Association Rules and Prediction Rules for Financial Data Mining

Zhuoyue WAN

The Hong Kong University of Science and Technology, Hong Kong SAR, China
zwanah@connect.ust.hk

Abstract. This report analyzes two stocks, AAPL and AMD, using data from Yahoo Finance. Daily returns are computed for both stocks. For AAPL, the report examines the w-day Exponential Moving Average (EMA), Cumulative Distribution Function (CDF), and Probability Density Function (PDF). Mean-Variance Analysis is then applied to AAPL and AMD to determine the optimal portfolio. Association rules are studied, focusing on four metrics: confidence, geometric mean, arithmetic mean, and rule power factor. The report concludes that geometric mean and rule power factor are better suited to assess association rule quality, as they consider both rule frequency and accuracy.

Keywords: Exponential moving average · CDF · PDF · Mean-Variance Analysis · Optimal portfolio.

1 Data Preprocessing

In this section, I choose the AAPL stock [1] and AMD stock [2] in Yahoo Finance as the data source. Then I compute the daily return of these two stocks with the given formula:

$$X(t) = ln\left[\frac{S(t)}{S(t-1)}\right] \tag{1}$$

2 Exponential Moving Average

In the context of financial data analysis, an exponential moving average (EMA) is often used to smooth out fluctuations and highlight trends in time-series data. In this section, I compute AAPL's w-day EMA with the given formula:

$$M(t,w) = aS(t) + (1-a)M(t-1,w) \tag{2}$$

Then I plot the $M(t,w)$ for $w = 30$, 100 and 300 when $a = 1/w$ and $a = 2/(w+1)$ in Fig. 1 and Fig. 2 respectively.

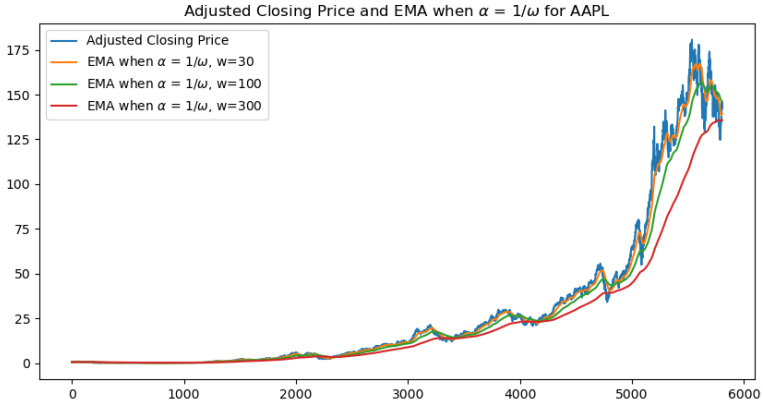

Fig. 1: $M(t, w)$ for $w = 30$, 100 and 300 when $a = 1/w$

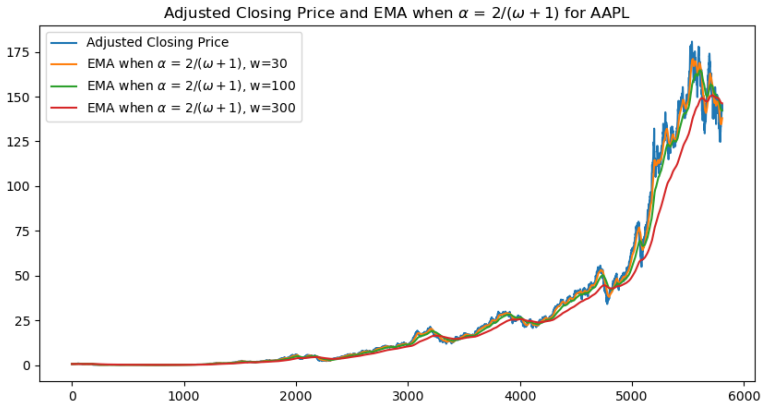

Fig. 2: $M(t, w)$ for $w = 30$, 100 and 300 when $a = 2/(w + 1)$

By observing the two figures, it becomes clear that the window length w and smoothing constant a play a significant role in determining the performance of EMA as a smoothing filter for financial data.

