AVOIDING OVERFITTING IN QUANTITATIVE TRADING MODELS

EDWARD YU

LUCAS SCHUERMANN

COLUMBIA UNIVERSITY, MARCH 2017

MOTIVATION

- Financial data is noisy and has complex structure
- It requires complex models to understand
- Unfortunately, complex models may fit historical data perfectly but not generalize well
- This is called overfitting. We want models that will work both on historical and out of sample data

SUMMARY

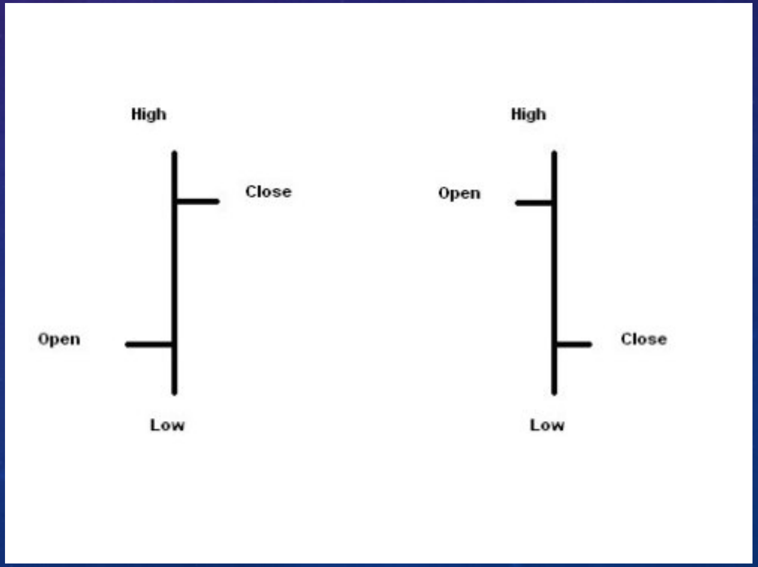
1. Get good data
2. Make model as simple as possible while maintaining profitability
3. Rigorously test model in as many scenarios as possible

BEST DATA PRACTICES

GET DATA AS GRANULAR AS POSSIBLE

- Minute bars will give you 1440x more data than daily bars
- More data → less overfitting
- Also prevents overfitting on unobtainable entry prices (daily open, close, etc)

Source: https://johnlivy.files.wordpress.com/2009/02/ohlc.jpg?w=500&h=375



IS YOUR DATA REPRESENTATIVE OF THE MARKET?

- Data set may contain survivorship bias
- Quant models may work well on old data (pre 2013), but not on newer data

MODEL SELECTION

SMOOTH THE PROBLEM

- Example: use principal component analysis to reduce dimension of data matrix X
- Example: smooth response variable using exponential moving average

Source: https://dev.tradingsim.com/wp-content/uploads/2011/06/Generating-a-Buy-Signal-while-Trading-with-the-Exponential-Moving-Average-1024x620.png

