Machine Learning for Finance, Algorithmic Trading and Investing

Brief introduction to machine learning, application to finance and investing, modelling technique and model validation.

Anthony Ng @ July 2017

Trends in Big Data, Machine Learning, and Finance

- Uncorrelated strategies
- Alpha
- Quantitative strategies
- Beyond traditional
  - Systematic
  - Cross-assets
  - Risk premia investing – value, momentum, quality, volatility, growth
- Big data revolution
  - Massive increase in data
  - Access to low cost computing power and data storage capacity
  - Advances in Machine Learning and Deep Learning

Categories of Machine and Deep Learning

- **Supervised**: Inferring a function from labelled training data
  - Regression
  - Classification
- **Unsupervised**: No labels are given to the learning algorithm, leaving it on its own to find structure in its input.
  - Clustering
  - Dimensionality Reduction
- **Deep Learning**. ANN, CNN, RNN, Reinforcement Learning.

Source: https://en.wikipedia.org/wiki/Machine_learning

Machine Learning / Deep Learning

Supervised Learning		Unsupervised Learning		Deep Learning	
Regression	Classification	Clustering	Factor Analysis	Time Series	Unstructured
Linear, Lasso, Ridge, KNN, XGBoost	Logistic, SVM, Random Forest, Hidden Markov	K-Means, Ward, Mean Shift	PCA, ICA		

Machine Learning for Finance and Investing Series

1. Introduction to machine learning, linear regression, financial time series stylised facts
2. Machine Learning for Algorithmic Trading - Pairs Trading
3. Machine Learning for Multi-assets Trend Following Strategies
4. Classification Based Machine Learning for Algo Trading
5. Ensemble Machine Learning for Algorithmic Trading
6. Unsupervised Machine Learning Hidden Markov for Algo Trading
7. Clustering and Unsupervised Machine Learning for Investing

Linear Regression

Linear Regression Model

$$y = \beta_0 + \beta_1 X$$

- Y is the dependent variable
- X is the independent variable

- Linear in nature
- Both X and Y are numeric

Source: https://en.wikipedia.org/wiki/Ordinary_least_squares

A scikit-learn Walk Through

- Machine Learning Workflow
- Using Linear Regression as an example

Assumptions of Linear Regression

- Linear relationship
- Multivariate normality
- No or little multicollinearity
- No auto-correlation
- Homoscedasticity

Source: http://www.statisticssolutions.com/assumptions-of-linear-regression/

Financial Data

Stylised Statistical Properties of Asset Returns

1. Absence of autocorrelations
2. Heavy tails
3. Gain/loss asymmetry
4. Aggregational Gaussianity
5. Intermittency
6. Volatility clustering
7. Conditional heavy tails
8. Slow decay of autocorrelation in absolute returns
9. Leverage effect
10. Volume/Volatility correlation
11. Asymmetry in time series

Source: http://www.proba.jussieu.fr/pageperso/ramacont/papers/empirical.pdf

Characterization of Financial Time Series

- **Dependence**. Autocorrelation in returns is largely insignificant, except at high frequencies when it becomes negative.
- **Distribution**. Approximately symmetric, increasingly positive kurtosis as the time interval decreases and a powerlaw or Pareto-like tail.
- **Heterogeneity**. Non-stationary (clustered volatility).
- **Non-linearity**. Non-linearities in mean and (especially) variance.

Source: http://www.cs.ucl.ac.uk/fileadmin/UCL-CS/images/Research_Student_Information/RN_11_01.pdf

Characterization of Financial Time Series

- **Scaling**. Markets exhibit non-trivial scaling properties.
- **Volatility**. Volatility exhibits positive autocorrelation, long-range dependence of autocorrelation, scaling, has a nonstationary log-normal distribution and exhibits non-linearities.
- **Volume**. Distribution decays as a power law, also calendar effects.

