Machine Learning and Applications in Finance

Christian Hesse\textsuperscript{1,2,*}

\textsuperscript{1} Autobahn Equity Europe, Global Markets Equity, Deutsche Bank AG, London, UK
christian-a.hesse@db.com

\textsuperscript{2} Department of Computer Science, University College London, London, UK
c.hesse@ucl.ac.uk

* The opinions and ideas expressed in this presentation are those of the author alone, and do not necessarily reflect the views of Deutsche Bank AG, its subsidiaries or affiliates.
Outline

• Machine Learning Overview
• Unsupervised Learning
• Supervised Learning
• Practical Considerations

• Recommended Reading:
Machine Learning

Machine learning is concerned with the design and development of data-driven algorithms able to identify and describe complex structure in data.

Machine learning algorithms are designed to tackle:

- High-dimensional data
- Noisy data
- Data corrupted by artifacts
- Data with missing values
- Data with small sample size
- Non-stationary data (i.e., structural changes in data generating process)
- Non-linear data

Machine learning techniques have been successfully applied in many areas of science, engineering and industry … including finance

Related fields: computer science, artificial intelligence, neural networks, statistics, signal processing, computational neuroscience
Machine Learning Approach

• Machine learning is generic
  – Data is “just numbers”
  – Data can be structured

• Machine learning is data-driven
  – Data structure is defined as a model
  – *Model has minimal or generic assumptions!*
  – Model parameters estimated from data

• Machine learning is robust
  – Maximize performance on unseen data
  – Consistent performance
  – Complexity control (avoid over-fitting)
  – Reliable and efficient parameter estimation
Machine Learning Problems

- **Unsupervised Learning**
  - Identifying component parts of data
  - Representing different data features

- **Unsupervised Learning Methods**
  - Dimension estimates/reduction
  - Decomposition methods
  - Clustering methods

- **Supervised Learning**
  - Mapping one data part onto another

- **Supervised Learning Methods**
  - Regression
  - Ranking
  - Classification
  - Feature Selection
Decomposition Methods

• Orthogonal de-correlating transforms (Gaussian mixtures): \( x = As + n \)
  – PCA, SVD, Whitening
  – Probabilistic PCA and Factor Analysis

• Non-orthogonal un-mixing transforms (non-Gaussian mixtures): \( x = As + n \)
  – Independent component analysis (ICA)
  – Probabilistic and “noisy” ICA

• Factorization and coding methods: \( X = WH \)
  – Non-negative matrix factorization (NMF), etc …
  – Dictionary learning and sparse coding

• Applications
  – Dimension reduction, regularization and de-noising
  – Feature extraction

• Applications in Finance (portfolio optimization)
  – Risk factor models, covariance matrix regularization
Clustering Methods

- **K-Means Clustering**
  - Distance metric (Euclidean, City-block, cosine)
  - Number of centres

- **Probabilistic Clustering**
  - Mixture of Gaussians (spherical covariance)
  - Mixture of Probabilistic PCA
  - Mixture of Factor Analysers

- **Non-Gaussian Clusters**
  - Mixture of ICA models
  - Mixture of Von-Mises Fisher distributions

- **Time Series Clusters**
  - Clusters reflect states
  - State transitions and Hidden Markov Models (HMM)
Application: Volume Analysis

Source: Bloomberg
Intra-day Volume Profiles

Source: Bloomberg
Volume Profile Analysis

• Motivation
  – Examination of market structure
  – Importance in algorithmic trading

• Data Set
  – Stocks: constituent names of the STOXX Europe 50 Index
  – Period: Dec 2010 - May 2014
  – Intra-day trading volumes from primary exchange aggregated over 5 minute buckets
  – Normalized volume profiles (density)

• Analysis Techniques
  – K-Means Clustering
  – Non-negative Matrix Factorization
  – Different initialization methods

Data Source: Thomson Reuters
Volume Profile Cluster Analysis

K-Means Cluster Analysis of Intraday Volume Profiles

Data Source: Thomson Reuters
Volume Profile Cluster Analysis

K-Means Cluster Analysis of Intraday Volume Profiles (Excluding Close Auction)

Samples from STOXX Europe 50 Index constituents (2010 - 2014)

Raw Volume Densities

Cluster Activations

Most Active Clusters

Least Active Clusters

Data Source: Thomson Reuters
Volume Profile Decomposition

Non-Negative Matrix Factorization of Intraday Volume Profiles (Initialization: Random)