A longer window length w can provide more smoothing, resulting in a reduced impact of noise or short-term fluctuations in the data. However, an extended window length can introduce more lag, which in turn reduces the responsiveness of the filter to changes in the data.

The weight assigned to recent versus older data points in the EMA calculation is determined by the smoothing constant (alpha). A smaller alpha value

gives more weight to older data points, producing a smoother output, while a larger alpha value gives more weight to recent data points, resulting in a more responsive output. Therefore, it is clear that EMA fluctuations with the same window length w are more noticeable when $a = 2/(w+1)$ than when $a = 1/w$.

3 Cumulative Distribution Function

Regard the values of $X(t)$ as realizations of a random variable X. Plot the cumulative distribution function (CDF) of X in Fig. 3.

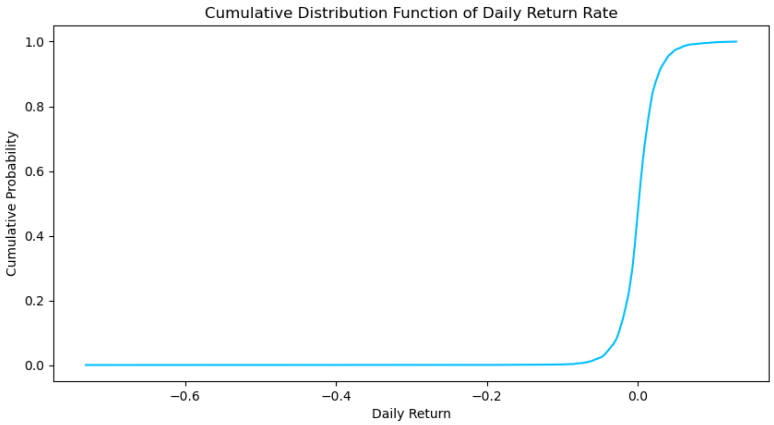

Fig. 3: Cumulative distribution function $F_X(x)$

3.1 Logistic Funcion

Fit $F_X(x)$ with the logistic function:

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

According to this fomular, we know $L(x^*) = 0.5$. By fitting the fomular with python, we know the empirical x^* is 1.09×10^{-3} and the empirical b is about 90.70. So we can calculate $L^{'}(0) = 22.62$. Then plot empirical $L(x)$ atop $F_X(x)$ in Fig. 4.

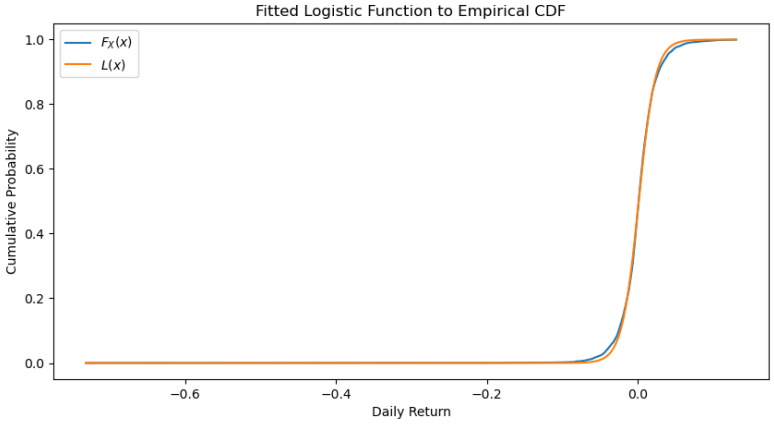

Fig. 4: Empirical$L(x)$ atop $F_X(x)$

3.2 Kolmogorov-Smirnov Test

We can evaluate the goodness of fit of $L(x)$ to $F_X(x)$ using the Kolmogorov-Smirnov test. Our null hypothesis is that $L(x)$ fits $F_X(x)$ well. To test this hypothesis, we define D as the maximum deviation between $F_X(x)$ and $L(x)$, i.e., $D = \max x\,|F_X(x) - L(x)|$. If $\sqrt{N}D > \eta\alpha$, we reject the null hypothesis at a significance level α. Here, the threshold η_α is determined by solving the following equation:

$$\frac{\sqrt{2\eta_\alpha}}{\eta_\alpha}\sum_{k=1}^{\infty}\exp\left[-\frac{(2k-1)^2\pi^2}{8\eta_\alpha^2}\right] = 1-\alpha \tag{4}$$

By performing the Kolmogorov-Smirnov test, we obtain a p-value of 0.007, which is less than the significance level of 0.05. As a result, we reject the null hypothesis in favor of the alternative, suggesting that there is insufficient evidence to support the claim that $L(x)$ fits $F_X(x)$ well.

