**Pj174\_StatArb SPX Value Chain**

1. **Define solution**

StatArb S&P500 constituents

|  |  |  |  |
| --- | --- | --- | --- |
| Daily | Long best k | 🡪 | 1 day ahead  trading signal |
| Short worst k |

1. **Gather data**

1989.12 ~ 2015.09 month end constituent binary matrix

1990.01 ~ 2015.10 daily total return indices (corp actions: stock split, dividend, …)

1992.12 ~ 2015.10 Trading

Eliminate survivorship bias

1. **Data preprocessing**

Eliminate holidays

1. **Split dataset**

|  |  |  |
| --- | --- | --- |
| Training window 750 days | 🡪 | 23 batches |
| Trading window 250 days |

1. **Feature engineering**

Input : 31 features

n(≒ 500) : num of stocks at end of training period with full data for prior 750 days

Output : 🡪 that cross-sectional median return

1. **ML algos / Ensemble**

DNN, GBT, RAF / ENS 🡪 later develop more advanced ML model

1. DNN

Backpropagation 31-31-10-5-2 layers

Maxout activation function, Softmax at output layer

Cross-entropy loss function

Dropout (input 0.1, hidden 0.5)

L1 regularization (shrinkage parameter λ 0.00001)

H2O ADADELTA Optimization (momentum learning, rate annealing)

400 epochs, random seed 1

1. GBT

Tree depth 3

15 features

Learning rate 0.1

100 iterations, random seed 1

1. RAF

1000 trees

Maximum tree depth 20

5 features (), random seed 1

1. ENS
2. **Performance metric**

Rate of return

Transaction costs = 0.05% per half-turn

1. **MODEL**
2. **REAL data**
3. **Reap**

A prototype for future development