Teaching Assistant: Pritti Agrawal

Email agrawalp@bu.edu

Lecture 2,(02/04/2019)

Download/update the following scripts/files:

1. Updated get stock data script – this will include the computation on the last digit for the Opening Price
2. Pdf file “Error Analysis” – this will include the 4 error metrics
3. Transactions from a bakery script – takes “bakery dataset 1”, adds pricing information and time (days, hours, periods, minutes, etc). The resulting dataset is “bakery dataset with prices and times”

Task 1:

Show that you stock follows (or does not follow standard distribution)

Compute mean and standard deviation and compute

The number of days outside mean +/- st\_dev

Compare this with 5% predicted by normal distribution

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Task 2: analyze the distribution of the last digit (cent position) in the opening price

Compute the statistics on the distribution of last (cent) digit for opening price

A contains ten frequencies (these frequencies add to 1)

A = ( 0.15, 0.05, 0.31, 011, …..)

P = (0.1, 0.1, 0.1, ……, 0.1)

How do you compute Error of your data?

Error Analysis:

Given A = (a1, a2, ……………….., an)

Actual values

And given P = (p1, p2, …., pn),

Compute the error in your model

Predicted Weather:

P1 = (30, 35, 38, 28, 18)

A = (55, 40, 31, 20, 16)

P2 = (16, 27, 31, 2, 10)

How to define Error?

N = 1

Actual = 100

Predicted = 90

Absolute Errror = |A –P| = 10

Relative Error abs( A-P/A)

Possibilities for Error Analysis:

P1 = (30, 35, 38, 28, 18)

A = (55, 40, 31, 20, 16)

P2 = (16, 27, 31, 2, 10)

Compute individual errors

P1 –A = (-26, ….)

P2 – A = (-39, …..)

Compute Average P1-A

Compute average absolute error P1 and A1

P2 and A2 -------------------🡪

A = (100, 120)

P1 = (90, 150), P2 = (95, 105)

|A – P1| = (10, 30) ---------------🡪 20

|A – P2| = (5, 15) -----------------🡪 10

Other: median median error

Average relative error

Rel\_err(A, P1)= (10%, 25%) ------------🡪 17.5

Rel\_err(A, P2) = (5, 12.5) ---------🡪 8.75

Maximum absolute error:

What is the most common error metric that we use?

A: root mean squared error

Sqrt( (a1-p1)\*\*2 + (a2-p2)\*\*2 + …………. + (an-pn)\*\*2/n)

I will put a file on error analysis

Task:

Generate a vector with frequencies for the last digit of the opening price for your stock

A = (a0, …………….., a9)

P = (0.1, ……………., 0.1)

Compute 4 error metrics for the above data:

1. Max absolute error
2. Median absolute error
3. Mean absolute error
4. Root mean squared error

Task 3:

Consider the transactions from a bakery dataset

Let us add one more time description to the dataset:

12am to 6 am ----🡪 night

6 am to 12 pm ----------🡪 morning

12 pm --- 6 pm ------------🡪 afternoon

6 p.m. to 12 am -🡪 evening

We will investigate the following questions:

Q: What is busiest:

1. Time of the day (hour)
2. Day of the week
3. Period of the day (morning, evening …..)

What is the most profitable time:

1. Hour
2. Day of the week
3. Period of the day

What is the most popular item?

Least popular item?

What combination of 2 items are most popular?

Least popular?

Assume that one barrista can handle 60 transactions per day.

How many “barista” professionals do we need to hire

For each day of the week?

---🡪

Divide all items into 3 groups:

1. Drinks
2. Food
3. Unknown

Q: average price of a drink?

Average price of a food item?

Does this coffee shop make most money from drinks or food?

Take the top 5 items by popularity

When (day of the week, hour, day period)

Take the bottom 5?

Estimate the group size (how many drinks per transaction)