
David Andre
Fall 2017
In this paper we propose a multi-level machine learning process for automating the role of the data scientist and the investment committee, allowing for the discovery and combination of investment strategies that utilize advanced machine learning techniques.
Outline

• Challenges in applying advanced ML techniques in asset management
• Background on automatic machine learning
• Data mining need not be harmful — it’s how you use info that matters
• Structural Stochastic Beam Search
• Empirical example using single-stock prediction vs. market
• The sieve (meta-machine-learning)
• Commentary and other results
AI History through Games

- **Chess**: 1997
- **Jeopardy**: 2011
- **Atari**: 2013
- **Go**: 2016
Artificial Intelligence Revolution

- Innovations in cloud computing, open-source software, and machine learning have made machine learning ascendant.
- Leading to AI/ML driven revolutions in ad-placement, speech and image recognition, self-driving cars, ....
Challenges in applying AI/ML to investing

1. Financial assets are correlated with one another and in time
2. Spurious correlations abound
3. Random signals might look surprisingly good
4. Non stationary relationships and only one history
5. In financial back-testing, “time travel” is dangerously easy
6. Many researchers succumb to trial-and-error tweaking of inputs, resulting in manual overfitting
7. Data scientists are expensive
Uses of ML in Asset Management

Retail Sales Per Company

Sentiment Per Company

Many quant firms are
A. hiring or outsourcing tons of ML experts, or
B. crowd-sourcing their ML
Traditional Quant Workflow

1) Pose Problem
2) Choose ML Method and Constraints
3) Run ML Method
4) Did it work?
5) What to do next?
Our Approach

1) Pose Problem
2) Choose ML Method and Constraints
3) Run ML Method
4) Did it work?
5) What to do next?

Traditional Quant Workflow
• Because our process is automated, we can run it in the past
• We can also run it with random or simulated data
• When a researcher launches a run, they can launch invention machines at every point in the past, so we know when the system could have discovered each signal or strategy
Let’s assume we have a method of generating investment strategies that fit within the given framework — we’ll call it the generation method.

The question is, given the in-sample and out-of-sample performance characteristics of the strategy, do we believe that it will perform well in the future?

We can use a Bayesian estimate of this likelihood.
Estimating good future performance

\[ P(F \mid g, s, x, y) = c^{-1} P(g)P(s \mid F)P(y)P(F \mid y) \sum_B P(x \mid F, B)P(B \mid g, s) \]
Estimating good future performance

\[
P(F \mid g, s, x, y) = c^{-1} P(g)P(s \mid F)P(y)P(F \mid y) \sum_B P(x \mid F, B)P(B \mid g, s)
\]

Constant, is estimable, but can be ignored for relative ranking

Can be estimated via lots of runs of the generator
Estimating good future performance

Strategy features can be estimated for various alphas using alpha (fake alpha)

Can be mostly estimated using a parametric prior and observed returns from generated strats

\[ P(F \mid g, s, x, y) = c^{-1} P(g)P(s \mid F)P(y)P(F \mid y) \sum_B P(x \mid F, B)P(B \mid g, s) \]
Estimating good future performance

Can be estimated using “nalpha” or random noise data as input to the generator.

Estimating alpha from observed unbiased returns

Estimating in-sample returns given different levels of true alpha and bias caused by fitting.

\[ P(F \mid g, s, x, y) = c^{-1} P(g)P(s \mid F)P(y)P(F \mid y) \sum_{B} P(x \mid F, B)P(B \mid g, s) \]
Structural Stochastic Beam Search

- SSBS is a generalization of simulated annealing

- **Structural** — SSBS searches over structures (often program trees)
- **Stochastic** — the search steps are probabilistic
- **Beam** — SSBS can use a population of candidates
- **Search** — SSBS requires both a proposal (what to try next) and an acceptance function (should we keep the thing we tried)
SSBS: Example 1 — Split Regression

If $X > 0.3$

- Proposal — select one randomly from:
  - change the classifier variable
  - add a variable to a regression
  - remove a variable from a regression
  - optimize the threshold for the given regressions
  - fit both regressions
  - iterate the previous two steps 10 times.
SSBS: Example 2 — Genetic Programming

- Crossover and mutation on tree-based computer programs
- Function and terminal set can be simple or complex
- Can represent network structures via “growth” functions
- Has been used to invent/discover many different kinds of structures or programs, including neural networks and electronic circuits.

