Quant Questions

Columbia Business School & Quant Markets and Trading Club

February 2025

Contents

Introduction

This document contains up-to-date quant-related real questions along with their solutions. The problems cover coding questions, brainteasers, probability, finance, case questions, and machine learning.

1 Coding Questions

1.1 Python

Question 1: Python's `zip` Function

Question: What is the `zip` function in Python, and how is it used?

Solution: The `zip` function combines two or more iterables element-wise into pairs or tuples. Example:

```
list1 = [1, 2, 3]
list2 = ['a', 'b', 'c']
print(list(zip(list1, list2)))  # Output: [(1, 'a'), (2, 'b'), (3, 'c')]
```

Question 2: Python's `tuple` and `list`

Question: What is the differences between `list` and `tuple` in Python?

Solution: The major differences include:

1. `list` is mutable while `tuple` is not.
2. `list` is defined using brackets `[]` while `tuple` is defined using parentheses `()`.
3. `list` is slower and consumes more memories compared with `tuple`.
4. `list` has more built-in methods such as `append()`, `remove()`, `pop()`, `sort()`, while `tuple` has limited methods, mainly `count()` and `index()`.

Question 3: Merging data with Pandas

Question: What are the differences between `merge`, `concat` and `join` in Python?

Solution:

1. `merge`
   (a) The `merge()` function used to merge the DataFrames with database-style join such as inner join, outer join, left join, right join.
   (b) Combining exactly two DataFrames.
   (c) The join is done on columns or indexes.
   (d) If joining columns on columns, the DataFrame indexes will be ignored.
   (e) If joining indexes on indexes or indexes on a column, the index will be passed on.
2. `join`

(a) The `join()` function used to join two or more pandas DataFrames/Series horizontally.

(b) `Join()` uses merge internally for the index-on-index (by default) and column(s)-on-index join.

(c) Aligns the calling DataFrame's column(s) or index with the other objects' index (and not the columns).

(d) Defaults to left join with options for right, inner and outer join.

3. `concat`

   (a) Concatenate two or more pandas DataFrames/Series vertically or horizontally.

   (b) Aligns only on the index by specifying the axis parameter.

   (c) Defaults to outer join with the option for inner join.

1.2 Machine Learning

Question 1: Logistic Regression Assumptions

Solution: Key assumptions include:

- The dependent variable is binary.
- Linear relationship between independent variables and the log-odds.
- Independent observations.
- No multicollinearity among predictors.

Introduction to Logistic Regression on YouTube.

Question 2: Understanding XGBoost

Solution: XGBoost is an optimized gradient boosting algorithm known for speed and accuracy in large datasets. It handles missing values and prevents overfitting through regularization. For further reference, see: **Introduction to XGBoost on YouTube**.

Question 3: Understanding PCA

Solution: PCA is a linear transformation of dataset to orthogonal components while remaining most of the original information.

Question 4: Understanding Underfitting and Overfitting

Solution:

1. Underfitting
   (a) **What is underfitting?** Underfitting happens when a model is too simple to capture the underlying patterns in the data. As a result, it performs poorly on both training data and test data, meaning it fails to learn effectively.

   (b) **Reasons for underfitting**
      i. High bias and low variance.
      ii. The model is too simple, So it may be not capable to represent the complexities in the data.
      iii. The input features which is used to train the model is not the adequate representations of underlying factors influencing the target variable.
      iv. The size of the training dataset used is not enough.
      v. Excessive regularization are used to prevent the overfitting, which constraint the model to capture the data well.
      vi. Features are not scaled.

   (c) **How to detect underfitting?**
      i. Holdout validation and cross-validation
      ii. Plot the learning curve (the training and validation error against training set size)

   (d) **How to address underfitting?**
      i. Increase model complexity.
      ii. Increase the number of features, performing feature engineering.
      iii. Remove noise from the data.
      iv. Increase the number of epochs or increase the duration of training to get better results.

2. Overfitting
   (a) **What is overfitting?** Overfitting happens when a model learns not only the underlying pattern in the training data but also noise and outliers, making it perform exceptionally well on training data but poorly on unseen test data.