Source: http://www.cs.ucl.ac.uk/fileadmin/UCL-CS/images/Research_Student_Information/RN_11_01.pdf

Characterization of Financial Time Series

- **Calendar effects**. Intraday effects exist, the weekend effect seems to have all but disappeared, intramonth effects were found in most countries, the January effect has halved, and holiday effects exist in some countries.
- **Long memory**. There is about a 30% chance that stock market returns exhibit long memory, a 50% chance that foreign exchange returns exhibit long memory and an 80% chance that market volatility exhibits long memory.
- **Chaos**. There is little evidence of low-dimensional chaos in financial markets.

Source: http://www.cs.ucl.ac.uk/fileadmin/UCL-CS/images/Research_Student_Information/RN_11_01.pdf

Effects of Outliers to Linear Regression

- Demo 1
- Demo 2

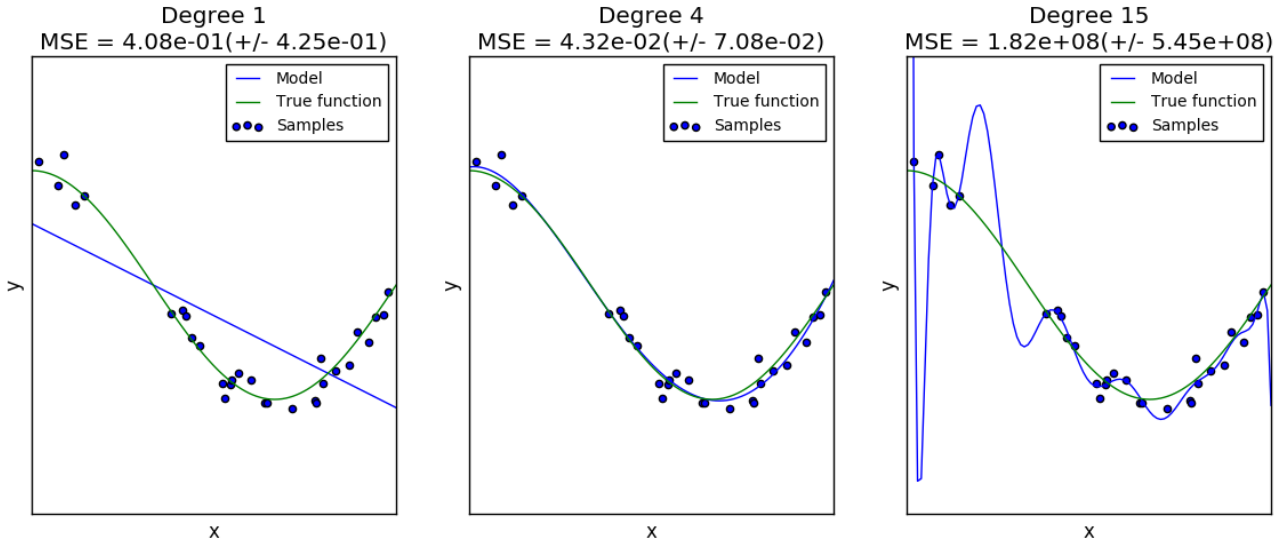
Model Selection

Hyperparameters and Model Validation

- Cross-Validation
  - Split into training set and testing set
- Tuning hyper-parameters
  - Parameters that are not directly learnt
  - Grid search
  - Randomized optimization

Selecting the Best Model

- The Bias-Variance tradeoff
- Model Evaluation
  - Mean absolute error
  - Mean squared error
  - Accuracy
  - Log loss
  - Adjusted rand score
- Validation Curves
- Learning Curves



High bias model

High variance model

Source: http://scikit-learn.org/stable/modules/learning_curve.html

Final Step

- 6 stages of quant strategy:
    1. Data
    2. Universe Definition
    3. Alpha discovery
    4. Alpha combination
    5. Portfolio Construction
    6. Trading
- Hypothesis and backtest

Source: https://blog.quantopian.com/a-professional-quant-equity-workflow/

MIT License