# QBUS1040: Foundations of Business Analytics
# Homework 5

Semester 2, 2019

This homework consists of **five** problems that require you to submit a written response and a coding component. The problems that require a written response are described in this paper. You need to print it and write your answers directly in this paper. For all problems where you are asked for a free-form answer, it must be written in the box below the problem. We won't read anything outside the boxes. You should use scratch paper (which you will not turn in) to do your rough work. You should submit a scanned copy of your written solution as a PDF via Canvas.

Please do not submit photos of your written solution as it is difficult for the marking system to recognise your submission. Also, please do not use a tablet to write your homework as it is very likely that your submission will not be processed correctly through the marking system. Please use a conventional scanner. You should double check your PDF before you submit the file. Your PDF submission should contain **all pages** of this document. The file size of your PDF document should not exceed 128MB, or it will not be accepted by the submission system.

The coding components are described in a separate Jupyter Notebook file. You should also download the Jupyter Notebook file for Homework 5 and enter your code in the space provided. You should submit your code as a Jupyter notebook file via Canvas.

This homework is due by 4pm on Friday, the 15th of November. **Late homework will not be accepted. Violation of the above submission instructions may incur a 30% penalty.**

**Tutorial time:** Friday 4 - 6pm

**Tutor's name:** Mr Kam Fung (Henry) Cheung

**Your SID:** 4 8 0 0 4 8 6 9 1

(For QBUS1040 staff only)

| Question: | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|
| Points: | 20 | 20 | 20 | 20 | 20 | 100 |
| Score: | | | | | | |

1. *Multi-class classifier via matrix least squares.* Consider the least squares multi-class classifier described in §14.3, with a regression model $\tilde{f}_k(x) = x^T\beta_k$ for the one-versus-others classifiers. (We assume that the offset term is included using a constant feature.) Show that the coefficient vectors $\beta_1, \dots, \beta_K$ can be found by solving the matrix least squares problem of minimizing $\|X^T\beta - Y\|^2$, where $\beta$ is the $n \times K$ matrix with columns $\beta_1, \dots, \beta_K$, and $Y$ is an $N \times K$ matrix.

(a) (10 points) Give $Y$, i.e., describe its entries. What is the $i$th row of $Y$?

The $N \times K$ matrix Y is the true classification matrix of all the data points given. Its entries are of Boolean type. For each row of Y, it's the real-valued classification of a data point. Say Y is a $10 \times 3$ matrix, ~~the~~ this lets us know that there are 10 data points of 3 different classes. The first row will be the classification of the first object. If that object is of class 1, $Y_{11}$ will be $+1$ and $Y_{12}$ and $Y_{13}$ will be $-1$.

(b) (10 points) Assuming the rows of $X$ (i.e., the data feature vectors) are linearly independent, show that the least squares estimate is given by $\hat{\beta} = (X^T)^\dagger Y$.

. Given that the rows of X are linearly independent

$\Rightarrow$ there exists a pseudo inverse for $X^T \Rightarrow (X^T)^+$

. For each individual class $i \in \{1, \dots, K\}$, we can estimate $\hat{\beta}_i$ by solving the least squares problem of minimising $\| X^T\hat{\beta}_i - y_i \|^2$ with $\beta_i$ the $i^{th}$ column of $\beta$ and $y_i$ the $i^{th}$ column of Y $\Rightarrow$ $\hat{\beta}_i = (X^T)^+ y_i$

. We have
$$\hat{\beta} = [\hat{\beta}_1 \quad \hat{\beta}_2 \quad \dots \quad \hat{\beta}_K]$$
$$= [(X^T)^+ y_1 \ (X^T)^+ y_2 \ \dots \ (X^T)^+ y_k]$$
$$= (X^T)^+ [y_1 \ y_2 \ \dots \ y_k] = (X^T)^+ Y$$

2. (20 points) Consider the regularized data fitting problem (15.7) of the textbook. Recall that the elements in the first column of $A$ are one. Let $\hat{\theta}$ be the solution of (15.7), i.e., the minimizer of

$$\|A\theta - y\|^2 + \lambda(\theta_2^2 + \cdots + \theta_p^2),$$

and let $\tilde{\theta}$ be the minimizer of

$$\|A\theta - y\|^2 + \lambda\|\theta\|^2 = \|A\theta - y\|^2 + \lambda(\theta_1^2 + \theta_2^2 + \cdots + \theta_p^2),$$

in which we also penalize $\theta_1$. Suppose columns 2 through $p$ of $A$ have mean zero (for example, because features $2, \ldots, p$ have been standardized on the data set; see page 269 of the textbook for details).

