Machine Learning & Time Series Analysis

A Dow Jones Trading Model

Martin Sewell
martin@martinsewell.com

Equitywire
Machine Learning for Equities
London

26 April 2018

Outline of Presentation

- The futility of bias-free learning
- Efficient-market hypothesis
- Financial markets stylised facts
- Bayesian model selection
- DJIA trading model
- Bayesian model averaging
- Conclusions

The Futility of Bias-Free Learning

- 'Even after the observation of the frequent conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience.' Hume (1739–40)
- Bias-free learning is futile (Mitchell, 1980; Schaffer, 1994; Wolpert, 1996).
- One can never generalise beyond one's data without making at least some assumptions.

No Free Lunch Theorem for Supervised Machine Learning

- Wolpert (1996) showed that in a noise-free scenario where the loss function is the misclassification rate, in terms of off-training-set error, there are no *a priori* distinctions between learning algorithms.
- On average, all supervised machine learning algorithms are equivalent (over all possible data sets).
- There is no free lunch for Occam's razor, overfitting avoidance or cross validation (they cannot be justified from first principles).
- We should only constrain our models according to our prior beliefs.
- Extract as much domain knowledge as possible.
- Carefully consider the assumptions being made.

Efficient Market Hypothesis

- A market is *efficient* with respect to an information set if the price 'fully reflects' that information set (Fama, 1970).
- Efficient if the price would be unaffected by revealing the information set to all market participants (Malkiel, 1992).
- The definitional 'fully' is an exacting requirement—strictly speaking the EMH is false.
- One of the strongest hypotheses in the whole of the social sciences—in spirit the EMH is profoundly true.
- Science concerns seeking the best hypothesis—the EMH is the best current hypothesis.

Market Predictability

Due to risk aversion, investors require a small positive expected return in risky markets. In long-only markets—like a stock market—this implies a positive upward drift. In symmetric markets which traders are as likely to be long as they are short, like futures and foreign exchange markets, the implication is that one would expect the price to be predictable to some degree. Government intervention in foreign exchange markets may provide a positive sum game for other participants in the short-term (LeBaron, 1999).

Market	Beat market?	Reason
Foreign exchange	hard	government intervention
Futures	harder	
Stock	hardest	positive sum game

Investment Newsletters

- Experiment to determine market efficiency
- Analysis of the Forbes/Hulbert investment letter survey
- 31 May 1990 to 31 December 2001
- Predominantly US equity focussed
- 8 purely fundamental newsletters, 2 beat the market
- 9 purely technical newsletters, none beat the market
- Other literature, mixed conclusions

Fund Performance

- Reviewed the literature on fund performance
- We can gauge market efficiency by identifying persistence in the returns of fund managers
- 18 papers found evidence of manager skill, 7 supported market efficiency
- 5 papers explicitly mentioned market timing, none of them found that fund managers were able to time the market
- Stock picking is a worthwhile activity—fundamental analysis works
- Market timing is not possible—technical analysis doesn't work

Non-linearities

- There exists empirical evidence that a non-linear process contributes to the dynamics of market returns (e.g. Scheinkman and LeBaron (1989)).
- Neftci (1991) showed that technical analysis relies on non-linearities being present.
- Park and Irwin (2004) found that, in terms of technical analysis, non-linear methods work best overall.

Financial Markets Stylised Facts 1

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 power-law or Pareto-like tail
Heterogeneity	Non-stationary (clustered volatility)
Non-linearity	Non-linearities in mean and (especially) variance
Scaling	Markets exhibit non-trivial scaling properties
Volatility	Volatility exhibits positive autocorrelation, long-range dependence of autocorrelation, scaling, has a non-stationary log-normal distribution and exhibits non-linearities

Financial Markets Stylised Facts 2

Volume Distribution decays as a power law, also calendar effects

Calendar effects Intraday effects exist, the weekend effect seems to have all but disappeared, intramonth effects have been found in most countries, the January effect has halved, and holiday effects exist in some countries

Long memory About 50 per cent of the articles analysing market returns concluded that they exhibit long memory, and about 80 per cent of those analysing market volatility concluded that it exhibits long memory

Chaos There is little evidence of low-dimensional chaos in financial markets

Red Herrings

Early claims made for the following turned out to be largely unfounded as higher-frequency data became available:

- stable distributions
- long memory in returns
- low-dimensional chaos

Model Selection

Model selection is the task of choosing a model with the correct inductive bias. In practice, select a model of optimal complexity for the given (finite) data.

