MSDM5058 Information Science
Computational Project II:
Portfolio Management with Prediction Tools

This computational project consists of nine parts. You need to complete the project in a group of two to four people. While you will analyse the same data as your groupmates and can share graphs with them, you need to write an individual report, which will be appraised independently from theirs. Your report must conform to the LNCS format, otherwise 10% will be deducted from its score. You need to submit your report onto Canvas at the latest on **11 May, 2024**.

1 Data Preprocessing

Choose two stocks whose prices are available on more than 4000 days from any credible sources like Yahoo Finance. Cite your source properly.

Denote the closing-price time series of the riskier stock by $S_1(t)$ and that of the safer stock by $S_2(t)$, where "today" $t = 0$ is defined so that the length ratio of the "past" to the "future" is around 3:1. Use the past data for learning up to Section 6, then use the future data for testing from Section 7 onwards.

- Plot $S_1(t)$, $S_2(t)$, and their daily-return

$$X_i(t) = \ln\left[\frac{S_i(t)}{S_i(t-1)}\right] \quad \text{for} \quad i = 1, 2. \tag{1}$$

2 Mean-Variance Analysis

Let us first construct the stocks' **minimum-risk portfolio** $S_0 = p_0 S_1 + (1 - p_0) S_2$. Although we know the formula of the fraction p_0, we cannot use it before estimating the stocks' population statistics. Hence, p_0 is practically not a constant but a function of time as our estimation also changes with time.

On one hand, we may infer $p_0(t)$ from all the data that we have observed so far, i.e. $\{\, S_{i=1,2}(\tau) \mid \tau \leq t \,\}$. On the other hand, since the relevance of old data should decay, we may limit ourselves and infer it only from the data on the h most recent days, i.e. $\{\, S_{i=1,2}(\tau) \mid \tau \in [t-h, t] \,\}$. The first method is in some sense equivalent to the second method by letting $h \to \infty$.

- Plot $p_0(t, h)$ for $h = 30$, 100, 300, and ∞. The three finite values roughly correspond to one month, one season, and one year.
- Plot $S_0(t, h)$ and its Sharpe ratio $\gamma_0(t, h)$ for the same values of h.
- How differently do the portfolios perform? Discuss the effects of h.

3 Moving Average

Now define $S \equiv S_1$ and $X \equiv X_1$. Hereafter when a task is stated in terms of S and X, you need to perform it on the riskier stock only.

We are going to examine the effectiveness of moving average, which some traders heavily rely on and regard as accurate indicators.

- Plot the w-day exponential moving average (EMA) of S for $w = 30$, 100, and 300, then describe the effects of w.
- Compare the EMA to the w-day simple moving average (SMA) of S for the same values of w. How are the two averaging methods similar or different?
- Compute the stock's **moving-average-convergence-divergence** (MACD) line and signal line. What happens to S when the MACD line crosses the signal line or zero? Does your observation match the empirical rules?

  You may freely adjust the window sizes used in the computation of the two lines for nicer results.

4 Probability Density Function

We are often interested in PDF $f(x)$ of X, but we are not going to construct it directly by binning due to its subjectivity. Instead, we may assume X is normally distributed and estimate $f(x)$ with a normal distribution

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{2}$$

after measuring the mean μ and the variance σ^2 of X. We may also choose to assume X is logistically distributed and fit a logistic function

$$L(x) = \frac{1}{1 + \exp[-b(x - x^*)]} \tag{3}$$

to the CDF $F(x)$ of X so that its derivative $L'(x)$ approximates $f(x)$. The parameters x^* and b can be determined by considering $L(x^*)$ and $L'(0)$.

- Report the values of the parameters (μ, σ^2) and (x^*, b).

- Plot your fits $G(x) \equiv \int g(x)\mathrm{d}x$ and $L(x)$ atop $F(x)$. Do the two assumptions look reasonable?

4.1 Digitization and Conditional Probability

Given the return $X(t)$ on a day, we would like to predict its value $X(t+1)$ on the next day. However, a quantitative prediction is too ambitious. It is more realistic to predict the trend qualitatively.