SSBS: Example 3 — Automatic Machine Learning

- Proposal function modifies learning pipeline
- Examples: add dimensionality reduction, add or remove variables, change machine learning method, change variable selection method
- Also, we need not just pick randomly — we can use bayesian optimization or reinforcement learning to choose
- Auto-sklearn, TPOT, and AutoML by Google are examples in the literature
Empirical Example

- Single stock prediction (hedged with SPY)
- Function set includes arithmetic operations, logical operators and conditionals, and operations such as trailing mean, standard deviation, and z-score.
- Terminal set includes technicals and fundamental company info
- Target is a Sharpe ratio with a hurdle rate
Example Programs from run on PFE

2002: \( \text{Logreturn}(4) \times [\text{HighLowSpread128}-8.1] \)

Sharpe Ratio: 1.44
Return: 13% / year

2005: \( \frac{\text{Receivables/Debt}}{} + \frac{\text{HighLowSpreadRatio256}}{\text{MACD}(6,12)} + \frac{\text{HighLowSpreadRatio256}}{\text{LinTrendRegSlope}(\text{HighLowSpread128})} - 2 \times \text{CloseHighRatio} \)

Sharpe Ratio: 1.79
Return: 17.5% / year
Leaderboards and the Sieve

• At each slice in time where we run SSBS, we save the best candidates that are not too correlated to one another onto what we call a leaderboard.

• Over time, we end up with hundreds or thousands of candidates per slice.

• We also have all the candidates that were invented earlier on older data.

• How to choose which to trade on?

• The Bayesian formula mentioned above gives us a clue — but what features of the generation process, the in and out of sample performance, and the strategy itself matter?
Machine learning on machine learning

• To approximate the necessary probabilities, we utilize a second layer of machine learning (again run at each of several points in time)
• The terminal set contains performance characteristics of the candidates, the generative process, and the in-and out-of sample performance.
• The function set contains different variable selection methods and different classifiers.
• We call this process the sieve, as it selects down to a manageable number of diverse candidates
Before the Sieve

1069 Strategies
Median OOS return: 0.08%
52% positive OOS
After the Sieve

16 Strategies
Median OOS return: 1.36%
75% positive OOS
Combination of Strategies

OOS Sharpe 1.5
All Stocks, OOS, through time

- Sieve doesn’t work perfectly for all stocks
- 75% are positive OOS.
- Single stock models are hard with only technicals and fundamentals
- We can improve the likelihood of success by taking correlations into account, and adding more stocks.
• Using an optimizer to allocate to stocks creates a good overall portfolio.
• Results are uncosted
• We can add many more stocks
• Overall correlation of returns is low
Moving past pedagogy

- We launched a fund in April using the techniques described in this talk
- Instead of single-stock predictors, it works with baskets of stocks and predicts 2-week return
- Performance has matched expectations
- Trades with 48 models selected each month using the sieve process
- Currently developing a long-only version; results for that shown in graph to left
Conclusions

• Instead of a timid approach that avoids looking at what works, take advantage of computational power by assessing overfitting risk using all the available information.
• Future work will utilize additional frameworks, utilize smarter search techniques, and of course new input data sets.
• Additionally, some of the information from the sieve process can inform the original strategy search process.
• Current research is on using deep learning to improve the diversity of candidates and to improve the sieve
Disclaimer

