Homework 1

Jake Lynn

Q1

In [ ]:
```
%pip install scikit-learn
```

```
305.78s - pydevd: Sending message related to process being replaced timed-out after 5 seconds
```

```
DEPRECATION: Configuring installation scheme with distutils config files is
deprecated and will no longer work in the near future. If you are using a Ho
mebrew or Linuxbrew Python, please see discussion at https://github.com/Home
brew/homebrew-core/issues/76621
Collecting scikit-learn
  Downloading scikit_learn-1.3.1-cp39-cp39-macosx_10_9_x86_64.whl (10.2 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 10.2/10.2 MB 42.6 MB/s eta 0:0
0:0000:010:01
Collecting scipy>=1.5.0
  Downloading scipy-1.11.2-cp39-cp39-macosx_10_9_x86_64.whl (37.2 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 37.2/37.2 MB 28.0 MB/s eta 0:0
0:0000:0100:01
Collecting joblib>=1.1.1
  Using cached joblib-1.3.2-py3-none-any.whl (302 kB)
Requirement already satisfied: numpy<2.0,>=1.17.3 in /usr/local/lib/python3.
9/site-packages (from scikit-learn) (1.19.4)
Collecting threadpoolctl>=2.0.0
  Using cached threadpoolctl-3.2.0-py3-none-any.whl (15 kB)
Collecting numpy<2.0,>=1.17.3
  Downloading numpy-1.26.0-cp39-cp39-macosx_10_9_x86_64.whl (20.6 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 20.6/20.6 MB 33.0 MB/s eta 0:0
0:0000:0100:01
Installing collected packages: threadpoolctl, numpy, joblib, scipy, scikit-l
earn
  DEPRECATION: Configuring installation scheme with distutils config files i
s deprecated and will no longer work in the near future. If you are using a
Homebrew or Linuxbrew Python, please see discussion at https://github.com/Ho
mebrew/homebrew-core/issues/76621
  Attempting uninstall: numpy
    Found existing installation: numpy 1.19.4
    Uninstalling numpy-1.19.4:
      Successfully uninstalled numpy-1.19.4
  DEPRECATION: Configuring installation scheme with distutils config files i
s deprecated and will no longer work in the near future. If you are using a
Homebrew or Linuxbrew Python, please see discussion at https://github.com/Ho
mebrew/homebrew-core/issues/76621
  DEPRECATION: Configuring installation scheme with distutils config files i
s deprecated and will no longer work in the near future. If you are using a
Homebrew or Linuxbrew Python, please see discussion at https://github.com/Ho
mebrew/homebrew-core/issues/76621
  DEPRECATION: Configuring installation scheme with distutils config files i
s deprecated and will no longer work in the near future. If you are using a
Homebrew or Linuxbrew Python, please see discussion at https://github.com/Ho
mebrew/homebrew-core/issues/76621
  DEPRECATION: Configuring installation scheme with distutils config files i
s deprecated and will no longer work in the near future. If you are using a
Homebrew or Linuxbrew Python, please see discussion at https://github.com/Ho
mebrew/homebrew-core/issues/76621
DEPRECATION: Configuring installation scheme with distutils config files is
deprecated and will no longer work in the near future. If you are using a Ho
mebrew or Linuxbrew Python, please see discussion at https://github.com/Home
brew/homebrew-core/issues/76621
Successfully installed joblib-1.3.2 numpy-1.26.0 scikit-learn-1.3.1 scipy-1.
11.2 threadpoolctl-3.2.0

[notice] A new release of pip available: 22.3.1 -> 23.2.1
```

```
[notice] To update, run: python3.9 -m pip install --upgrade pip
Note: you may need to restart the kernel to use updated packages.
```

Q2

In [ ]:
```
from sklearn.datasets import load_diabetes
data = load_diabetes()
x, y = data.data, data.target
print(data.DESCR)
```

```
.. _diabetes_dataset:

Diabetes dataset
----------------

Ten baseline variables, age, sex, body mass index, average blood
pressure, and six blood serum measurements were obtained for each of n =
442 diabetes patients, as well as the response of interest, a
quantitative measure of disease progression one year after baseline.

**Data Set Characteristics:**

  :Number of Instances: 442

  :Number of Attributes: First 10 columns are numeric predictive values

  :Target: Column 11 is a quantitative measure of disease progression one ye
ar after baseline

  :Attribute Information:
      - age     age in years
      - sex
      - bmi     body mass index
      - bp      average blood pressure
      - s1      tc, total serum cholesterol
      - s2      ldl, low-density lipoproteins
      - s3      hdl, high-density lipoproteins
      - s4      tch, total cholesterol / HDL
      - s5      ltg, possibly log of serum triglycerides level
      - s6      glu, blood sugar level

Note: Each of these 10 feature variables have been mean centered and scaled
by the standard deviation times the square root of `n_samples` (i.e. the sum
of squares of each column totals 1).

Source URL:
https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html

For more information see:
Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani (2004) "L
east Angle Regression," Annals of Statistics (with discussion), 407-499.
(https://web.stanford.edu/~hastie/Papers/LARS/LeastAngle_2002.pdf)
```

