Statistical Inference and Learning- Ex-3

Yiqing Wang and Boaz Nadler

due on Sun. Dec. 27 2015

This exercise deals with the analysis of *real* data. The files `GIS.txt, K.txt, INTC.txt` contain the stock prices of three well known companies (K-Kellogs, GIS-General Mills, and INTC - Intel), from Jan-1-2010 till Dec-31-2014. The structure of each row is as follows, with all entries delimited by commas:

```
2015,11,23,34.66,34.849998,34.41,34.48,20097100,34.48
```

The first three entries are the date (year,month,day), the next fields are the Open, High, Low, Close, Volume, Adj Close stock price. The first row in each file contains precisely this information.

We will be mostly interested in the last field in each row: the adjusted closing stock price at each day. Overall we have $n = 1258$ trading days in each of these data files.

Q1 Load the data into your favorite statistical software. Let P_t denote the price of stock P at trading day t. Compute the mean and standard deviation of the daily (more precisely, between consecutive trading days) return rate, defined as $100 \cdot (P_{t+1} - P_t)/P_t$ for each of the three stocks.

Clearly each stock has an empirical different return, and different standard deviation. Assume that for each stock P, the process $(P_{t+1} - P_t)/P_t$ is stationary over time, with some unknown underlying population mean and standard deviation parameters. Is it reasonable to assume that $\mu_{\mathrm{K}} = \mu_{\mathrm{GIS}}$? What about $\sigma_{\mathrm{K}} = \sigma_{\mathrm{GIS}}$? Similarly, it is reasonable to assume that $\sigma_{\mathrm{K}} = \sigma_{\mathrm{INTC}}$? Explain your answer.

Q2 A fundamental question of interest is what is the *distribution* of daily returns of a given stock, and in particular do different stocks have different daily return distributions.

Suppose that the daily return rates of K and GIS have probability distribution functions F and G. Let $\hat{F}_n(x)$ denote the *empirical distribution function* of stock K, computed from samples $x_1, \ldots, x_n$. It is defined as

$$\hat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}(x_i \leq x)$$

(a) Assume that the samples x_i are all i.i.d. from $F(x)$. Prove that any fixed x, $\hat{F}_n(x) \to F(x)$. What is the distribution of $\hat{F}_n(x)$ at a fixed x ?

(b) Compute $\hat{F}_n$ and $\hat{G}_n$ and plot them on the same figure. Find what is the value of $\max_x |\hat{F}_n(x) - \hat{G}_n(x)|$.

(c) We would like to know how informative is this value for assessing or rejecting the hypothesis that $F = G$. To this end recall the Dvoretzky-Kiefer-Wolfowitz inequality. It states that if x_i are all i.i.d. samples from a random variable X with c.d.f. $F(x)$, then

$$\Pr[\sup_x |\hat{F}_n(x) - F(x)| > \epsilon] \leq 2e^{-2n\epsilon^2}$$

Prove that if x_i and y_j are all i.i.d. from the same distribution $F(x)$, with empirical distributions $\hat{F}_n$ and $\hat{G}_n$ respectively, then

$$\Pr[\sup_x |\hat{F}_n(x) - \hat{G}_n(x)| > \epsilon] \leq \Pr[\sup_x |\hat{F}_n(x) - F(x)| > \epsilon/2] + \Pr[\sup_x |\hat{G}_n(x) - F(x)| > \epsilon/2]$$

What does this imply on the hypothesis that $F = G$?

Q3 Compute a non-parametric kernel density estimate for the density of the daily price change rates $100 \cdot (P_{t+1} - P_t)/P_t$ for the stock $INTC$. Choose the kernel of your choice and find the bandwidth h via cross-validation. (a) Plot the unbiased estimate of the MSE as a function of h.

(b) Plot the resulting density and compare it to the parametric Maximum likelihood density corresponding to the assumption that the daily returns follow a Gaussian distribution $N(\mu, \sigma^2)$.

Some Matlab Tips:

Matlab is a common software for scientific calculations. While Matlab is not free, Weizmann has a site licence, so all computers in the department should have it. Octave is a free version mostly compatible with it.

Reading comma separated data: `A = csvread(filename,start-row,start-col)` will read a comma separated text file and put the numbers into the matrix `A`. If the first row has text you can start from some other row by specifying `start-row` variable.

Handling Matrices: `size(A)` will give you the number of rows and columns in `A`. Rows and columns start with the index 1, namely `A(1,1)` the first entry of `A` in its first row. `A(1,:)` returns a row vector with all entries in the first row. Similarly `A(:,3)` returns a column vector with all entries in the third column of `A`. You can guess what is `A(:,end)`. For a vector or matrix `A`, its transpose is `A'`.

Reversing Order: Since the dates in the file are from most recent to oldest, you may wish to *reverse* their order. If `A` is a matrix, then `A(end:-1:1,:)` will reverse the order of its rows.

Finite Difference: If `A` is a vector of length n, then `diff(A)` is a vector of length $n-1$ whose entries are $A_{t+1} - A_t$.

Entry-wise operations: If `A` and `B` are two vectors of same length, then `A.*B` and `A./B` yield vectors of same length, whose entries are $A_j \cdot B_j$ and A_j/B_j respectively.

Mean and Standard Deviations: for a vector `A`, `mean(A)` and `std(A)` give the mean and standard deviation.