The Machine Learning Workflow

Data Collection and Preparation

- What Data to collect?
  - What problem to solve?
  - Business Understanding is crucial for the success of any Data Mining/Machine Learning project
- Combining datasets
  - Structured, unstructured sources
  - Excel sheets and CSV files
  - Legacy datasets

Exploratory Data Analysis

- Statistical Summary
- Univariate Explorations of Data
  - Histogram
  - Box plot
- Multivariate Explorations of Data
  - Scatter plot
  - Correlation and Correllograms
  - Facet plots

Data Preprocessing

- Centering and Scaling
- Missing Value Treatment
  - Univariate methods
  - Multivariate methods
- Outlier Treatment
  - Univariate methods
  - Multivariate Supervised methods
  - Multivariate Unsupervised methods

Feature Engineering

Feature Transformation

- Rescaling
  - Centering
  - Scaling
  - Normalization, Standardization
- Algebraic Transformations
  - Polynomial
  - Exponential and Logarithmic
- Combining Existing Features

Feature Engineering

Feature Creation/Extraction

- Recoding
    - Categorical to Ordinal/Numeric
        - Hashing Trick
        - One-hot Encoding
        - Label Encoding
        - Counting Level Frequency
    - Numeric to Categorical
        - Binarization
        - Discretization
- Brainstorming
    - Combining Features
    - Applying Domain Expertise
- Feature Space Reengineering
    - Dimensionality Reduction
    - Dimensionality Explosion

Feature Engineering

Feature Selection

- Model-based
  - Regularization
  - Ensemble
- Filter
  - Pearson Correlation Coefficient
  - Mutual Information
  - embedded methods (feature selection is a part of model construction)
- Wrappers
  - Forwards
  - Backwards

Feature Engineering

Example

- Predict if a team in IPL will qualify for the semi-final
  - Total runs scored against the team
  - Total runs scored by the team
- What features can we build from this data?

Modeling

Selecting the right algorithm

- The Task
  - Classification vs Regression etc
- Dimensionality of Data
  - p>>n
  - n>>p
- Linear Separability
- Interpretability vs Predictive Power
- Online vs Offline
- Start by choosing the simplest hypothesis

Modeling

Sampling

- Train-Test-Validation Split
  - Simple Validation
  - Cross-Validation
- Dealing with Imbalanced Data

Modeling

- Select the right error metric
- Training

Modeling

Ensembling

- Generating an ensemble of models
    - Bagging
    - Subspace Sampling
    - Boosting
- Combining an ensemble of models
    - Voting and Aggregation
    - Stacking

Modeling

Hyperparameter Training

- Grid Search
- Random Search
- Bayesian Strategies

Modeling

Model Assessment

- Model Evaluation and Selection