Statistical Inference and Learning- Ex-3

Yaniv Tenzer and Boaz Nadler

Due on Tuesday, June 25, 2019,
email submissions: `stat.wisdom@gmail.com`

Q1 In this problem we will look at the daily returns of 31 stocks (companies) over the period 2010-2014. The file `A.txt` contains 1258 lines each with 31 columns. Each line contains the (closing) price of 31 stocks on a specific day. The list of stock tickers is available in the file `stock_list.txt`. For each stock we are interested in its daily return in percentage points, defined as

$$R_i(t) = 100 \times (P_i(t) - P_i(t-1))/P_i(t-1).$$

where $P_i(t)$ is the closing price of stock i on trading day t.

This gives a matrix of size 1257x31 that appears in the file `B.txt`. The corresponding trading days [year month day] appear in the file `dates.txt`

In matlab, for example, the operation to compute `B` from `A` is simply

```
B = 100 * diff(A)./A(1:end-1,:).
```

In future exercises we will also look at individual stocks, but for now, we will only consider the daily return averaged over these 31 stocks.

1. Plot the mean of daily returns (namely a vector of length 1257). What is the average and standard deviation of this random variable ?
2. A crucial part of data analysis is to detect outliers / abnormal points in the data. Find the date with the lowest return - which date was it? How many standard deviations was this return far from the mean daily return? Would you consider this as an outlier/abnormal observation, explain your answer! If interested at what happened during the few days around that date, take a look at
`en.wikipedia.org/wiki/August_2011_stock_markets_fall`
3. Compute a non-parametric density estimate of the average daily returns. Choose the kernel of your choice and find the bandwidth h via cross validation. Provide details on how precisely the bandwidth h was found.

   Compare the estimated density to a fit assuming the random variable in question was distributed as a Gaussian $N(\mu, \sigma^2)$, with their parameters estimated by Maximum Likelihood. Namely, plot $(x, \hat{p}_{KDE}(x))$ and $(x, \hat{p}_N(x))$ on the same graph. In your opinion, is a Gaussian distribution a good fit to the daily returns? Explain your answer.

Q2 Let $\{(\mathbf{x}_i, Y_i)\}_{i=1}^n$ be n observations from the following regression model:

$$Y_i = \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + \epsilon_i, \; 1 \le i \le n,$$

where $\epsilon \sim N(\mathbf{0}, \sigma^2 I)$. Let $\hat{\beta}$ be the OLS estimator, and for each i, let $\hat{Y}_i$ be the corresponding predicted value $\mathbf{x}_i^T \hat{\beta}$. Show that $\hat{\sigma}^2 \equiv \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-k-1}$ is an unbiased estimator of σ^2.

Q3 [OLS solution invariant to scaling] Let $D \equiv Diag(\lambda_1, \ldots, \lambda_p)$ denote a diagonal matrix with λ_i on its i'th entry (i.e., $D_{ii} = \lambda_i$). Denote by $\hat{\beta}_X$ the OLS estimator based on the n observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, with $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,k})$, $1 \le i \le n$. Similarly, denote by $\hat{\beta}_Z$, the OLS estimator based on $\{(y_i, \mathbf{z}_i)\}_{i=1}^n$, where $z_i \equiv Dx_i$. For a new observation $\mathbf{x}^*$, show that $\hat{\beta}_X^T \mathbf{x}^* = \hat{\beta}_Z^T \mathbf{z}^*$, where $\mathbf{z}^* \equiv D\mathbf{x}^*$.

Q4 [OLS solution is a maximum likelihood estimator] Similar, but slightly different from the linear settings of Q1, assume that $\epsilon \sim N(\mathbf{0}, Diag(\sigma_1^2, \ldots, \sigma_n^2))$. Namely, each observed Y_i has a different and *known* noise level σ_i.

1. Show that ML estimator of β is given by the solution of the following optimization problem:
$$\text{argmin}_{\beta \in \mathbb{R}^k} \sum_{i=1}^n (y_i - \beta^T \mathbf{x}_i)^2 / \sigma_i^2$$
2. Solve for β. Show that $\hat{\beta} = (\mathbf{X}^T D^{-1} \mathbf{X})^{-1} \mathbf{X}^T D^{-1} \mathbf{y}$, where $D = Diag(\sigma_1^{-2}, \ldots, \sigma_n^{-2})$ and $\mathbf{X}$ is the $n \times k$ design matrix whose rows are the observations $\mathbf{x}_i^T$.
3. More generally, let $\mathbf{w} \equiv (w_1, \ldots, w_n)$ be a weight vector (i.e., $w_i > 0$ for all i). Write a formula for the solution for the following optimization problem
$$\text{argmin}_{\beta \in \mathbb{R}^\mathbf{k}} \sum_{i=1}^n w_i (y_i - \beta^T x_i)^2.$$