Homework 2
MATH 191 Topics in Data Science
Posted: Friday, October 16, 2015
Updated: Wednesday, October 21, 2015
Due in class, Friday , October 30, 2015

Problem 1 (10 points)

(a) Let X be a random variables distributed uniformly $\sim$ Unif($[-1, 1]$) and let $Y = -X$ if $X < 0$, and $Y = X$ if $X > 0$. Are X, Y correlated? Are they independent?

(b) Let α be a random variables distributed uniformly $\alpha \sim$ Unif($[0, 2\pi]$) and let $X = \cos\alpha$ and $Y = \sin\alpha$. Are X, Y correlated? Are they independent?

Problem 2 (10 points)

(a) Simple Linear Regression: Show that the OLS estimator β_0 computed in class is unbiased, i.e., show that $\mathbb{E}[\hat{\beta}_0] = \beta_0$

(b) Multiple Linear Regression: Show that the OLS estimator $\beta \in \mathbb{R}^p$ computed in class is unbiased, i.e., show that $\mathbb{E}[\hat{\beta}] = \beta$

Problem 3 (10 points)
Simple Linear Regression: Compute the Variance of the estimator $\hat{\beta}_0$ obtained in class.

Problem 4 (10 points)
Compute the characteristic function of the following distribution

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}$$

Problem 5 (10 points)
Let $H(X)$ denote the entropy of the random variable X, and $H(X, Y)$ denote the joint entropy of (X, Y). Show that the following holds

$$H(X, Y) = H(X) + H(Y|X)$$

Problem 6 (10 points)
Assuming that (X, Y) are jointly distributed according to Table 1

$Y\backslash X$	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

Table 1: Joint distribution of X and Y.

(a)(2p) Compute the marginal distributions of X and Y
(b)(3p) Compute $H(X)$ and $H(Y)$
(c)(5p) Compute $H(X|Y)$, $H(Y|X)$ and $H(X,Y)$.

Problem 7 (10 points)
Let $I(X;Y)$ denote the mutual information of X and Y. Show that the following hold true
(a)(4p)

$$I(X;Y) = H(X) - H(X|Y)$$

(Note that $I(X;Y) = H(Y) - H(Y|X)$)
(b)(2p)

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

(c))2p)

$$I(X;Y) = I(Y;X)$$

(d)(2p)

$$I(X) = H(X)$$

Problem 8 (10 points)
Programming Assignment 1: Problem 8 from the textbook, Chapter 3, page 121

Problem 9 (50 points)
Programming Assignment 2: [Note: this problem is shorter than it seems... I just added a number of additional clarifications along the way.]

This assignment uses the (Yahoo) financial data set from the previous homework. Consider the following set of $n = 8$ stocks

{’SPY’,’ˆVIX’, ’ˆTNX’, ’OIL’,’GLD’,’GOLD’,’ˆN225’,’ˆFTSE’}

As in the previous homework, let P denote the prices matrix, of size $T \times n$, where T is the number of days in history, and s is the number of instruments. For each time series of a given instrument, we compute the so called log-returns

$$R_i(t) = log(R_i(t)/R_i(t-1))$$

You can do so, for all instruments at once, in R via the command

$$R = \log(P[2:T,]/P[1:(T-1),])$$

as in the previous homework. Denote by n the resulting number of samples (day) in your history ($n = T - 1$). For the current data set there should be $n \approx 1262$ days (after removing the rows that contain NAs).

(a) We will use a linear model to explain the return of a stock on day t as a function of the return of all $s = 8$ stocks on day $t - 1$. In other words, we fit the following model

$$r_{t+1,i} \sim \beta_0 + \sum_{i=1}^{s=8} \beta_i r_{t,i}$$

where $r_{t,i}$ denotes the return of stock i on day t. Note the number of samples in your regression will be $n - 1$, and you will have to separately fit a total of $s = 8$ models, once for each stock. For now, let us fix $i =$ SPY, in other words we are trying to forecast the price of SPY for tomorrow, based on the state of the world available today.

(i) Compute and record the β's from your fit, and compute the estimated $\hat{r}_{t+1,SPY}$ value after the fit, for $t = 1, \ldots, n$

(b) To measure the success of our forecast $\hat{r}_{t+1,SPY}$ made on day t for tomorrow's return $r_{t+1,SPY}$, we will use the popular Sharpe Ratio of the PNL time series, defined as follows. The PNL_t (Profit & Loss) of our forecast for day $t + 1$ is

$$\text{PNL}_{t+1} = \begin{cases} |r_{t+1,SPY}|, \text{ if } (\hat{r}_{t+1,SPY} > 0 \text{ and } r_{t+1,SPY} > 0) \text{ or } (\hat{r}_{t+1,SPY} < 0 \text{ and } r_{t+1,SPY} < 0) \\ -|r_{t+1,SPY}|, \text{ otherwise} \end{cases}$$

Note that, in more compact notation, the PNL can be simply computed as

$$\text{PNL}_{t+1} = \text{sign}(\hat{r}_{t+1,SPY}) \times r_{t+1,SPY} \tag{1}$$

In other words, on any given day, if we forecast that $\hat{r}_{t+1,SPY}$ is positive, we will buy \$1 of SPY, and hold our position for 1 day. If our forecast is indeed correct and $r_{t+1,SPY}$ is positive, our profit will be $r_{t+1,SPY} \times \$1$. (You can think on your own about the other 3 scenarios, and convince yourself that the PNL is indeed give by (1)).