1. Choose model space — difficult
2. Model selection — difficult
   Choose from

$$m_1 = a_{11}x_1 + a_{10}$$
$$m_2 = a_{22}x_2 + a_{21}x_1 + a_{20}$$

3. Parameter estimation — easy
   Given

$$m_1 = a_{11}x_1 + a_{10}$$

   find a_{11} and a_{10}

Bayesian Model Selection

- Finding the most probable model with noisy data was solved in principle by Harold Jeffreys nearly 80 years ago (Jeffreys, 1939)
- Explicitly trades model complexity, as determined by prior probabilities, against the (probabilistic) fit to the data
- Informs how much structure can be justified by the given data and the assumed model space
- Chooses the model with the largest posterior probability
- Works with nested or non-nested models
- No need for a validation set
- No ad hoc penalty term or assumptions (except the prior)
- Consistent—if one of the entertained models is true, with enough data it will be chosen

Bayesian Model Selection

$P(\text{model}|\text{data}) \propto \text{prior} \times \text{likelihood}$

H a hypothesis (or model)
D data

$P(H)$ the prior—subjective
$P(D|H)$ the likelihood—objective
$P(H|D)$ the posterior

$P(H|D) \propto P(H)P(D|H)$

Only interested in the *relative* probability of different hypotheses

Prior

- Choosing the model space is a very important first step.
- The prior, $P(H)$, is our subjective assessment of how surprising a model is.
- The Bayesian approach makes incorporation of prior information relatively easy and explicit.
- Apply the principle of indifference, and thus a uniform prior, not to models but to functions (instances of models).

Prior and Complexity

Model	Complexity	Volume in parameter space	prior
$a_{11}x_1 + a_{10}$	simple	n^2	cn^2
$a_{22}x_2 + a_{21}x_1 + a_{20}$	complex	n^3	cn^3

n is arbitrary, c is a normalising constant

Likelihood

- To compare models of different complexity, it is necessary to marginalise the parameters.
- The marginal likelihood, $P(D|H)$, is the probability of the data given the model with a random choice of parameters.

Model	Complexity	Fit to data	Likelihood
$a_{11}x_1 + a_{10}$	simple	poor	large
$a_{22}x_2 + a_{21}x_1 + a_{20}$	complex	good	small

Posterior

The Bayesian approach automatically gives the necessary trade-off between model complexity and fit to the data.

Model	Prior	Likelihood	Posterior
Simple	small	large	?
Complex	large	small	?

Pedagogical Example: Data

- Dow Jones Industrial Average
- Daily data
- Log returns
- Source: Yahoo Finance

Data set	Start	End	No. data points
Training	3 Jan 2000	31 Dec 2009	2515
Test	4 Jan 2010	19 April 2018	2088

Model Inputs

- 5 orthogonal potential inputs, x_n, and a target, y
- p_n is closing price n days in the future

$$
\begin{aligned}
x_1 &= log(p_0/p_{-1}) \\
x_2 &= log(p_{-1}/p_{-5}) \\
x_3 &= log(p_{-5}/p_{-10}) \\
x_4 &= log(p_{-10}/p_{-20}) \\
x_5 &= log(p_{-20}/p_{-40}) \\
y &= log(p_1/p_0)
\end{aligned}
$$

Model Space

Models are nested, but need not be

$$
\begin{aligned}
m_1 &= a_{11}x_1 + a_{10} \\
m_2 &= a_{22}x_2 + a_{21}x_1 + a_{20} \\
m_3 &= a_{33}x_3 + a_{32}x_2 + a_{31}x_1 + a_{30} \\
m_4 &= a_{44}x_4 + a_{43}x_3 + a_{42}x_2 + a_{41}x_1 + a_{40} \\
m_5 &= a_{55}x_5 + a_{54}x_4 + a_{53}x_3 + a_{52}x_2 + a_{51}x_1 + a_{50}
\end{aligned}
$$

Volume of Parameter Space

Model	No. params	Volume
m_1	2	10^2
m_2	3	10^3
m_3	4	10^4
m_4	5	10^5
m_5	6	10^6

The choice of 10 is arbitrary

Priors

- Tobler's first law of geography informs us that 'everything is related to everything else, but near things are more related than distant things' (Tobler, 1970).
- Autocorrelation in market returns is largely insignificant, but Campbell, Lo, and Mackinlay (1996) found that autocorrelations of daily stock index returns was positive.