4 Probability Density Function

On one hand, we can estimate X''s PDF $f_X(x)$ with the derivative of its fitted CDF. On the other hand, we can estimate $f_X(x)$ with a k-bin normalized histogram, where each bin is $h = (\max x - \min x)/k$ units wide. In general, the i th bin measures the frequency of $x \in [\min x + (i-1)h, \min x + ih)$.

Firstly, we can derive the formula of $L^{'}(x)$ as follows:

$$L^{'}(x) = \frac{b\exp\left[-b(x-x^*)\right]}{1+2*\exp\left[-b(x-x^*)\right]+\exp\left[-2b(x-x^*)\right]} \tag{5}$$

We can plot $L^{'}(x)$ in Fig.5 and compare it with five histograms generated using different numbers of bins. Specifically, we plot histograms with k values of 20,

100, and 400, as well as histograms with k values determined using the Sturges formula and the Freedman-Diaconis formula. By observing Fig.5, we can see that the number of bins can significantly affect the appearance and interpretation of the resulting histogram.

Firstly, the number of bins directly affects the width of the bins in the histogram. A smaller number of bins will result in wider bins, while a higher number of bins will result in narrower bins. Secondly, the number of bins also affects the representation of the data distribution in the histogram. A small number of bins may not capture important features of the data distribution, such as multiple modes or skewness. In contrast, a large number of bins can introduce noise and overemphasize minor fluctuations in the data. Therefore, choosing an appropriate number of bins is crucial in generating an accurate and informative histogram.

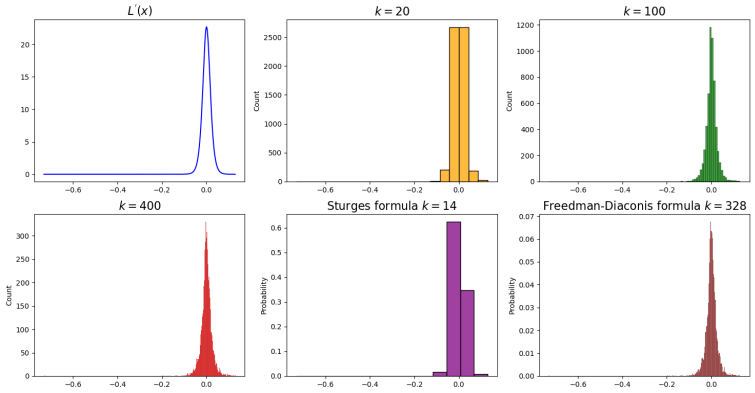

Fig. 5: $L^{'}(x)$ and histograms with different bins

5 Descriptive Statistics

In this part, we calculate the daily return of the AMD stock. Based on the data, we have computed the following all-time statistics for the two return rates:

- Means: $\mu_1 = 0.000884$ and $\mu_2 = 0.000272$
- Variances: $\sigma_1^2 = 0.000676$ and $\sigma_2^2 = 0.001553$
- Sharpe ratios: $\gamma_1 = 0.726745$ and $\gamma_2 = 0.393288$
- Covariance: $\sigma_{12} = 0.000374$

Then repeat last step using only the data on the K most recent days for K = 30, 100, 300. The result is as follows:

