SFU faculty of science department of mathematics

MACM 316- D100 COMPUTING ASSIGNMENT 5

Crowdmark Due Date: Monday 11 March, 2019 at 11:55pm

- Please follow the *Guidelines for Computing Assignments* found on the Canvas FAQ.
- Submit a one-page report via Crowdmark, grading scheme is also found on the FAQ.
- Acknowledge any and all collaborations and assistance from classmates/TAs/instructor.

Deductions from Data

QR Warm-up: Before beginning the data analysis part of this assignment, you are asked to benchmark Matlab's *backslash* version of the least-squares solve. From lecture, we know that least-squares solutions to $[A]\,\vec{c} = \vec{y}$ also satisfy linear solves of either:

- the normal equation, $[A^T A]\,\vec{c} = \vec{y}$,
- the QR normal equation, $[R]\,\vec{c} = [Q^T]\,\vec{y}$.

As part of this week's download, the Matlab script *lsq_red.m* loads a datafile *red.csv* with $M = 1499$ rows of data for red wines. The k^{th} row of the array corresponds to a different wine product where y_k is a *quality variable*, in addition to 10 other columns of numerical attributes labelled $(x_1)_k, \ldots (x_{10})_k$ (names of these columns are identified by the string variables in the script). (You do not have to know any wine chemistry for this assignment, except that a higher quality number is desirable.)

The script then constructs the matrix $[A]$ as

$$[A] = \begin{bmatrix} \vdots & \vdots & & \vdots \\ 1 & \vec{x}_1 & \ldots & \vec{x}_{10} \\ \vdots & \vdots & & \vdots \end{bmatrix},$$

and then computes the backslash solution $\vec{c}$ to the multi-variable least-squares fit

$$c_0 + \sum_{j=1}^{10} c_j\,\vec{x}_j = \vec{y}$$

so that $Y(\vec{a}_k) = \vec{a}_k^T \vec{c}$ is the best-fit linear estimate for the quality y_k based on the attribute vector $\vec{a}_k$.

Begin by modifying this script to recompute the best-fit coefficients by three other numerical algorithms:

- $\vec{c}_{NE}$, by directly solving the normal equations;
- $\vec{c}_{QR}$, using the *qr(A,0)* function call;
- $\vec{c}_{QRp}$, using the permuted *qr(A,0)* function call ($AP = QR$).

Recall that the inverse of a permutation matrix is simply its transpose. (Note: if you call *[Q,R,E] = qr(A,0)* in Matlab, E is a "permutation vector". To generate P in your script, you will need to uncomment the line *I=eye(11,11); P = I(:,E)* in *lsq_red.m*). For the benchmark testing, present the values of the least-square residuals (RMS) for all four calculations, as well as the magnitudes of the coefficient differences $|\vec{c} - \vec{c}_{test}|$, where $\vec{c}$ is the backslash solution.

Based on these results, discuss what Matlab's backslash might involve.

Quality prediction challenge: For this week, you will act as a data consultant for a restaurant chain who have tasked you to produce a low-cost predictor for its purchasing of white wines. As the testing of wines has an associated cost, you have been asked to develop a **3-factor** predictor to help them identify wines of potentially high quality. You have available to you a data array in the file *white.csv* that has 4748 evaluated white wines. This is the so-called training data — from which you choose three columns of data $j = \{j_1, j_2, j_3\}$ from among the 10 columns of wine attributes (same attributes as the red data) to develop your predictor of quality $Y(a_{j_1}, a_{j_2}, a_{j_3})$. Your expertise is wanted to determine which three factors, out of the ten, give a relatively small RMS residual from the evaluated quality values. For this part of the assignment, you may use *backslash* for all least squares solves.

Finally, from the Canvas page, you will download one of 3 data files corresponding to a catalogue of white wines that are the ones available for purchase – student numbers ending in 0-3 should download *whitelist1.csv*, numbers ending in 4-6 should download *whitelist2.csv*, and numbers ending in 7-9 should download *whitelist3.csv*. **Please include in your report which list you are using.** These files contain all of the attribute columns, but the quality values are NOT included — the **11th column is now the bottle number**. Use your 3-factor predictor to give your recommendations (by bottle number) for the top five white wines available from your catalogue. Make sure you also include the predicted ratings for these wines.

Note that you will not be graded on finding optimal factors, but your 1-page report to the restaurant chain should convince of the logic behind your selection. (Disclaimer: although this is real wine data, this exercise does not imply there is science to support that a 3-factor analysis applies to the selection of quality wines.)

Recall from lecture that you should aim to communicate critical ideas with clarity. To this end, please underline the **three** sentences in your report that you feel convey the biggest impact to the reader.