Model Priors

Priors determined according to complexity and significance of one-day lag

$$P(m_1) = c_1 \times 11^2 \times 0.6 = 0.000540$$
$$P(m_2) = c_1 \times 11^3 \times 0.1 = 0.000900$$
$$P(m_3) = c_1 \times 11^4 \times 0.1 = 0.008996$$
$$P(m_4) = c_1 \times 11^5 \times 0.1 = 0.089960$$
$$P(m_5) = c_1 \times 11^6 \times 0.1 = 0.899604$$

c_1 is a normalising constant

Bayesian Information Criterion (BIC)

BIC is easy to calculate and enables us to approximate the marginal likelihood.

n = number of data points
k = number of free parameters
RSS is the residual sum of squares

$\text{BIC} = n \log(\text{RSS}/n) + k \log(n)$
(marginal) likelihood $\propto e^{-0.5\text{BIC}}$

Model Likelihoods

Model	n	k	RSS	BIC	Likelihood
m_1	2515	2	0.432518	-21784.8	0.928762
m_2	2515	3	0.432061	-21779.6	0.069811
m_3	2515	4	0.432061	-21771.8	0.001395
m_4	2515	5	0.432029	-21764.1	0.000030
m_5	2515	6	0.431846	-21757.4	0.000001

Model Posteriors

$P(\text{model}|\text{data}) \propto \text{prior} \times \text{likelihood}$

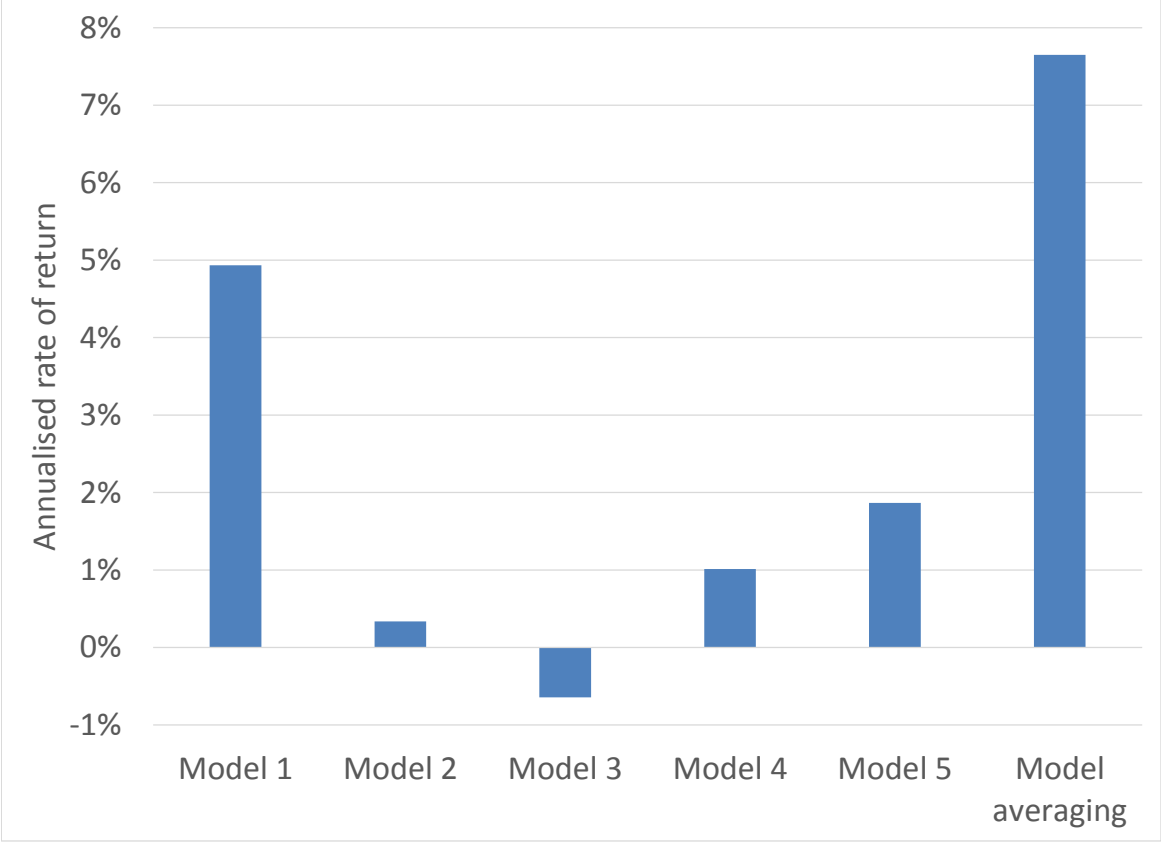
$$P(m_1|D) = c_2 \times 0.000540 \times 0.928762 = 0.863823$$
$$P(m_2|D) = c_2 \times 0.000900 \times 0.069811 = 0.108217$$
$$P(m_3|D) = c_2 \times 0.008996 \times 0.001395 = 0.021630$$
$$P(m_4|D) = c_2 \times 0.089960 \times 0.000030 = 0.004722$$
$$P(m_5|D) = c_2 \times 0.899604 \times 0.000001 = 0.001607$$

