

Homework 5

(Total 200 pts)

Due 11:59 pm on December 2, 2022 (Friday)

Canvas submission of your answers (with required plots and source codes attached) in a single

PDF file ('hw5.pdf'), and then submit the following files:

Q1.m, winequality-matlab.csv, Q2.m, Q3.py, Q4.m, and Q5.py

1. (60 pts) Linear Regression on the white wine quality using a reduced dataset:

<https://archive.ics.uci.edu/ml/datasets/wine+quality>

We seek to use the first two attributes (x_1 : fixed acidity, and x_2 : volatile acidity) of four instances (with indices 11, 14, 252, and 775) selected from the following dataset to predict the wine quality (with values in the rightmost column).

<https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv>

x_1 (Fixed Acidity)	x_2 (Volatile Acidity)	y (Quality)
8.1	0.27	5
6.6	0.16	7
8.5	0.26	3
9.1	0.27	9

We attempt to fit these four instances of (x_1, x_2) to the corresponding quality using linear

regression. That is, we seek to determine the content of a 3×1 matrix $p = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}$, so that we

can best (in the least-square sense) “predict” the quality by:

Quality = $p_1 \times$ Fixed Acidity + $p_2 \times$ Volatile Acidity + p_3 .

To this end, we solve the equation $Vp = y$, such that the solution vector \hat{p} minimizes $\|V\hat{p} - y\|^2$.

- (a) Show in the table below the content of the matrix V and vector y in the equation: $Vp = y$.

V	p	y (4×1) vector
	$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}$	

- (b) Write Matlab code to solve the equation using four different methods given in the table below and fill in the solutions you obtain for these methods:

Pseudo-Inverse	Economy-Size QR Decomposition	Economy-Size Singular Value Decomposition	<i>fitlm</i> function in Matlab
$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} =$	$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} =$	$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} =$	$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} =$

- (c) Compare the solutions using these methods and comment on your results.
 (d) What is the value of $\|V\hat{p} - y\|^2$, where \hat{p} is the solution using the *fitlm* function in Matlab. How is the value related to the “Root Mean Squared Error” of the regression model obtained by using *fitlm*? Show the math. Note: Do your own research to find out.
 (e) Attach and upload your code ‘Q1.m’. Also attach a screenshot of running the code.

2. (30 pts) Linear Regression using the *fitlm* function in Matlab. This time, we will predict the white wine quality using all the instances in the dataset.

(a) In Matlab, read in the dataset file in Question 1.

```
>> data = readmatrix('winequality-white.csv');
```

(b) Use the *writematrix* function to export the variable “data” to a file named “winequality-matlab.csv”, which will be used in Question 3.

Upload the file “winequality-matlab.csv” to Canvas.

(c) Use the first two attributes (fixed acidity, and volatile acidity) of **all instances** to best fit the wine quality: $Quality = p_1 \times \text{Fixed Acidity} + p_2 \times \text{Volatile Acidity} + p_3$. What is the solution to the least-square problem? What is the “Root Mean Squared Error” of the best fit regression model?

Fill in the table with your answers.

Solution to least-square fit (\hat{p})	Root Mean Squared Error
$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} =$	

(d) Now we use all eleven attributes (fixed acidity, volatile acidity, ..., alcohol) other than the quality attribute of **all instances** to best fit the quality attribute. Regarding the best fit regression model obtained, what is “Error degrees of freedom”? What is the “Root Mean Squared Error”? Fill in the table below with your answers.

Error degrees of freedom	Root Mean Squared Error

(e) Compare the Root Mean Squared Errors in (c) and (d) and comment on your results.

(f) Attach and upload your script ‘Q2.m’.

3. (30 pts) Linear Regression in sklearn.

(a) Load the file “winequality-matlab.csv” generated in Question 2.

(b) Use the first two attributes (fixed acidity, and volatile acidity) of **all instances** to best fit the wine quality:

$$Quality = p_1 \times \text{Fixed Acidity} + p_2 \times \text{Volatile Acidity} + p_3.$$

(c) What is the solution to the least-square problem?

Fill in the table with your answer.

Solution to least-square fit (\hat{p})
$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} =$

(d) Compare the above solution to that in Q2(c). Comment on your results.

(e) Attach a screenshot of running the code. Attach and upload the code “Q3.py”.

4. (40 pts) Classification using Logistic Regression on the Wine Quality dataset.
 - (a) Read into Matlab the following dataset “winequality-white.csv”.
 - (b) Create a new column named “Class”, by converting the values of the “quality” into a class label based on the following rule:
 - Class = 1, if quality < 7;
 - Class = 2, if quality >= 7.

These “Class” values will serve as labels used to train a classifier in the next step so that we can classify the wine into either Class 1 (Poor) or Class 2 (Good) based on all the attributes except for the quality.

- (c) Use all eleven attributes (fixed acidity, volatile acidity, ..., alcohol) other than the quality attribute of **all instances** to train a binary classifier using the *mnrfit* function.
- (d) Show and attach the Confusion Matrix Chart of the trained classifier.
- (e) What is the classification accuracy of the trained classifier? What are the precisions of the Good and Poor classes, respectively. Fill in the table with your answers:

Accuracy (%)	Precision (%) for Good Wine	Precision (%) for Poor Wine

- (f) Attach and upload the code “Q4.m”.

5. (40 pts) Linear Perceptron Classifier in sklearn.
 - (a) Train a classifier with a single perceptron unit using the following dataset: http://www.ece.uah.edu/~dwpan/course/ee610/hw/dataset_hw.csv
Use the *Perceptron* function without specifying any parameter.
 - (b) How many iterations were actually used when the training stopped?
 - (c) Fill in the table below with your answers regarding the trained classifier:

Weighs Assigned to the Features	Constant in Decision Function

- (d) Show and attach the Confusion Matrix Chart.
- (e) Print and attach the Classification Report with the “digits” parameter set to 4.
- (f) Compare the overall classification accuracy, as well as the precisions for the two classes with the trained classifier based on Discriminant Analysis in Question 5 of HW4. Fill in the table below with your answers. Comments on your results.

	Overall Classification Accuracy (%)	Precision (%) for Class 1	Precision (%) for Class 2
Perceptron Classifier			
Discriminant Analysis Classifier			

- (g) Attach and upload your code ‘Q5.py’.