1. When $K = 30$, the means, variances, Sharpe ratios and covariance are:

  - Means: $\mu_1 = -0.000281$ and $\mu_2 = -0.002414$
  - Variances: $\sigma_1^2 = 0.000676$ and $\sigma_2^2 = 0.001553$
  - Sharpe ratios: $\gamma_1 = -0.089444$ and $\gamma_2 = -0.743423$
  - Covariance: $\sigma_{12} = 0.000588$
2. When $K = 100$, the means, variances, Sharpe ratios and covariance are:
  - Means: $\mu_1 = -0.000761$ and $\mu_2 = -0.000577$
  - Variances: $\sigma_1^2 = 0.000568$ and $\sigma_2^2 = 0.001349$
  - Covariance: $\sigma_{12} = 0.000635$
  - Sharpe ratios: $\gamma_1 = -0.369944$ and $\gamma_2 = 0.004296$
3. When $K = 300$, the means, variances, Sharpe ratios and covariance are:
  - Means: $\mu_1 = -0.000281$ and $\mu_2 = -0.002414$
  - Variances: $\sigma_1^2 = 0.000475$ and $\sigma_2^2 = 0.001440$
  - Covariance: $\sigma_{12} = 0.000588$
  - Sharpe ratios: $\gamma_1 = -0.089444$ and $\gamma_2 = -0.743423$

6 Mean-Variance Analysis

We would like to perform a mean-variance analysis on $X_1(t)$ and $X_2(t)$ and accordingly construct the minimum-risk portfolio $S_p(t) = pS_1(t) + (1-p)S_2(t)$. For some fraction of investment $p \in [0, 1]$. We can determine the value of p according to the following formula:

$$\begin{cases} a = \dfrac{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}{(u_1 - u_2)^2} \\ b = \dfrac{-2\left[u_2\sigma_1^2 + u_1\sigma_2^2 - (u_1 + u_2)\,\sigma_{12}\right]}{(u_1 - u_2)^2} \\ c = \dfrac{u_2^2\sigma_1^2 + u_1^2\sigma_2^2 - 2u_1u_2\sigma_{12}}{(u_1 - u_2)^2} \end{cases}. \tag{6}$$

At $u_p^* = -\dfrac{b}{2a}$, the portfolio possesses the minimum risk $\sigma_p^{*2} = c - \dfrac{b^2}{4a}$ and $p = \dfrac{\mu_p^* - \mu_2}{\mu_1 - \mu_2}$. Using these formulas, we can get $p \approx 0.7961$. Using this value of p, we can plot the resultant portfolio $S_p(t)$ atop $S_1(t)$ and $S_2(t)$ in Fig. 6.

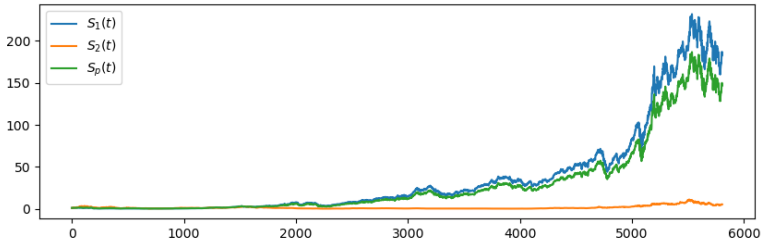

Fig. 6: The resultant portfolio $S_p(t)$ atop $S_1(t)$ and $S_2(t)$

6.1 A K-Day Analysis

As the relevance of old data should decay, it is more sensible to consider the stock's performance on the K most recent days only. In other words, we only infer information from $\{X_{i=1,2}(t-\tau+1) \mid \tau \in [1,K]\}$ at evey moment t. Hence, the fraction of investment, now denoted by $p(t,K)$, varies with time and depends on K. Complete the following tasks for $K = 30, 100, 300$. Then plot $p(t,K)$ for $K = 30, 100, 300$ in Fig. 7, plot the resultant portfolio $S_p(t)$ atop $S_1(t)$ and $S_2(t)$ for $K = 30, 100, 300$ in Fig. 8, plot $S_p(t,k)$'s K-day Sharpe ratio $R_p(t,K)$ for $K = 30, 100, 300$ in Fig. 9.