Show that $\hat{\theta}_k = \tilde{\theta}_k$ for $k = 2, \ldots, p$.

. We have $\| A\theta - y\|^2 + \lambda(\theta_2^2 + \dots + \theta_p^2)$ can be rewritten as

$$J = \| A\theta - y\|^2 + \lambda \| B.\theta \|^2$$

with $B$ the $(p-1) \times p$ selector matrix $[0 \; e_1 \; e_2 \; \cdots \; e_p]$

we can turn this multi-objective problem into single-objective

$$\begin{bmatrix} \| A\theta - y \|^2 \\ \lambda \| B.\theta \|^2 \end{bmatrix} \Leftrightarrow \left\| \begin{bmatrix} A\theta - y \\ \sqrt{\lambda} B\theta \end{bmatrix} \right\|^2 \Rightarrow \| \tilde{A}\theta - \tilde{b} \|$$

with $\tilde{A} = \begin{bmatrix} A \\ \sqrt{\lambda} B \end{bmatrix}$ and $\tilde{b} = \begin{bmatrix} y \\ 0 \end{bmatrix}$

We can find ~~θ~~ the minimiser of this newly created objective

by $\hat{\theta} = (\tilde{A}^T \tilde{A})^{-1} \tilde{A}^T \tilde{b}$

$= (1.\, A^T A + \lambda B^T B)^{-1} (A^T y)$

. The same method applies when we also penalise $\theta_1$

$$\Rightarrow J = \| A\theta - y \|^2 + \lambda \|\theta\|^2$$

$$\begin{bmatrix} \| A\theta - y \|^2 \\ \lambda . I \end{bmatrix} \Leftrightarrow \left\| \begin{bmatrix} A\theta - y \\ \sqrt{\lambda} . I \end{bmatrix} \right\|^2 \Rightarrow \| \tilde{A}\theta - \tilde{b} \|$$

with $\tilde{A} = \begin{bmatrix} A \\ \sqrt{\lambda} . I \end{bmatrix}$ and $\tilde{b} = \begin{bmatrix} y \\ 0 \end{bmatrix}$

The minimiser for $J$ is $\tilde{\theta} = (\tilde{A}^T \tilde{A})^{-1} \tilde{A}^T \tilde{b}$

$= (A^T A + \lambda . I)^{-1} (A^T y)$

3. (20 points) Please see Jupyter Notebook file `Homework5.ipynb` for details.

4. *Least squares classification with regularization.* The file `lsq_classifier_data.ipynb` contains feature $n$-vectors $x_1, \ldots, x_N$, and the associated binary labels, $y_1, \ldots, y_N$, each of which is either $+1$ or $-1$. The feature vectors are stored as an $n \times N$ matrix $X$ with columns $x_1, \ldots, x_N$, and the labels are stored as an $N$-vector $y$. We will evaluate the error rate on the (training) data $X$, $y$ and (to check if the model generalizes) a test set $X_{\text{test}}$, $y_{\text{test}}$, also given in `lsq_classifier_data.ipynb`.

   (a) (10 points) *Least squares classifier.* Find $\beta, v$ that minimize $\sum_{i=1} N(x_i^T\beta + v - y_i)^2$ on the training set. Our predictions are then $\hat{f}(x) = \mathbf{sign}(x\beta + v)$. Report the classification error on the training and test sets, the fraction of examples where $\hat{f}(x_i) \neq y_i$. There is no need to report the $\beta, v$ values.

Rewrite the minimising problem as $\| X^T\beta + v.\mathbb{1} - y\|^2$

with $A = [\mathbb{1} \ \ X^T]$, I once again rewrite my problem as

$\|A\theta - y\|^2$ with $\beta = \theta_{2:p}$ and $v = \theta_1$

Solving $\theta$ using QR factorisation and back sub $\Rightarrow$ get $\beta$ and $v$

Find $\tilde{y} = x.\beta + v$, using the found $\beta$ and $v$ and training set and testing set

$\Rightarrow \hat{y} = \text{sign}(\tilde{y})$. Compare this value against y-train and y-test gives the error rate (training) = 0.2067 and (testing) = 0.21

   (b) (10 points) *Regularized least squares classifier.* Now we add regularization to improve the generalization ability of the classifier. Find $\beta, v$ that minimize

$$\sum_{i=1}^{N}(x_i^T\beta + v - y_i)^2 + \lambda\|\beta\|^2,$$

where $\lambda > 0$ is the regularization parameter, for a range of values of $\lambda$. Please use the following values for $\lambda$: $10^{-1}, 10^0, 10^1, 10^2, 10^3$. Suggest a reasonable choice of $\lambda$ and report the corresponding classification error on the training and test sets. Again, there is no need to report the $\beta, v$ values. *Hint:* plot the training and test set errors against $\log_{10}(\lambda)$.