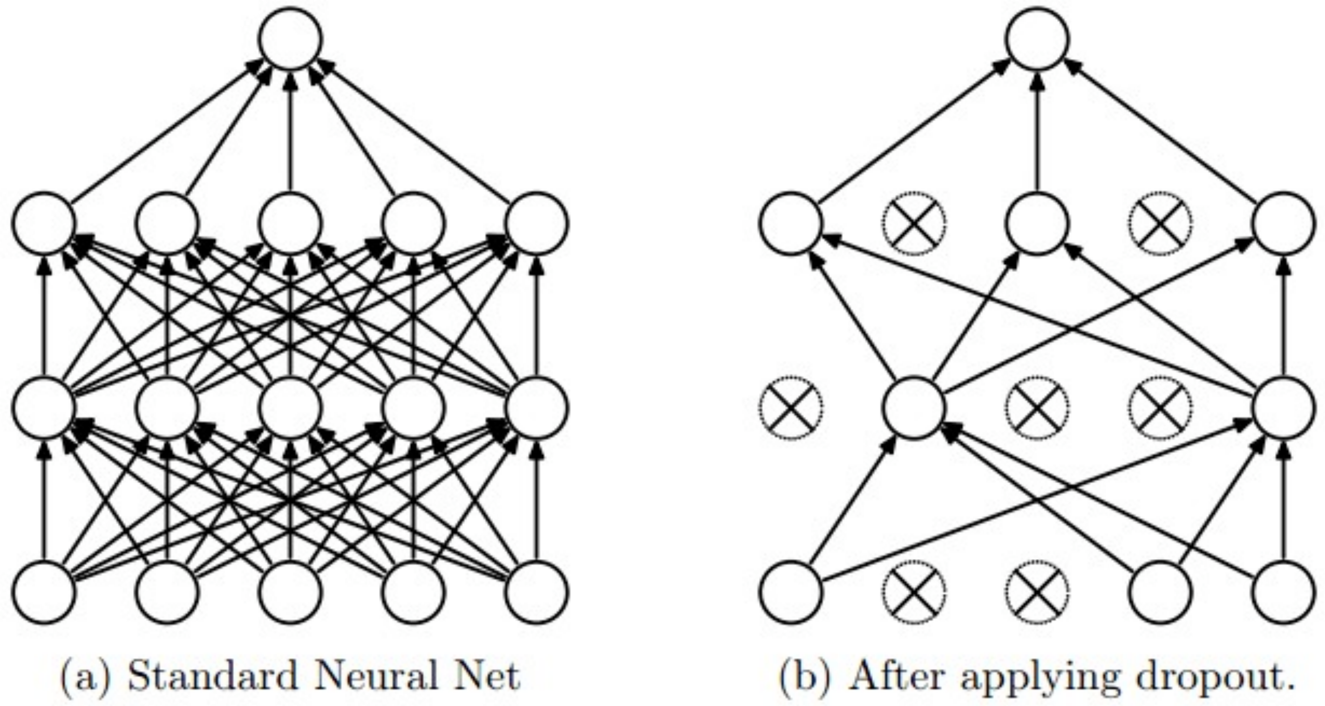
PENALIZE COMPLEXITY

- Example: LASSO

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Source: Linxi Liu, Statistical Machine Learning lecture notes

- Example: Dropout



Source: http://cs231n.github.io/assets/nn2/dropout.jpeg

STATISTICALLY RIGOROUS BACKTESTING

TRAIN/TEST/VALIDATION SPLIT

1. Optimize model parameters on training set. If backtest performs well, proceed to step 2.
2. Backtest model on test set. If backtest performs well and there is no large drop in performance, proceed to step 3. Otherwise, return to step 1.
3. Backtest model on validation set. If backtest performs well and there is no large drop in performance, proceed to realtime paper trading. Otherwise, return to step 1.

- Backtest on validation set as few times as possible! Preferably only once.
- Test/train/validation datasets must be in chronological order to avoid look-ahead bias.
- Corollary: do not use k-fold cross validation on timeseries data.