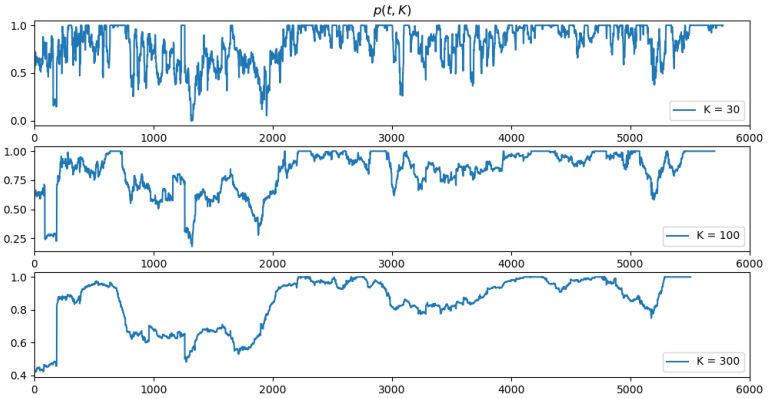

Fig. 7: $p(t,K)$ for $K = 30, 100, 300$

Fig. 8: The resultant portfolio $S_p(t)$ atop $S_1(t)$ and $S_2(t)$

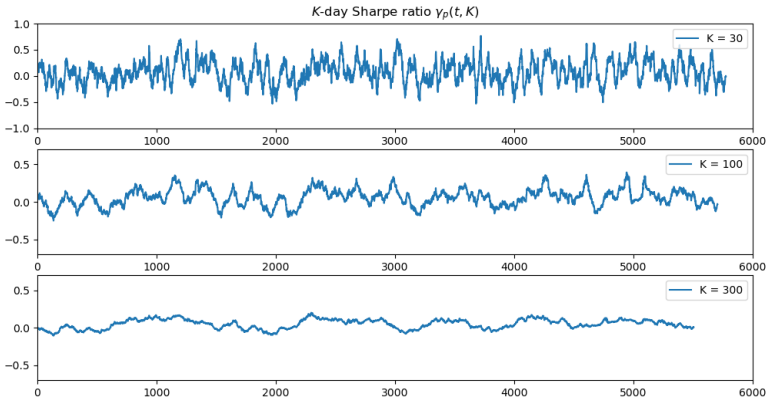

Fig. 9: $S_p(t,k)$'s K-day Sharpe ratio $R_p(t,K)$ for $K = 30, 100, 300$

By observing Fig.7, we can see that the fluctuations of $p(t,K)$ decrease as the value of K increases. This is because more recent data is more relevant to the current stock price and contributes to more violent and frequent fluctuations in $p(t,K)$. In other words, as the value of K decreases, portfolios will respond more frequently to the market. This can be verified by observing the other two figures. In Fig.8, the resultant portfolio $S_p(t)$ when $K = 30$ has the best performance since it is more sensitive to the market. In Fig.9, $R_p(t,30)$ fluctuates more violently than $R_p(t,100)$ and $R_p(t,300)$, which is consistent with the observation in Fig.7. However, in another view, the more frequent fluctuations in $p(t,30)$ means that the portfolio takes higher risks in pursuit of higher returns compared to the other two portfolios.

7 Digitization of Time Series

Let's focus on the first stock AAPL. We can digitize $X(t)$ as $Y(t)$ using three alphabets: D for "down", U for "up", and H for "hold". The following rules can be used:

$$Y(t) = \begin{cases} \mathrm{D} & [X(t) < -\varepsilon] \\ \mathrm{U} & [X(t) > +\varepsilon] \\ \mathrm{H} & (\text{ otherwise}) \end{cases} \tag{7}$$

Here, we will use $\varepsilon = 0.002$. We can then calculate the probability $P[Y(t) = y]$ for $y \in \mathrm{D}, \mathrm{U}, \mathrm{H}$. The result is as follows:

- U (up): 0.471409
- D (down): 0.424733
- H (hold): 0.103858

Then I calculate the conditional probability $P[Y(t+1) = y_1 \mid Y(t) = y_2]$ for all nine possible pairs of $(y_1, y_2) \in \{\mathrm{D}, \mathrm{U}, \mathrm{H}\} \times \{\mathrm{D}, \mathrm{U}, \mathrm{H}\}$. The result is as follows:

- DD: 0.1772609819121447
- DU: 0.2053402239448751
- DH: 0.04220499569336779
- UD: 0.20378983634797587
- UU: 0.2186046511627907
- UH: 0.04892334194659776
- HD: 0.043583118001722654
- HU: 0.047545219638242896
- HH: 0.012747631352282515

8 Association Rules

We would like to find out a five-day pattern A that associates well with an immediate down (D). Formally, a rule R can be written as

$$R : A = \{Y(t-4), Y(t-3), Y(t-2), Y(t-1), Y(t)\} \to Y(t+1) = \mathrm{D} \quad (8)$$

There are $3^5 = 243$ possible rules that could be used to digitize the AAPL stock's price movement.