THIS DOCUMENT SHALL NOT CONSTITUTE AN OFFER TO SELL OR THE SOLICITATION OF ANY OFFER TO BUY AN INTEREST IN THE CEREBELLUM MULTI-STRAIGHT FUND OR THE CEREBELLUM MACHINE LEARNING FUND (THE “FUNDS”), OR CEREBELLUM CAPITAL, INC (THE “MANAGER”) WHICH MAY ONLY BE MADE AT THE TIME A QUALIFIED OFFEREES RECEIVES AN OFFERING MEMORANDUM (“PPM”), WHICH CONTAINS IMPORTANT INFORMATION (INCLUDING INVESTMENT OBJECTIVE, POLICIES, RISK FACTORS, FEES, TAX IMPLICATIONS AND RELEVANT QUALIFICATIONS, AND ONLY IN THOSE JURISDICTIONS WHERE PERMITTED BY LAW). IN THE CASE OF ANY INCONSISTENCY BETWEEN THE DESCRIPTIONS OR TERMS IN THIS DOCUMENT AND THE PPM, THE PPM SHALL CONTROL. THESE SECURITIES SHALL NOT BE OFFERED OR SOLD IN ANY JURISDICTION IN WHICH SUCH OFFER, SOLICITATION OR SALE WOULD BE UNLAWFUL UNTIL THE REQUIREMENTS OF THE LAWS OF SUCH JURISDICTION HAVE BEEN SATISFIED. THIS DOCUMENT IS NOT INTENDED FOR PUBLIC USE OR DISTRIBUTION. WHILE ALL THE INFORMATION PREPARED IN THIS DOCUMENT IS BELIEVED TO BE ACCURATE, CEREBELLUM CAPITAL MAKES NO EXPRESS WARRANTY AS TO THE COMPLETENESS OR ACCURACY OF THE DOCUMENT, NOR CAN IT ACCEPT RESPONSIBILITY FOR ERRORS APPEARING IN THE DOCUMENT.

AN INVESTMENT IN THE FUND IS SPECULATIVE AND INVOLVES A HIGH DEGREE OF RISK. OPPORTUNITIES FOR WITHDRAWAL/REDEMPTION AND TRANSFERABILITY OF INTERESTS ARE RESTRICTED, SO INVESTORS MAY NOT HAVE ACCESS TO CAPITAL WHEN IT IS NEEDED. THERE IS NO SECONDARY MARKET FOR THE INTERESTS AND NONE IS EXPECTED TO DEVELOP. THE PORTFOLIO, WHICH IS UNDER THE SOLE AUTHORITY OF CEREBELLUM CAPITAL, IS PRIMARILY INVESTED IN THE DOMESTIC EQUITIES MARKETS AND THIS LACK OF DIVERSIFICATION MAY RESULT IN HIGHER RISK. LEVERAGE MAY BE EMPLOYED IN THE PORTFOLIO, WHICH CAN MAKE INVESTMENT PERFORMANCE VOLATILE. AN INVESTOR SHOULD NOT MAKE AN INVESTMENT UNLESS IT IS PREPARED TO LOSE ALL OR A SUBSTANTIAL PORTION OF ITS INVESTMENT. THE FEES AND EXPENSES CHARGED IN CONNECTION WITH THIS INVESTMENT MAY BE HIGHER THAN THE FEES AND EXPENSES OF OTHER INVESTMENT ALTERNATIVES AND MAY OFFSET PROFITS.

THERE IS NO GUARANTEE THAT THE INVESTMENT OBJECTIVE WILL BE ACHIEVED. MOREOVER, THE PAST PERFORMANCE (IF ANY) OF THE INVESTMENT TEAM SHOULD NOT BE CONSTRUED AS AN INDICATOR OF FUTURE PERFORMANCE. ANY PROJECTIONS, MARKET OUTLOOKS OR ESTIMATES IN THIS DOCUMENT ARE FORWARD-LOOKING STATEMENTS AND ARE BASED UPON CERTAIN ASSUMPTIONS. OTHER EVENTS WHICH WERE NOT TAKEN INTO ACCOUNT MAY OCCUR AND MAY SIGNIFICANTLY AFFECT THE RETURNS OR PERFORMANCE OF THE FUNDS. ANY PROJECTIONS, OUTLOOKS OR ASSUMPTIONS SHOULD NOT BE CONSTRUED TO BE INDICATIVE OF THE ACTUAL EVENTS WHICH WILL OCCUR.

THE INFORMATION PROVIDED HEREIN, INCLUDING, WITHOUT LIMITATION, INVESTMENT STRATEGIES, INVESTMENT RESTRICTIONS AND PARAMETERS, AND INVESTMENT AND OTHER PERSONNEL, MAY BE MODIFIED, TERMINATED OR SUPPLEMENTED AT ANY TIME WITHOUT FURTHER NOTICE IN A MANNER WHICH CEREBELLUM CAPITAL BELIEVES IS CONSISTENT WITH ITS OVERALL INVESTMENT OBJECTIVE OF LONG-TERM CAPITAL APPRECIATION AND REDUCED RISK.