Raw Volume Densities

Factors Activations

Most Active Factors

Least Active Factors

Samples from STOXX Europe 50 Index constituents (2010 - 2014)

Time (HH:MM)

08:00 12:00 16:00

Factor Index

1 6 12

Time (HH:MM)

08:00 12:00 16:00

Data Source: Thomson Reuters
Volume Profile Decomposition

Non-Negative Matrix Factorization of Intraday Volume Profiles (Initialization: Furthest)

Samples from STOXX Europe 50 Index constituents (2010 - 2014)

Data Source: Thomson Reuters
Volume Profile Analysis

• Summary
  – Both approaches are sensitive to fundamental characteristics of the volume data, e.g., special days, US market effects, intra-day skews
  – K-means provides an exemplar-based representation
  – NMF provides a reduced sum-of-parts representation
  – Unclear which is more desirable/useful

• Open issues and ongoing work
  – Intelligent initialization of both methods is important
  – What is the best distance metric to use for k-means here
  – What is the most appropriate model order selection approach here
  – “Vanilla” NMF results exhibit spurious zeros > consider extensions of NMF
  – Behaviour on data from less liquid stocks

• Applications
  – Feature extraction for intra-day volume prediction
Supervised Learning

- **Data structure**
  - Most of the data X is “just numbers”
  - A part of the data Y are annotation
  - Pairs reflect a mapping of X onto Y

- **Learning task: find mapping Y = F(X)**
  - The nature of the learning task depends on the scale that Y is measured on
  - Y is metric >> regression task
  - Y is ordinal >> ranking task
  - Y is nominal >> classification task

- **What kind of mapping is F?**
  - Linear or non-linear map
  - Exemplar or kernel based

- **How complex and reliable is the map**
  - Feature (variable) selection
  - Regularization and stability
Application: Index Forecasting

Data Source: Bloomberg, Thomson Reuters
Classifying Future Index Moves

Hang Seng Index from 20050404 to 20090204

Return

Labels

Time Axis [t]

HSI

SX5E

SPX

GOLD

...

USDJPY

EMG

ITRX

VIX

Data Source: Bloomberg, Thomson Reuters

Time Axis [t-1]
Classifier Evaluation

• “Out-of-Sample” Evaluation Procedure

Step 1  
[ ] train [ ] test

Step 2  
[ ] train [ ] test

... 

Step T  
[ ] train [ ] test

data time line

• Measure aggregate proportion of correct predictions (hit rate)
• Compare with guessing, naïve benchmarks and/or other classifiers
Feature Selection: Linear Methods

Hang Seng Index from 20050404 to 20090204

Data Source: Bloomberg, Thomson Reuters
Feature Selection: Kernel Methods

Hang Seng Index from 20050404 to 20090204

Return

Labels

HSI

SX5E

SPX

GOLD

...

USDJPY

EMG

ITRX

VIX

discriminative features  train  test

discriminative features  train  test

Time Axis [t]

Time Axis [t-1]

Data Source: Bloomberg, Thomson Reuters
Classification Results

- **HSI 2-class problem**
  - Best “out-of-sample” hit rate for 2-class case is 0.67
  - Statistically significantly better than guessing and naïve benchmarks

- **Confusion Matrix**

<table>
<thead>
<tr>
<th>predicted class</th>
<th>-1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>observed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>0.6429</td>
<td>0.3571</td>
</tr>
<tr>
<td>1</td>
<td>0.3029</td>
<td>0.6971</td>
</tr>
</tbody>
</table>

- **Classification seems to perform well, but is it good enough?**

![Performance of Classifier versus Benchmarks and Guessing](image)

- `best` = best classification method
- `bear` = always decrease
- `bull` = always increase
- `same` = always same as day before
- `freq` = most frequent label in training window
Practical Considerations

• Data Quality and Quantity
  – “Garbage in > garbage out”
  – Missing values, Repeated values
  – Some methods impractical for very large datasets
  – Regularization can help fitting large models to small datasets
  – Sample biases affect model estimates

• Performance Evaluation
  – Cross-validation tests
  – “Out-of-sample” tests (moving window, growing window)

• Application Domain Knowledge
  – Remains critically important
  – Required to define learning problem (e.g., labels for classification)