ST326 Week 7

Kaixin Liu[1]

[1]PhD Student in Statistics, LSE

Nov 14, 2025

Table of Contents

Table of Contents

Concept and pipeline of ML

Machine learning

Detect/Learn certain "signals" from data by training a "machine" – a class of model which can depend on various **tuning parameters**. Fine tune the machine by twitching the tuning parameters until an optimization criterion is reached. In order to avoid **overfitting**, divide data into three **mutually exclusive** sets: training, validation and test set.

1. Define the prediction problem.
2. Data engineering and preprocessing.
3. Choose a model family.
4. Training / validation / test split.
5. Evaluation and interpretation.

ML pipeline

1. **Define the prediction problem.** Predict Y_{t+1} (FTSE return) using information available at time t.
2. **Data engineering and preprocessing.**
   - Align calendars of all indices. Missing days get filled with random small returns; everything standardised to variance 1.
   - Volatility adjustment using exponential filtering.

$$\sigma_t^2 = (1-\lambda)x_{t-1}^2 + \lambda\sigma_{t-1}^2.$$

   (This is actually a simple ML model for volatility.)
3. **Choose a model family.**
   - Single-index / multi-index models as a way to reduce dimension (project high-dimensional $\mathbf{z}_t$ onto a low-dimensional index $\mathbf{u}^\top \mathbf{z}_t$.
   - But for this course a linear regression model is settled, with rolling windows.
   - To deal with high dimensionality or multicollinearity, variable screening or penalised regression can be implemented.

ML pipeline (Cont.)

4. **Training / validation / test split.**
   - Training: estimate regression parameters (and volatility model) for each candidate D, λ.
   - Validation: for each candidate D, re-run the rolling strategy on validation data and compute Sharpe ratio as performance measure.
   - Test: once picked D, λ, evaluate again on the held-out test set to assess out-of-sample performance.

   (Classic ML story: don't judge your model by training performance.)

Table of Contents

Ridge regression

Consider the original OLS problem in Step 3, which is to solve

$$\widehat{\boldsymbol{\alpha}}_D(t) = \arg\min_{\alpha} \|\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha}\|^2.$$

To restrict the magnitudes of $\boldsymbol{\alpha}_D(t)$, we can solve

$$\min_{\alpha} \|\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha}\|^2, \qquad \text{subject to } \|\boldsymbol{\alpha}\|^2 \leq c,$$

where $c > 0$. Using Lagrange multiplier, the above problem is equivalent to

$$\min_{\alpha} \left\{ \|\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha}\|^2 + \delta \|\boldsymbol{\alpha}\|^2 \right\}$$

for some $\delta > 0$. Solving this, we get

$$\widehat{\boldsymbol{\alpha}}_\delta = (\mathbf{Z}^\top \mathbf{Z} + \delta \mathbf{I}_p)^{-1} \mathbf{Z}^\top \mathbf{Y}.$$