OPTIMIZATION TIPS AND TRICKS

WALK-FORWARD OPTIMIZATION

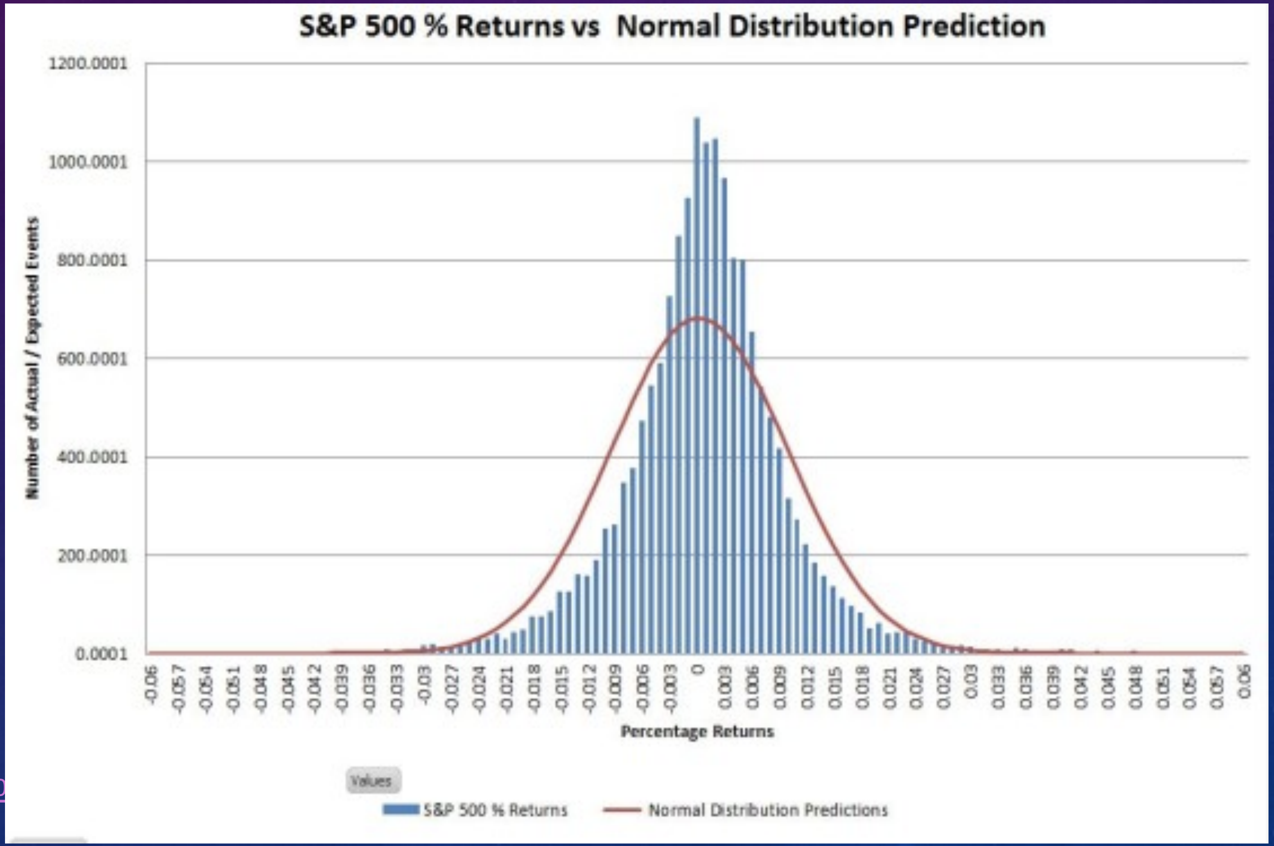
- Model at time t_n should only have access to data from $[t_0, t_{n-1}]$
- This requires that your algorithm is adaptive
- Better likelihood of generalizing well to new market conditions

DON'T ONLY OPTIMIZE FOR RETURNS

- Optimizing for portfolio return places high weight on anomalous events.
- Sharpe ratio: $\frac{r_p - r_f}{\sigma_p}$ (excess portfolio return / standard deviation)
- Optimizing for Sharpe ratio favors strategies that deliver consistent returns with high probability.
- Natural penalty for variance (bias-variance tradeoff).
- Mathematically: comparing returns is not a good equivalence relation.

AVOID NORMALITY ASSUMPTIONS

- In general, don't assume financial data is generated by a Gaussian distribution.



Source:
https://sixfigureinvesting.com/2016/0

- Assuming a normal distribution of S&P 500 daily returns, how often do we expect the index to move 8.76% (9 standard deviations)?
- Twice in 20000000000000000 years.
- Reality: in just under 70 years, this has happened *7 times*.

- Implication: if confidence intervals are generated with assumption that errors are iid Gaussian, they may be far too optimistic.
- See: 2008 financial crisis.
- “But in the CDO market, people used the Gaussian copula model to convince themselves they didn't have any risk at all, when in fact they just didn't have any risk 99 percent of the time. The other 1 percent of the time they blew up.” – WIRED

Source: http://archive.wired.com/techbiz/it/magazine/17-03/wp_quant?currentPage=all

- Tentative advice: use a fat-tailed distribution or use bootstrap samples to generate confidence intervals.
- One commonly used trick is to assume extreme events happen C times more often than a normal distribution implies.
- Not mathematically rigorous but works well in practice.

SUMMARY

1. Get good data
2. Make model as simple as possible while maintaining profitability
3. Rigorously test model in as many scenarios as possible