   (b) **Reasons for overfitting**
      i. High variance and low bias.
      ii. The model is too complex.
      iii. Limited size of the training data.

   (c) **How to detect overfitting?**
      i. Holdout validation and cross-validation
      ii. Plot the learning curve (the training and validation error against training set size)

   (d) **How to address overfitting?**

i. Improving the quality of training data reduces overfitting by focusing on meaningful patterns, mitigate the risk of fitting the noise or irrelevant features.

ii. Increase the training data can improve the model's ability to generalize to unseen data and reduce the likelihood of overfitting.

iii. Reduce model complexity.

iv. Early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).

v. Ridge Regularization and Lasso Regularization.

vi. Use dropout for neural networks to tackle overfitting.

2 Brainteaser & Probability

2.1 Probability

Question 1: Probability with Two Dice

Question: Two dice are rolled. What is the probability that the value on the second die is greater than the value on the first die?

Solution: There are $6 \times 6 = 36$ possible outcomes. The cases where $D_2 > D_1$ include:

$$\begin{bmatrix} - & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ & - & (2,3) & (2,4) & (2,5) & (2,6) \\ & & - & (3,4) & (3,5) & (3,6) \\ & & & - & (4,5) & (4,6) \\ & & & & - & (5,6) \\ & & & & & - \end{bmatrix}$$

There are 15 favorable outcomes, so:

$$P(D_2 > D_1) = \frac{15}{36} = \frac{5}{12}$$

Question 2: Points on a Circle

Question: Three points are randomly chosen on the circumference of a circle. What is the probability that the triangle formed by these points contains the center of the circle?

Solution: We show that $\frac{1}{4}$ of the time, the center is inside the triangle, based on ideas of symmetry:

1. **Special "Square" Scenario:** Place Points 1 and 2 so that they resemble a square diagonal (near 90° apart). In this configuration, the *red arc* (where Point 3 can land to ensure the center is inside) takes up $\frac{1}{4}$ of the circle's circumference.

2. **General Scenario:** For any random arrangement of Points 1 and 2, there exists a unique symmetrical configuration by *flipping* or rotating them to match the original case. The "opposite" configuration again pinpoints a $\frac{1}{4}$ arc where Point 3 must lie. When we combine both scenarios (the original and its flip), a total of $\frac{1}{2}$ of the circle is covered. However, because each scenario has an equal chance of occurring, the overall probability remains $\frac{1}{4}$ that Point 3 falls within the necessary arc for the center to be contained.

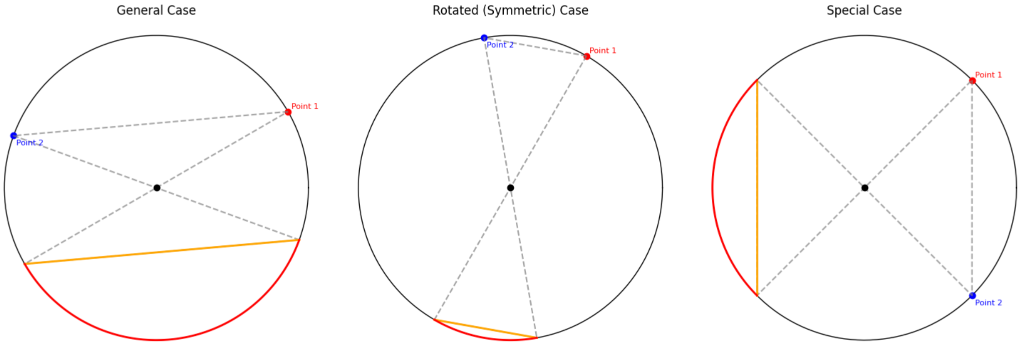

Figure 1: Special and Symmetric Cases of points distribution

Hence, the probability that the triangle formed by three randomly chosen points on the circle contains the center is

$$\boxed{\frac{1}{4}}.$$

Question 4: Drawing Socks from Drawers

Question: There are 3 drawers:

- The first drawer contains infinite red and green socks in a 1:1 ratio.
- The second drawer contains infinite green and blue socks in a 1:1 ratio.
- The third drawer contains infinite red and blue socks in a 1:1 ratio.