To analyze the stock's behavior, we can divide $Y(t)$ into the "past" and the "future" at $t = M \approx 3N/4$, where N is the total number of days. By examining the history of the previous $3N/4$ days, we hope to identify good association rules that can be applied to the coming $N/4$ days to make a profit.

To identify the best association rules, we can calculate the confidence of all 243 rules in the past and report the top 10 rules with the highest confidence, denoted as $\{R_{conf}\}_p$ as follows:

- $DDDHH$
- $DUHHD$
- $DHDHU$
- $DHUHD$
- $DHUHH$
- $DHHDH$
- $DHHUD$
- $DHHUU$
- $UDDHH$
- $UDHDH$

9 Verification of Association Rules

We may verify a rule's goodness by doing a betting experiment with the future data like this: as time passes, we bet on an immediate down every day. If a down

indeed comes out, we earn $\$u$, and we ascribe the profit to the just appeared five-day pattern A and its rule $R : A \to \mathrm{D}$; otherwise, we lose $\$v$, and we similarly ascribe the loss to the just appeared pattern.

Consider $u = 1$ and $v = 0$ for simplicity. In this regard, a rule's profit is merely the number of times that it works. I record the profits due to all the 243 rules, then report the 10 most profitable rules $\{R_{\text{expt}}\}_f$ as follows:

- $DUUDD$ (Work times = 21)
- $UDDDU$ (Work times = 21)
- $UDUUD$ (Work times = 21)
- $UUDUU$ (Work times = 21)
- $UUUUU$ (Work times = 20)
- $DDDUD$ (Work times = 19)
- $DUUUD$ (Work times = 18)
- $DDUUU$ (Work times = 17)
- $DUDDU$ (Work times = 17)
- $DUUUU$ (Work times = 17)

Upon observation, we found that there are no association rules in $\{R_{\text{conf}}\}_p$ that share the same pattern with $\{R_{\text{expt}}\}_f$.We computed the Pearson correlation coefficient between a rule's confidence in the past and its profit in the future and found it to be approximately 0.1178. Although this value is not zero, it is still considered small, which means that the confidence of a rule is not associated with $\{R_{\text{expt}}\}_f$ well. This may be due to that $\{R_{conf}\}_p$ does not account for the frequency of a rule's pattern appearing in the whole dataset. In other words, a rule may have a high confidence, but its pattern may appear very rarely in the dataset. When we apply such a rule to future data, it may not work well since the pattern is not frequent enough to make accurate predictions. Therefore, it is important to consider the frequency of a rule's pattern in the dataset and evaluate its effectiveness in predicting future outcomes.

10 Further Analysis of Association Rules

Let us analyse the betting experiment more carefully. One can prove that a rule's profit π is related to its future support s_f and future confidence c_f via

$$\pi \sim s_f \left[uc_f - v\left(1 - c_f\right)\right] \tag{9}$$

where the proportionality depends on the length of the future. This matches our intuition: a rule is good if it is both frequent (thus a high support) and accurate (thus a high confidence). As $v = 0$, the formula becomes $\pi \sim s_f c_f$.

The remaining problem is that we can never know s_f and c_f but are only able to estimate them with the rule's past support s_p and past confidence c_p, which may hugely deviate from their future counterparts.

10.1 Geometric mean and Arithmetic mean

Let us first boldly assume that all rules obey $s_p = s_f$ and $c_p = c_f$, yielding $\pi \sim s_p c_p$. Since the ranking of π does not change when we take a square root on the right-hand side, we may predict a rule's goodness with the geometric mean between its support and confidence. Out of curiosity, we may also guess whether their arithmetic mean is useful.