Rewrite the minimise problem $\|X^T\beta + v.\mathbb{1} - y\|^2 + \lambda.\|\beta\|^2$

With $A = [\mathbb{1} \ \ X^T]$ and B the selector matrix from question 2

I once again rewrite my problem as

$\|A\theta - y\|^2 + \lambda.\|B\theta\|^2$ with $\beta = \theta_{2:p}$ and $v = \theta_1$

Find $\tilde{y}$ and $\hat{y}$ using the same method above for different values of $\lambda$. A reasonable choice for $\lambda$ would be $10^0$ since it's the largest $\lambda$ that minimises The error rate ~~for~~ of testing data set (0.21), while for the training set it's 0.2067.

5. *Estimating the elasticity matrix.* In this problem you create a standard model of how demand varies with the prices of a set of products, based on some observed data. There are $n$ different products, with (positive) prices given by the $n$-vector $p$. The prices are held constant over some period, say, a day. The (positive) demands for the products over the day is given by the $n$-vector $d$. The demand in any particular day varies, but it is thought to be (approximately) a function of the prices. The units of the prices and demands don?t really matter in this problem. Demand could be measured in 10000 units, and prices in \$100.

   The nominal prices are given by the $n$-vector $p^{\text{nom}}$. You can think of these as the prices that have been charged in the past for the products. The nominal demand is the $n$-vector $d^{\text{nom}}$. This is the average value of the demand, when the prices are set to $p^{\text{nom}}$. (The actual daily demand fluctuates around the value $d^{\text{nom}}$.) You know both $p^{\text{nom}}$ and $d^{\text{nom}}$. We will describe the prices by their (fractional) variations from the nominal values, and the same for demands. We define $\delta^p$ and $\delta^d$ as the (vectors of) relative price change and demand change:

$$\delta_i^p = \frac{p_i - p_i^{\text{nom}}}{p_i^{\text{nom}}}, \quad \delta_i^d = \frac{d_i - d_i^{\text{nom}}}{d_i^{\text{nom}}}, \quad i = 1, \ldots, n.$$

   So $\delta_3^p = +0.05$ means that the price for product 3 has been increased by 5% over its nominal value, and $\delta_5^d = -0.04$ means that the demand for product 5 in some day is 4% below its nominal value. Your task is to build a model of the demand as a function of the price, of the form

$$\delta^d \approx E\delta^p,$$

   where $E$ is the $n \times n$ elasticity matrix.

   You don't know $E$, but you do have the results of some experiments in which the prices were changed a bit from their nominal values for one day, and the day's demands were recorded. This data has the form

$$(p_1, d_1), \ldots, (p_N, d_N),$$

   where $p_i$ is the price for day $i$, and $d_i$ is the observed demand.

   (a) (10 points) Explain how you would estimate $E$, given this price-demand data. Be sure to explain how you will test for, and (if needed) avoid over-fit.
   *Hint.* You might find it easier to separately fit the models $\delta_i^d \approx \tilde{e}_i^T \delta^p$, where $\tilde{e}_i$ is the $i$th row of $E$. (We use the tilde above $e_i$ to avoid conflict with the notation for unit vectors.)

Since $\delta^P$ is a $5 \times 75$ matrix, its rows are linearly independent, so we can use its right inverse in conjunction with $\delta^d$ to estimate E.

We have $\delta^d \approx E . \delta^P$

$$E \approx \delta^d . (\delta^P)^{-1}$$

$$\approx \delta^d . (\delta^P)^T . [\delta^P . (\delta^P)^T]^{-1}$$

To test for over-fit, we can divide the data set into two parts : $\delta_A^P$ & $\delta_A^d$ and $\delta_B^P$ & $\delta_B^d$. Using our method above and set A to estimate E. Then use E and $\delta_B^P$ to estimate $\tilde{\delta}_B^d$. The difference between $\delta_B^d$ and $\tilde{\delta}_B^d$ will let us know if our model generalises.

(b) (10 points) Carry out your method using the price and demand data in the matrices `Prices` and `Demands`, found in `Homework5.ipynb`. Give your estimate $\hat{E}$ in the same file.

Here are some facts about elasticity matrices that might help you check that your estimates make sense (but you don't need to incorporate this information into your estimation method). The diagonal entries of $E$ are always negative, and typically on the order of one. (This means that when you raise the price of one product only, demand for it goes down by a similar fractional amount as the price increase.) The off-diagonal entries can have either sign, and are typically (but not always) smaller than one in magnitude.