You do not know which two colors are in any of the drawers. What is the expected number of socks you need to draw to determine the combination of colors in each drawer?

2.2 Strategy

Question 3: Number Choosing Game

Question: You and I play a game of choosing numbers in turn. In each round, one of us picks an integer from 2 to 30 but cannot pick a number that shares a common divisor with any previously picked numbers (e.g., if 4 is chosen, no even numbers can be picked). The player who cannot choose a number loses. Would you start first? What is your strategy?

2.3 Discrete Math

Question 1: Chocolate Bar

Question: There is a 6x8 rectangular chocolate bar made up of small 1x1 bits. We want to break it into 48 bits. We can break one piece of chocolate horizontally or vertically, but cannot break two pieces together! What is the minimum number of breaks required?

Solution: For a chocolate of size mxn, we need mn - 1 steps. By breaking an existing piece horizontally or vertically, we merely increase the total number of pieces by one. Starting from 1 piece, we need mn - 1 steps to get to mn pieces. Another way to reach the same conclusion is to focus on ”bottom left corners of squares”: Keep the chocolate rectangle in front of you and start drawing lines corresponding to cuts. Each cut ”exposes” one new bottom left corner of some square. Initially, only one square's bottom left corner is exposed. In the end, all mn squares have their bottom left corners exposed.

Question 2: Secret Safe

Question: A group of 5 people want to keep their secret document in a safe. They want to make sure that in the future, only a majority (¿=3) can open the safe. So they want to put some locks on the safe, each of the locks have to be opened to access the safe. Each lock can have multiple keys; but each key only opens one lock. How many locks are required at the minimum? How many keys will each member carry?

Solution: For each group of 2 ppl, there must be a lock which none of them have a key to. But the key of such a lock will be given to the remaining 3 ppl of group. Thus, we must have atleast 5C2 = 10 Locks. Each lock has 3 keys, which is given to unique 3-member subgroup. So each member should have 10*3/5 = 6 keys.

Question 3: Trailing Zeros

Question: Calculate the number of trailing zero of 1000!.

Solution: Firstly, we know that the combination of 2 and 5 can generate one trailing zero. We should calculate the number of factors of 5 because the number of factors of 2 is much larger. Then, we know that combination of 4 and 25 can generate an additional trailing zero. Similarly, the combination of 8 and 125, 16 and 625 can also generate an additional one.

$$Z = \left\lfloor \frac{1000}{5} \right\rfloor + \left\lfloor \frac{1000}{25} \right\rfloor + \left\lfloor \frac{1000}{125} \right\rfloor + \left\lfloor \frac{1000}{625} \right\rfloor = 249$$

3 Finance Questions

3.1 Fixed Income

Question 1: Callable Bond Pricing

Question: One bond is priced at $500, and an identical but callable bond is also issued. Should the callable bond have a higher or lower price?

Solution: The callable bond should have a lower price because the call option benefits the issuer, allowing them to redeem the bond early when rates drop, creating reinvestment risk for the bondholder.

Question 2: Valuation of a Perpetual Bond

Question: A perpetual bond pays $1 per year indefinitely. What is the present value of this bond if the discount rate is r?

Solution: The present value of a perpetuity with a fixed payment D_0 is given by:

$$PV = \frac{D_0(1+g)}{r-g}$$

For a perpetual bond with no growth ($g = 0$) and $D_0 = 1$, this simplifies to:

$$PV = \frac{1}{r}$$

Question 3: Duration and Convexity

Question: What is duration in bond pricing, and is higher convexity better?

Solution: Duration measures a bond's sensitivity to interest rate changes. Convexity measures how the duration itself changes as interest rates change. Higher convexity is generally desirable because when interest rates rise, the bond price decreases less, and when interest rates fall, the bond price increases more compared to a bond with lower convexity.