Firstly, report the 10 rules with the highest geometric mean between support and confidence in the past as $\{R_{geo}\}_p$:

- $DDUDU$: Geometric mean: 0.1109
- $DUDUU$: Geometric mean: 0.1109
- $UDDDU$: Geometric mean: 0.1096
- $DDUUD$: Geometric mean: 0.1083
- $DUDDU$: Geometric mean: 0.1083
- $UDUUD$: Geometric mean: 0.1083
- $UDDUU$: Geometric mean: 0.1056
- $UDDUD$: Geometric mean: 0.1056
- $UDUDD$: Geometric mean: 0.1043
- $UUDDU$: Geometric mean: 0.1043

Secondly, report the 10 rules with the highest arithmetic mean between support and confidence in the past as $\{R_{ari}\}_p$:

- $UUHHU$: Arithmetic mean: 0.5004
- $HDHDU$: Arithmetic mean: 0.5004
- $DDDHH$: Arithmetic mean: 0.5003
- $DUHHD$: Arithmetic mean: 0.5003
- $DHUHD$: Arithmetic mean: 0.5003
- $UDHDH$: Arithmetic mean: 0.5003
- $UUDHH$: Arithmetic mean: 0.5003
- $HDDHD$: Arithmetic mean: 0.5003
- $HDUUH$: Arithmetic mean: 0.5003
- $HHDDD$: Arithmetic mean: 0.5003

Then we can find that for $\{R_{geo}\}_p$ and $\{R_{expt}\}_f$:

- The rule $UDDDU$ is in both dictionaries
- The rule $DUDDU$ is in both dictionaries
- The rule $UDUUD$ is in both dictionaries

And we can also find that $\{R_{ari}\}_p$ and $\{R_{expt}\}_f$ do not share any rules. So it seems that the geometric mean assesses a rule better than the arithmetic mean and confidence.

10.2 A Gneralized Mean

Now we will account for the discrepancy between the past and the future, so $s_p \neq s_f$ and $c_p \neq c_s$. We are pessimistic, so we expect that a rule's confidence depreciates over time, but we are also optimistic, so we expect that a more confident rule in the past remains more confident in the future.

The two assumptions combine to suggest $c_f = c_p^m$ for some $m > 1$. With the scale of π maintained, we may formulate π with a generalized mean

$$\pi = \left(s_p c_p^m\right)^{\frac{1}{1+m}} \equiv s_p^\lambda c_p^{1-\lambda}$$

for some tuning parameter $\lambda \equiv 1/(1+m) \in [0,1]$. As λ rises, the emphasis of $s_p^\lambda c_p^{1-\lambda}$ smoothly slides from support to confidence. When λ strikes $1/3, \pi = \sqrt[3]{s_p c_p^2} \equiv \sqrt[3]{r_p}$, where r_p is the rule's rule power factor (RPF).

Then, report the 10 rules with the highest RPF in the past as $\{R_{RPF}\}_p$:

- $UDUUD$: RPF: 0.1892
- $DUDUU$: RPF: 0.1877
- $DUDDU$: RPF: 0.1866
- $DDUDU$: RPF: 0.1862
- $UDDDU$: RPF: 0.1848
- $DDUUD$: RPF: 0.1834
- $UDUDD$: RPF: 0.1790
- $UDDUU$: RPF: 0.1774
- $UDDUD$: RPF: 0.1724
- $DDDDU$: RPF: 0.1710

Then we can find that for $\{R_{RPF}\}_p$ and $\{R_{expt}\}_f$:

- The rule $UDUUD$ is in both dictionaries
- The rule $DUDDU$ is in both dictionaries
- The rule $UDDDU$ is in both dictionaries

Determining the optimal value for the rule power factor (RPF) λ that best predicts a rule's goodness can be challenging, as it may depend on the specific task and dataset. The choice of RPF can significantly impact the performance of the association rule mining algorithm. However, what constitutes an optimal value may vary depending on the characteristics of the data and the task at hand. Typically, a higher RPF prioritizes rule confidence over support, while a lower RPF prioritizes support overconfidence. The appropriate value of RPF for a given problem may depend on factors such as the dataset's complexity, the number of items involved, and the desired balance between rule support and confidence.

References

1. AAPL. (n.d.). Apple Inc. [Yahoo homepage]. Retrieved from https://finance.yahoo.com/quote/AAPL?p=AAPL&.tsrc=fin-srch
2. AMD. (n.d.). Advanced Micro Devices, Inc. [Yahoo homepage]. Retrieved from https://finance.yahoo.com/quote/AMD?p=AMD&.tsrc=fin-srch