- Digitize $X(t)$ as $Y(t)$ with three alphabets, viz. D for "down", U for "up", and H for "hold". You need to choose a sensible value for the threshold ε so that you will not be distracted by transient fluctuations in X.

$$Y(t) = \begin{cases} \mathrm{D} & [X(t) < -\varepsilon] \\ \mathrm{U} & [X(t) > +\varepsilon] \\ \mathrm{H} & (\text{otherwise}) \end{cases} \tag{4}$$

- Plot $F(\,x \mid y\,) \equiv \mathrm{CDF}[\,X(t) = x \mid Y(t+1) = y\,]$, i.e. the conditional CDF of X given that Y will become y one day later, for $y =$ D, U, and H.
- Follow the method of this section to extract the three corresponding conditional PDFs $f_y(x)$, then plot them on the same graph.

  You may freely choose the underlying form of the distributions. If you think it is neither normal nor logistic, you may use other assumptions.

5 Bayes Detector

We will now construct a Bayes detector with the just obtained PDFs to predict $Y(t+1)$ after observing $X(t) = x$. Formally, we have three hypotheses, namely "$H_\mathrm{D} : Y(t+1) = \mathrm{D}$", "$H_\mathrm{U} : Y(t+1) = \mathrm{U}$", and "$H_\mathrm{H} : Y(t+1) = \mathrm{H}$".

- Compute the probabilities $P[Y(t+1) = y]$ for $y =$ D, U, and H. These values can be used as the **prior probabilities** $q(y)$ of the hypotheses.
- Despite its more tedious algebra, a three-hypothesis detector follows the same principle as a two-hypothesis detector: assuming no reward for right decisions but equal cost for wrong decisions, we choose to believe in H_{y^*} for

$$\begin{aligned} y^* &= \operatorname*{argmax}_y \; P[\,Y(t+1) = y \mid X(t) = x\,] \\ &= \operatorname*{argmax}_y \; q_y\, f_y(x) \quad . \end{aligned} \tag{6}$$

  Compute $y^*(x)$ numerically or analytically. At what values of x does y^* change? Mark these critical values $\{x^*\}$ on the graph of the PDFs.

6 Association Rules

We would also like to predict Y after observing its k most recent values. Considering $k = 5$, we have $3^{k+1} = 729$ possible association rules in the form of

$$\{Y(t-4), Y(t-3), Y(t-2), Y(t-1), Y(t)\} \to Y(t+1) = y\ . \tag{7}$$

- Report the 10 rules with the highest support and the 10 rules with the highest confidence in two tables.

6.1 Usefulness of Association Rules

The usefulness u of a rule should be proportional to its support s and its confidence c because a useful rule is both frequent (thus yielding a high support) and accurate (thus yielding a high confidence). Still, support and confidence may not be equally important, so we may assume $u \sim s^{1-\lambda} c^{\lambda}$ for some tuning parameter $\lambda \in [0, 1]$. As λ rises, the emphasis of this quantity smoothly slides from support to confidence.

If $\lambda = \frac{1}{2}$, $u \sim \sqrt{sc}$, i.e. the geometric mean of support and confidence. If $\lambda = \frac{2}{3}$, $u \sim \sqrt[3]{sc^2} = \sqrt[3]{r}$, where r is a rule's **rule power factor** (RPF).

- Report the 10 rules with the highest geometric mean and the 10 rules with the highest RPF in two tables.
- Do you think there is an optimal value for λ, which best measures a rule's usefulness? If yes, what may it be? If no, why not?
- Overall, can you observe any useful rules?

7 A Portfolio with One Stock and Money

We have studied three investment tools, viz. MACD, Bayes detector, and association rules, with the past data. It is time to do some simulations with the future data and check if they really work. In this section, we will first build a market model to trade the riskier stock with money.

Let $M(t)$ be the amount of your money and $N(t)$ the number of your shares at the end of day t, so your portfolio's value $V(t)$ is defined as $V = M + NS$. You are initially given the minimum-risk portfolio that has $V(0) = 100{,}000$ dollars (or whatever currency that describes your data); in other words, you only have $M(0) = 100{,}000$ dollars at the beginning. Then you start trading at $t = 1$ according to the prediction by the tools.

You are allowed to trade at most once per day. (After all, this market deals with daily data, and the stock's price does not change during a day.) If you decide to trade on day t,

- you may spend $m = gM(t-1)$ dollars buying $m/S(t)$ shares, or
- you may sell $n = gN(t-1)$ shares to get $nS(t)$ dollars.

The parameter $g \in (0, 1)$ quantifies your "greed". The greedier you are, the more you want to earn and thus the more you trade per transaction.