3.2 Equities

Black Scholes

3.3 Miscellaneous

Question 1: Swaps

Question: Which leg of a floating-to-fixed interest rate swap is at an advantage when interest rates rise?

Solution: When interest rates rise, the fixed leg of a swap benefits while the floating leg is disadvantaged, as the fixed leg provides a stable payment while the floating leg will increase with rising market rates.

4 Econometrics Questions

Question 1: Autoregression and Stationarity

Question: What is autoregression, and what does it mean for a time series to be stationary?

Solution: Autoregression models a variable based on its past values (e.g., $y_t = \alpha + \beta y_{t-1} + \epsilon_t$). A stationary time series has constant mean, variance, and autocorrelation over time.

Question 2: Multicollinearity

Question: How to detect multicollinearity?

Solution: Calculate Variance Inflation Factor(VIF). If VIF is larger than 10, there is strong multicollinearity.

Question: How to address multicollinearity?

Solution:

1. **Remove Highly Correlated Variables** Identify and remove one of the correlated variables. This can be done using a correlation matrix or Variance Inflation Factor (VIF) analysis. A VIF value above 5 or 10 typically indicates problematic multicollinearity.

2. **Combine Variables** If two variables are measuring similar constructs, consider combining them into a single variable through methods like averaging, summation, or creating an index.

3. **Principal Component Analysis (PCA)** Use PCA to transform the correlated variables into a set of uncorrelated components. You can then use these components in your regression model.

4. **Regularization Techniques** Employ regularization methods like Ridge regression or Lasso regression. These techniques add a penalty to the loss function, which can help mitigate the impact of multicollinearity by shrinking the coefficients of correlated predictors.

5 Case Questions

Question 1: Predicting Bank Account Closures (Extreme Bank Account Withdrawals)

Question: What parameters and model would you use to predict large-scale bank account closures driven by customer?

Solution: Key parameters to consider include factors that may trigger customer panic and liquidity concerns:

- **Bank liquidity indicators:** Exposure to long-term Treasuries, liquidity ratios, and duration mismatch between assets and liabilities.

- **Market news sentiment:** Panic-inducing news such as rumors about bank solvency, sector-wide crises, or bank downgrades.

- **Customer withdrawal patterns:** Unusually high or frequent withdrawals, spikes in ATM or online banking transactions.
- **External economic factors:** Rising interest rates, inflation, or unexpected monetary policy changes.

Models to consider:

- **Logistic regression or decision trees:** For identifying key variables driving withdrawal probability.
- **Sentiment analysis:** To quantify panic through news articles and social media using natural language processing (NLP).
- **Early warning indicators:** Develop a bank-specific "panic score" based on liquidity metrics, abnormal withdrawals, and sentiment signals.

Question 2: Police Stations and Crime

Question: A simple linear regression is used to model the relationship between the number of police stations (independent variable) and the number of thieves (dependent variable). What issues could arise with this model?

Solution: The main issue is **simultaneity bias**, where the dependent variable (number of thieves) may influence the independent variable (number of police stations). This violates the assumption that the independent variable is exogenous.

Follow-up: How can we address this issue?

Solution: One approach is using an **instrumental variable (IV)**, which is correlated with the number of police stations but not directly with the number of thieves. For example, government funding for police expansion can be a valid instrument.

Question 3: Yen Depreciation and Japanese Stocks

Question: In the past year, the yen depreciated, and Japanese stocks rose. What is the connection?

Solution: Yen depreciation makes Japanese exports cheaper for foreign buyers, boosting export-driven company revenues. Many Japanese companies rely on exports, so the depreciation positively affects their financial performance, driving stock prices up.

Question 4: Japanese Company with Dollar-Denominated Revenue

Question: If a Japanese company's revenue and costs are both denominated in U.S. dollars, how does yen depreciation affect its stock price?

Solution: Yen depreciation is beneficial because when the company reports earnings in yen, the dollar-denominated revenue translates to a higher amount of yen, increasing profits in domestic currency terms and boosting the stock price.