c_2 is a normalising constant

Bayesian Model Averaging

- The goal is to make as accurate predictions as possible about future data.
- It is optimal to take an average over all models, with each model's prediction weighted by its posterior probability.

Out-of-Sample Results



Ignoring transaction costs

Conclusions

- Science is essentially applied Bayesian analysis.
- Bayesian analysis formalises the fact that a surprising result requires more evidence.
- Machine learning can be viewed as an attempt to automate 'doing science'.
- Everyone should be a Bayesian (willing to put a probability on a hypothesis).
- Use domain knowledge to infer an appropriate bias.
- Prediction using Bayesian model averaging beats the best model.

Bibliography

Campbell, J. Y., Lo, A. W., & Mackinlay, A. C. (1996). *The Econometrics of Financial Markets*. Princeton, NJ: Princeton University Press.

Fama, E. F. (1970, May). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383–417.

Hume, D. (1739–1740). *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*. Oxford: Oxford University Press. (Edited by Norton, D. F. and Norton, M. J., published 2000.)

Jeffreys, H. (1939). *Theory of Probability*. Oxford: Oxford University Press. (Third ed., Oxford Classic Texts in the Physical Sciences, 1998.)

LeBaron, B. (1999). Technical trading rule profitability and foreign exchange intervention. *Journal of International Economics*, *49*(1), 125–143.

Malkiel, B. (1992, October). Efficient market hypothesis. In P. Newman, M. Milgate, & J. Eatwell (Eds.), *The New Palgrave Dictionary of Money and Finance.* London: Macmillan.

Mitchell, T. M. (1980). *The need for biases in learning generalizations* (Technical report No. CBM-TR-117). New Brunswick, NJ: Rutgers University.

Neftci, S. N. (1991, October). Naive trading rules in financial markets and Wiener-Kolmogorov prediction theory: A study of "technical analysis". *The Journal of Business*, *64*(4), 549–571.

Park, C.-H., & Irwin, S. H. (2004, October). *The profitability of technical analysis: A review* (AgMAS Project Research Report No. 2004-04). Urbana: University of Illinois at Urbana-Champaign.

Schaffer, C. (1994). A conservation law for generalization performance. In W. W. Cohen & H. Hirsh (Eds.), *Proceedings of the Eleventh International Conference on Machine Learning* (pp. 259–265). San Francisco, CA: Morgan Kaufmann.

Scheinkman, J. A., & LeBaron, B. (1989, July). Nonlinear dynamics and stock returns. *The Journal of Business*, *62*(3), 311–337.

Tobler, W. R. (1970, June). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*(Supplement), 234–240.

Wolpert, D. H. (1996, October). The lack of *a priori* distinctions between learning algorithms. *Neural Computation*, *8*(7), 1341–1390.