Meanwhile, M is compounded with a daily interest rate r at the beginning of each day, so the minimum-risk portfolio's value amounts to $V_0(t) = M(0) \times (1+r)^t$. Set $r = 0.001\%$ to model the fact that many ten-year governmental bonds yield around 4% simple interest.

- When do you buy, sell, or do nothing? How do you follow the various prediction tools?
- Plot $V(t, g)$ for two values of g atop $V_0(t)$. Let the larger g be the "aggressive greed" g_A and the smaller g be the "conservative greed" g_C. Are the greedy portfolios better than the minimum-risk portfolio?
- You are Alice, Bob, and Charlie's banker. Alice and Charlie invest with $g = g_A$ and $g = g_C$ respectively, whereas Bob wants to strike a balance between them and lets you decide when to use $g = g_A$ and when to use $g = g_C$.

  Discuss how you should vary Bob's greed g_B so that his portfolio may outperform Alice's and Charlie's, then plot $V(t, g_B)$ on the previous graph.

8 A Portfolio with Two Stocks

Now consider a market model in which you are going to trade the riskier stock using the safer stock instead of money.

Let $N_1(t)$ and $N_2(t)$ be the number of your shares in the riskier stock and the safer stock, so your portfolio's value is $V = N_1 S_1 + N_2 S_2$. Again, you are initially given the minimum-risk portfolio that has $V(0) = 100{,}000$ dollars, then you can trade once per day like last section. You may sell $n_i = gN_i(t-1)$ shares in either stock to buy $n_i S_i(t)/S_{3-i}(t)$ shares in the other stock on day t.

However, here is a deliberate assumption: you will trade solely according to the prediction about S_1, and you do not try to predict S_2 at all. It means that you must buy or sell the riskier stock on the days when you bought or sold it last section.

- Following your findings in Section 1, compute the initial number of shares $N_1(0) = p_0 V(0)/S_1(0)$ and $N_2(0) = (1 - p_0)V(0)/S_2(0)$. How do you estimate the riskier stock's fraction p_0 in the minimum-risk portfolio?
- Plot $V(t, g_A)$ and $V(t, g_C)$ atop the minimum-risk portfolio's value $V_0(t)$. Is it worth being greedy this time?

8.1 Efficient Frontier

The discussed scheme of investment is probably suboptimal since we may therewith buy too many shares in the safer stock, resulting in a portfolio that yields a return even lower than the minimum-risk portfolio. We can prevent this situation by taking the stocks' efficient frontier into account.

Since the portfolio's risk $\sigma(t)$ is always between the minimum-risk portfolio's risk σ_0 and the riskier stock's risk σ_1, instead of setting $n_i = gN_i(t-1)$, you may set n_i in a way so that after each trade,

- the portfolio becomes safer and yields $\sigma(t) = \sigma(t-1) - g[\sigma(t-1) - \sigma_0]$, or
- the portfolio becomes riskier and yields $\sigma(t) = \sigma(t-1) + g[\sigma_1 - \sigma(t-1)]$.

Hence, the portfolio's return is always at least equal to that of the minimum-risk portfolio.

- Plot $p(t) = N_1(t)S_1(t)/V_1(t)$ for the original scheme of investment, then compare it to the minimum-risk fraction $p_0(t)$. Are there undesirable moments at which $p(t) < p_0(t)$?

  [Hint: while $p(t)$ changes with time because of trading, $p_0(t)$ changes with time just because of updated estimation, so it is expected to change much slower than $p(t)$.]
- Plot $V(t, g_A)$ and $V(t, g_C)$ atop $V_0(t)$ again for the new scheme of investment. Do they perform better after considering the efficient frontier?

9 A Portfolio with Two Stocks and Money

Finally, we will work on a more realistic market model: you can now predict both stocks' prices and trade them with money independently, so you are going to manage a portfolio that comprises three assets.

You are initially given $M(0) = 100{,}000$ dollars and $N_1(0) = N_2(0) = 0$ shares in the stocks, then you can trade each stock at most once per day. Note that the order of trading matters. For example, if you have sold the shares in a stock beforehand, you will have extra money to buy more shares in the other stock.

And like Section 7, the money is compounded with a rate $r = 0.001\%$ at the beginning of each day.

- Combining all the techniques that you have learnt from this report, this course, or other sources, devise a scheme of investment in terms of greed g, then plot $V(t, g_A)$ and $V(t, g_C)$. Comment on the performance of your scheme.

Acknowledgments. While your groupmates are not your co-authors, you must acknowledge their contribution at the end of the report for academic etiquette.