(i) Plot the cumulative PNL of your above investment strategy, across time. In other words, if $\text{PNL}_t, t = 1, \ldots, n$ denotes the time series of daily PNL, the cumulative PNL on day t is given by $\sum_{i=1}^{t} \text{PNL}_i$ (in other words, the profit or loss accumulated thus far).

(ii) Compute the Sharpe Ratio of your time series of daily $\text{PNL}_t, t = 1, \ldots, n$, defined as the following normalized z-score

$$\text{Sharpe} = \frac{\mu_{PNL}}{\sigma_{PNL}} \times \sqrt{252},$$

where μ_{PNL} denotes the average of the daily PNL times series, and σ_{PNL} the standard deviation of the daily PNL time series. The normalization comes from there being 252 trading days per year.

(iii) Compute the average annualized return of your investment strategy as

$$\text{Ret}_{ann} = \mu_{PNL} \times 252$$

(iv) Compute the total return of your investment strategy as

$$\text{Ret}_{total} = \sum_{t=2}^{n} \text{PNL}_t$$

(c) Repeat part (b), for every stock i in your list, and record your results in a table, such as

Stock name	Sharpe	μ_{PNL}	Ret_{ann}	Ret_{total}
1. SPY	0.98	6.2×10^{-4}	0.156	0.779
2. VIX				
$\vdots$				
8. FTSE				

Table 2: Table with in-sample performance statistics

For ease of visualization, it is customary to show the above μ_{PNL} as 6.2×10^{-4} as opposed to 0.00062. The quantity 10^{-4} is often referred to as *1 basis point (denotes as bpts)*, i.e., $1\% = 100$ *bpts*.

(d) Repeat part (b), but this time we will be using a sliding window approach, to avoid using future information. In other words, on any given day t, we will be building our training set from the past m days of data. For simplicity, we fix the parameter $m = 100$.

(i) for each and every day $d = 101, \ldots, n-1$, repeat part (b), where the regression will be performed on a training set with the previous 100 days of data, and their corresponding response given by the return next day. In other words you, for a given day, say $d = 533$, the training data will be given by

$$x = (r_{t,1}, r_{t,2}, \ldots, r_{t,8}) \text{ with response } \quad y = r_{t+1,'SPY'}, \quad t \in \{433, 435, \ldots 532\}$$

After doing the multiple linear regression and obtaining the coefficients β's, apply them to today's information $x = (r_{533,1}, r_{533,2}, \ldots, r_{533,8})$ and compute the estimate for tomorrow return of SPY

$$\hat{r}_{534,SPY} = \beta_0 + \beta_1 r_{533,1} + \beta_2 r_{533,2} + \ \ldots \ + \beta_8 r_{533,8}$$

Compute the various performance statistics defined above: Sharpe, μ_{PNL}, Ret_{ann}, Ret_{total}.

(e) Repeat part (d) for each of the remaining stocks in your list, and gather in a single plot the $s = 8$ cumulative PNL graphs of your investment strategy (each one with a different color; you may use the provided function for this). Gather your results in a table of performance statistics, similar to Table 2. What can you conclude from looking at the results in Table 2 and Table 3?

Stock name	Sharpe	μ_{PNL}	Ret$_{ann}$	Ret$_{total}$
1. SPY				
2. VIX				
⋮				
8. FTSE				

Table 3: Table with out-of-sample performance statistics

Additional clarifications:

Note 1: I provided for you the function compute_linear_regression(), which takes as INPUT:
- Training X matrix: X_{train} of size n by p
- Training y vector: y_{train} of size n by 1
- Test X matrix: X_{test} of size k by p

and performs multiple linear regression (without an intercept) on the training data, thus fitting

$$y_{train} \sim X_{train},$$

and applies the obtained coefficients $\beta_1, \ldots, \beta_p$ to a test matrix X_{test}. The output is the forecast $\hat{y}_{test}$ based on X_{test}. This function will come in hand many times throughout the course, and especially in your projects.

Note 2: I provided for you the function plot_cumsum(), which takes as input a matrix X of size $T \times n$ and plots the time series of the n different columns, on the same plot. Note that you can use this function to plot the cumulative PNL from part (b)(i), if the argument is X - a matrix where entry $X_{d,i}$ is the cumulative PNL on stock i on day d.

Note 3: I provided for you the function compute_Sharpe_Ratio(), which takes as input the time series of PNL, and computes the Sharpe ratio.

Note 4: The function main_linear_regression() should get you started.

Note 5: You should hand in whatever additional code you write, in addition to Table 2, Table 3, and the cumulative PNL plots of your investment strategy, one each for in-sample (Part (b)(i)) and out-of-sample (Part (e)).