickuptime','dropofftime'};
ds.SelectedVariableNames = {'pickuptime','dropofftime','trip_distance','payment_type','fare_amount'};
ds.SelectedFormats(1:2) = {'%{yyyy-MM-dd HH:mm:ss}D'};
```

Create a tall array

```matlab
tt = tall(ds)
```

Determine trip duration

```matlab
tt.hr_of_day = hour(tt.pickuptime);
tt.trip_minutes = minutes(tt.dropofftime - tt.pickuptime)
```
Data Access and pre-processing – challenges and solution

**Challenges**

- **Data aggregation**
  - Different sources (files, web, etc.)
  - Different types (images, text, audio, etc.)

- **Data clean up**
  - Poorly formatted files
  - Irregularly sampled data
  - Redundant data, outliers, missing data etc.

- **Data specific processing**
  - Signals: Smoothing, resampling, denoising, Wavelet transforms, etc.
  - Images: Image registration, morphological filtering, deblurring, etc.

- **Dealing with out of memory data (big data)**

**Solutions**

- Point and click tools to access variety of data sources
- High-performance environment for *big data*
- Built-in algorithms for data preprocessing including sensor, image, audio, video and other real-time data
Consider Machine/Deep Learning When

**Problem is too complex for hand written rules or equations**
- Speech Recognition
- Object Recognition
- Engine Health Monitoring

*Because algorithms can*
- learn complex non-linear relationships

**Program needs to adapt with changing data**
- Weather Forecasting
- Energy Load Forecasting
- Stock Market Prediction

*update as more data becomes available

**Program needs to scale**
- IoT Analytics
- Taxi Availability
- Airline Flight Delays

*learn efficiently from very large data sets*
Different Types of Learning

**Type of Learning**
- **Supervised Learning**
  - Develop predictive model based on both input and output data
- **Unsupervised Learning**
  - Discover an internal representation from input data only

**Categories of Algorithms**
- **Classification**
  - Output is a choice between classes (True, False) (Red, Blue, Green)
- **Regression**
  - Output is a real number (temperature, stock prices)
- **Clustering**
  - No output - find natural groups and patterns from input data only
Machine Learning with Big Data

**R2016b**

- Descriptive statistics (skewness, tabulate, crosstab, cov, grpstats, …)
- K-means clustering (kmeans)
- Visualization (ksdensity, binScatterPlot; histogram, histogram2)
- Dimensionality reduction (pca, pcacov, factoran)
- Linear and generalized linear regression (fitlm, fitglm)
- Discriminant analysis (fitcdiscr)

**R2017a**

- Linear classification methods for SVM and logistic regression (fitclinear)
- Random forest ensembles of classification trees (TreeBagger)
- Naïve Bayes classification (fitcnb)
- Regularized regression (lasso)
- Prediction applied to tall arrays
Regression Learner
Demo: Training a Machine Learning Model
Demo: Training a Machine Learning Model
Regression Learner
App to apply advanced regression methods to your data

- Added to Statistics and Machine Learning Toolbox in R2017a
- Point and click interface – no coding required
- Quickly evaluate, compare and select regression models
- Export and share MATLAB code or trained models
Classification Learner
App to apply advanced classification methods to your data

- Added to Statistics and Machine Learning Toolbox in R2014a
- Point and click interface – no coding required
- Quickly evaluate, compare and select classification models
- Export and share MATLAB code or trained models
Tuning Machine Learning Models
Get more accurate models in less time

Automatically select best machine learning “features”

Automatically fine-tune machine learning parameters

R2016b
NCA: Neighborhood Component Analysis

R2016b
Hyperparameter Tuning
Machine Learning Hyperparameters

Tune a typical set of hyperparameters for this model

Tune all hyperparameters for this model
Bayesian Optimization in Action
MATLAB Production Server

- Server software
  - Manages packaged MATLAB programs and worker pool

- MATLAB Runtime libraries
  - Single server can use runtimes from different releases

- RESTful JSON interface

- Lightweight client libraries
  - C/C++, .NET, Python, and Java
Integrate analytics with systems
Key Takeaways

1. MATLAB Analytics work with **business and engineering data**

2. MATLAB enables **domain experts to do Data Science**

3. MATLAB Analytics **run anywhere**

- Utilize all of your data.
- Apply advanced analytics techniques.
- Operationalize analytics to enterprise systems and embedded devices.
Resources to learn and get started

mathworks.com/big-data

mathworks.com/machine-learning