Mohammad Emtiyaz Khan, Timur Bagautdinov, Carlos Becker, Ilija Bogunovic & Ksenia Konyushkova

1. Basic Exploratory Data Analysis using Matlab

1.1 The goal

The goal of this exercise is to,

- Learn about basic matrix operations.
- Learn matrix manipulations in Matlab.
- Learn plotting in Matlab.
- Learn basic exploratory data analysis.
- Get familiar with a regression dataset.

1.2 Installing Matlab

If you do not have Matlab, please install and download it from http://poseidon.epfl.ch/english/software/article/professio softwares.htm if you are a master student, or http://distrilog.epfl.ch for PhD students.

1.3 Matlab Basics

If you are not familiar with Matlab, we recommend you to check out the guide by Matt Dunham and Kevin Murphy at https://ubcmatlabguide.github.io/.

For this exercise session we recommend to read the following sections:

- Getting Started
- Matrix Operations: only up to *More Linear Algebra*
- Plotting: scatter, subplots, and histograms

1.4 Basic Vector and Matrix Algebra

After finishing the tutorial, you are ready to practice. Do the following exercises.

Exercise 1.1 Practice creating matrices.

1. Create an Identity matrix of size 3×3.
2. Create a diagonal matrix of size 3×3 with diagonal values 1, 2, and 3.
3. Create a matrix containing all zeros of size 3×3.
4. Create a matrix $\mathbf{M}$ of size 4×3 containing random values.
5. Reshape the matrix $\mathbf{M}$ to have dimensionality 6×2 and assign the results to a new matrix $\mathbf{T}$. How does MATLAB re-arrange the elements when a matrix is reshaped?
6. Create a vector $\mathbf{a}$ of size 3×1 containing values 1, 2, and 3.

Hint: Use the functions: `eye()`, `diag()`, `zeros()`, `rand()`, `reshape()` ■

Exercise 1.2 Practice matrix-vector multiplication.

1. Compute the new vector $\mathbf{b} = \mathbf{Ma}$, i.e. by multiplying $\mathbf{M}$ to $\mathbf{a}$.
2. What is the size of $\mathbf{b}$? Is this correct and why?
3. Compute the inner product $e = \mathbf{b}^T\mathbf{b}$. What is the size of e? What are the values? Check with others if they get same values.

Hint: Use the function `length()` to check the length of a vector. ■

Exercise 1.3 Practice computing the inverse of a matrix.

1. Create a matrix $\mathbf{A}$ of size 3×3 containing random values.
2. Compute the new vector $\mathbf{c} = \mathbf{A}\mathbf{a}$.
3. Compute $\mathbf{d} = \mathbf{A}^{-1}\mathbf{c}$, i.e. multiplying *the inverse* of $\mathbf{A}$ by the vector $\mathbf{c}$.
4. What is the size of $\mathbf{d}$?
5. The value of $\mathbf{d}$ should be same as $\mathbf{a}$. Why? Discuss with others if you do not understand.

Hint: Compute inverse using the function `inv()`. ■

Exercise 1.4 Practice selecting sub-blocks of a matrix. Compute the inverse of the 3×3 sub-matrix from the top-left side of $\mathbf{M}$. If you are not sure how to access a sub-matrix, check the Matlab tutorial, section *Basic Indexing* from Matrix Operations. ■

Exercise 1.5 Let's explore floating point numbers:

1. Set `A = [1,2; 3,4]`
2. Define `k = 3.1`
3. Add and subtract it to $\mathbf{A}$: `B = A + k - k`
4. Check the difference: `C = A - B`

Are `A` and `B` equal?
How would you check for approximate equality to take into account such nuances? ■

Exercise 1.6 Now, let's learn about some matrix identities using Matlab. First, create a constant $c = 5$. Then create two random matrices $\mathbf{A}$ and $\mathbf{B}$ of size 3×3. Let $\mathbf{C} = \mathbf{AB}$, i.e. the multiplication of $\mathbf{A}$ and $\mathbf{B}$. Check the following identities:

1. $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
2. $(c\mathbf{A})^{-1} = c^{-1}\mathbf{A}^{-1}$
3. $(\mathbf{A}^{T})^{-1} = (\mathbf{A}^{-1})^{T}$
4. $\mathbf{B} = \mathbf{A}^{-1}\mathbf{A}\mathbf{B} = \mathbf{A}^{-1}\mathbf{C}$
5. $\mathbf{C} = \mathbf{A}\mathbf{B} = (\mathbf{C}\mathbf{B}^{-1})(\mathbf{A}^{-1}\mathbf{C}) = \mathbf{C}\mathbf{B}^{-1}\mathbf{A}^{-1}\mathbf{C}$
6. $\mathbf{C}\mathbf{B}^{-1}\mathbf{A}^{-1} = \mathbf{I}$
7. $\mathbf{C}^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} = (\mathbf{A}\mathbf{B})^{-1}$

■

1.5 Generating, Manipulating and Plotting Data

Here we will do some basic exercises with data and plotting.

