Machine Learning Pipelines

Rayid Ghani

THE UNIVERSITY OF
CHICAGO

Slides liberally borrowed and customized from lots of excellent online sources

Things we will cover

- What is a ML Pipeline?
- What components should it have?
- Best Practices
- Examples

Why a pipeline?

- Reusable across projects
- Test new ideas/components easily
- Reduce bug/errors

Components

- Read/Load Data (from csv, db, api)
- Integrate Data (dedupe, link)
- Explore Data (descriptives, correlations, outliers, over time, clustering)
- Process Data
  - Missing values (fill/impute, create dummy)
  - Transformations (scale/normalize, log, square, root)
  - Feature Generation (
- Modeling
  - Create training and test sets
  - Define metric(s)
  - Build model
  - Validate model
- Model Selection and Validation
- Communication
- Field Trial

Data Acquisition & Integration

- Get Data
  - API, CSV, Database
- Store Data
  - Database
- Integrate Data
  - Record Linkage

Explore and Prepare data

- Data Exploration
  - Distributions
  - Missing Values
  - Correlations
  - Other Patterns
- Pre-Processing
  - Leakage
  - Deal with Missing values
  - Scaling
  - Data errors

Feature Creation

- Common Features
  - Discretization
  - Transformations
  - Interactions/Conjunctions
  - Disaggregation
  - Aggregations
    - Temporal
    - Spatial

Method Selection

- Select pool of methods applicable for task
- For loop over a large number of methods
  - For loop over parameters

Validation

- Using historical data
  - Methodology
  - Metric

- Field Experiment
  - Methodology
  - Metric

Deployment

- Re-training
  - How often?
  - Re-select methods?
- Scoring

What do you want to test

- Different models
- Model parameters
- Labels/Outcomes
- Feature (Groups)
- Metrics

Best Practices

- Config files (yaml, json, py)
- Store models as pickles
- Store predictions in databases
- Store evaluation metrics in databases
- Sample results schema

Config file example

- https://github.com/dssg/san_jose_housing/blob/master/example_experiment_config.yaml

What should a simple pipeline do?

- Build a simple, modular, extensible, machine learning pipeline with functions to do the following:
  - ETL and exploration
    - Load Data
    - Explore data
    - Pre-process data
  - Matrix Creation
    - Create rows
    - Create labels for each row
    - Create one feature
  - Train Test Set Creation
    - Generate one training set
    - Generate one validation set
  - Modeling
    - Build 1 classifier on training set
    - Run the 1 classifier on the validation set