Dataset Introduction

The 442 instances of data in this dataset represent information about people with diabetes. The target variable in this dataset is a number which represents progression of the disease in the year after a baseline value was collected. Ensuring that data points are easy to use, ex. ensuring all numerical points are represented as numbers instead of strings, will be important in preprocessing. I would like to do regressions from each individual data point to the target variable using scatter plots to see which may be the most descriptive features. Correlation matricies will be helpful in this instance as well, as I suspect many of these features may be related to one another.

Q3

In [ ]:
```
%pip install numpy
import numpy as np
```

```
1274.06s - pydevd: Sending message related to process being replaced timed-o
ut after 5 seconds
DEPRECATION: Configuring installation scheme with distutils config files is
deprecated and will no longer work in the near future. If you are using a Ho
mebrew or Linuxbrew Python, please see discussion at https://github.com/Home
brew/homebrew-core/issues/76621
Requirement already satisfied: numpy in /usr/local/lib/python3.9/site-packag
es (1.26.0)
DEPRECATION: Configuring installation scheme with distutils config files is
deprecated and will no longer work in the near future. If you are using a Ho
mebrew or Linuxbrew Python, please see discussion at https://github.com/Home
brew/homebrew-core/issues/76621

[notice] A new release of pip available: 22.3.1 -> 23.2.1
[notice] To update, run: python3.9 -m pip install --upgrade pip
Note: you may need to restart the kernel to use updated packages.
```

In [ ]:
```
a = np.array([[1, 2], [12, 5], [6, 9], [4, 3]])
b = np.array([[4, 5, 6, 7], [9, 8, 7, 6]])
```

Matricies can only be added if they are the same shape, so we check if their shapes are the same.

In [ ]:
```
if a.shape == b.shape:
  print("Can add. Result: ", np.add(a, b))
else:
  print('Cannot add.')
```

```
Cannot add.
```

Lets transpose A and try again.

In [ ]:
```
aT = np.transpose(a)
if aT.shape == b.shape:
  print('Can add. Result: ', np.add(aT, b))
else:
  print('Cannot add.')
```

```
Can add. Result:  [[ 5 17 12 11]
 [11 13 16  9]]
```

Multiplication:

In [ ]:
```
print(np.matmul(a, b))
print(np.matmul(b, a))
```

```
[[ 22  21  20  19]
 [ 93 100 107 114]
 [105 102  99  96]
 [ 43  44  45  46]]
[[128 108]
 [171 139]]
```

In order to check if a matrix is invertible, we need to check if it is square, and its determinant is 0.

In [ ]:
```
def check_matrix_invertible(matrix):
  if matrix.shape[0] != matrix.shape[1]:
    return False
  else:
    d = np.linalg.det(matrix)
    return np.abs(d) > 0
```

In [ ]:
```
print(check_matrix_invertible(a))
print(check_matrix_invertible(np.matmul(a, b)))
```

```
False
True
```

Q4

a: .6*.6 = .36

b: .4*.4 = .16

c: 1 - (.36 + .16) = .48

Q5

a:

P(B) = 0.3

P(BB) = 0.1

P(N) = 1 - (0.3 + 0.1) = 0.6

P(D|BB) = 0.73

P(D|B) = 0.4

P(D|N) = 0.19

Law of Total Prob:

P(D) = P(B) * P(D|B) + P(BB) * P(D|BB) + P(N) * P(N|D)

P(D) = 0.3 * 0.4 + 0.1 * 0.73 + 0.6 * 0.19 = 0.307

30.7%

b:

Find P(BB|D) and P(N|D)

Bayes:

P(BB|D) = (P(D|BB) * P(N))/(P(D))

P(BB|D) = (0.73 * 0.1) / 0.307 = 0.237

If a white cat is deaf, there is a 23.7% chance it has BB.

P(N|D) = (P(D|N) * P(N))/P(D)

P(N|D) = (0.19 * 0.6) / 0.307 = .371

There is a 37.1% chance a deaf white cat does not have blue eyes.