Exercise 1.7 Practice plotting a simple function and manipulating the plot.

1. Plot the function $y(x) = 100\,(\cos(x))^2 + x^3$ for the interval $-10 \leq x \leq 10$.
2. Changing the color of the line to red.
3. Plot the function using the `plot()` function, but with the `'.'` option, i.e. use the command `plot(x,y,'.')` where `y` contains the value of the function. What do you see? This style of plotting is called *scatter plot*.

Hint: You can use functions `linspace()` and `plot()`. Remember that element-wise operations carry a dot before the operand. For example, element-wise multiplication between `a` and `b` is done with `a .* b`. ■

Exercise 1.8 Plot a histogram. Generate 10,000 random numbers, sampled from a uniform distribution between 0 and 1, and then plot their histogram.

How does it look like? Does it makes sense? How does the number of bins affect the histogram?

Hint: You can use the functions `rand()` and `hist()` to generate random numbers and plot the histogram, respectively. ■

Exercise 1.9 Now a pretty histogram. Repeat the exercise above, but now sampling from a normal distribution with mean $\mu = 10$ and standard deviation $\sigma = 2$. Compute the empirical mean and standard deviation for this data sample. Check Wikipedia about Normal Distribution if you do not understand this exercise.

Hint: You can use the function `randn()` to generate random numbers from a normal distribution of mean zero and standard deviation 1.

Hint: `mean()` and `std()` compute the empirical mean and standard deviation of a set of observations. ■

1.6 Exploratory Data Analysis (EDA) with Matlab

We will now do some very basic Exploratory Data Analysis (EDA). Such analysis is always required before you start applying machine learning algorithms to the data. In this exercise, we will look at some basic statistics such as averages.

We use a dataset that contains heights, weights and genders of 10,000 people. We will compute the average weight and height of males and females, and compare them. You already know how the averages would compare (think about it).

The dataset is taken from the book *Machine Learning for Hackers* by Conway and Myles White, and is provided in the file `height_weight_gender.mat` that accompanies this file on the course website.

First download the dataset and change the directory in Matlab to where the data is. You can use the function `cd()` to do this. Then load the data as described below.

```
>> clearvars;    % clear workspace
>> load('height_weight_gender.mat');
>> whos          % show variables in workspace
  Name            Size              Bytes  Class      Attributes

  gender      10000x1             10000  logical
  height      10000x1             80000  double
  weight      10000x1             80000  double
```

This tells us that `gender`, `height` and `weight` are all vectors with 10,000 elements each. Gender is coded as 1 for male, and 0 for females. You should check how many males and females are in the dataset (you can use the function `sum()`).

Heights and weights are in inches and pounds respectively, so let's convert them to the metric system, to meters and kilograms respectively:

```
>> height = height * 0.025;
>> weight = weight * 0.454;
```

Exercise 1.10 Do the following and discuss with other students.

1. The average height of females and the average height of males.
2. The average weight of females and the average weight of males.
3. Do these numbers make sense?
4. What is the average weight of females whose height is between 1.6m to 1.7m?

Hint: Use the function `mean()` to compute the average of numbers. ■

Exercise 1.11 Compare the histograms of males and females. For a good visualization, it is useful to make all the plots in the same figure. Use `subplot()` to show multiple plots in the same figure. Make a grid of 3×2 for a total of 6 plots. You can use `axis()` to set the axis limits of each plot (after plotting the histograms).

1. Plot the histogram of weights for the whole population (what does 'population' mean?).
2. Plot the histogram of weights for females.
3. Plot the histogram of weights for males.
4. Repeat above for the height measurements.
5. Compare these distributions. Do they look like what you expect them to? Discuss with other students.

■

So far we looked at the weight and height independently, but we expect them to be correlated.

Exercise 1.12 Make a scatter plot of height vs weight. Think about the following:

1. How well can we predict the height of a person given his/her weight, and vice-versa?
2. How well can we predict gender of a person given the weight and height?
3. Which task is more difficult: first or second?

Hint: By default `plot()` will join points with lines, so make sure that you specify that you only want it to draw a point for each using `'.'`. ■