HW1

September 18, 2023

0.0.1 Question 1

Created conda env and ran `conda install scikit-learn`

0.0.2 Question 2

[2]:

```python
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 year
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) "Least
Angle Regression," Annals of Statistics (with discussion), 407-499.
(https://web.stanford.edu/~hastie/Papers/LARS/LeastAngle_2002.pdf)
```

The dataset is about diabetes patients and their attributes, relating to how much the disease progresses one year after measurement. There are 442 data points in total, and the target feature is the quantitative measurement of disease progression. Some plots that could be included in the EDA are scatter plots between each of the numerical attributes and the target measurement, as well as a boxplot for sex vs. target. For pre-processing, we would need to consider missing data and decide whether to remove or impute it, find the outliers and decide whether to ignore them, and scale the many numerical attributes to work with each other.

0.0.3 Question 3

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{bmatrix}, B = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}$$

$A + B$ cannot be calculated because the dimensions do not match.

But, $A^T + B$ can be, because A^T is a 2-by-4 matrix.

$$A^T + B = \begin{bmatrix} 1 & 3 & 5 & 7 \\ 2 & 4 & 6 & 8 \end{bmatrix} + \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix} = \begin{bmatrix} 2 & 5 & 8 & 11 \\ 7 & 10 & 13 & 16 \end{bmatrix}$$

$$AB = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix} = \begin{bmatrix} 11 & 14 & 17 & 20 \\ 23 & 30 & 37 & 44 \\ 35 & 46 & 57 & 68 \\ 47 & 62 & 77 & 92 \end{bmatrix}$$

$$BA = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 50 & 60 \\ 114 & 140 \end{bmatrix}$$

We cannot calculate A^{-1} because A is a 4-by-2 matrix and you can only find the inverse of square matrices. We also cannot calculate $(AB)^{-1}$ because the determinant of $AB = 0$.

0.0.4 Question 4

Part a: Probability that two successive nucleotides are G-C:

$$P(2GC) = P(GC) * P(GC) = 0.6 * 0.6 = 0.36$$

Part b: Probability that two successive nucleotides are A-T:

$$P(2AT) = P(AT) * P(AT) = 0.4 * 0.4 = 0.16$$

Part c Probability that two successive nucleotides are different:

$$P(2\text{diff}) = P(GC) * P(AT) + P(AT) * P(GC) = 0.6 * 0.4 + 0.4 * 0.6 = 2 * 0.6 * 0.4 = 0.48$$

0.0.5 Question 5

Part a: Let B be the event that the cat has one blue eye, let BB be the event that the cat has two blue eyes and let N be the event that the cat does not have blue eyes. Let D be the event that the cat is deaf.

$$\begin{gathered}
P(D) = P(D \cap N) + P(D \cap B) + P(D \cap BB) \\
P(D) = P(D|N)P(N) + P(D|B)P(B) + P(D|BB)P(BB) \\
P(D) = (0.19)(0.6) + (0.4)(0.3) + (0.73)(0.1) \\
P(D) = 0.114 + 0.12 + 0.073 \\
P(D) = 0.307
\end{gathered}$$

Part b:

$$\begin{gathered}
P(BB|D) = \frac{P(BB \cap D)}{P(D)} \\
P(BB|D) = \frac{0.073}{0.307} = 0.2378
\end{gathered}$$

$$\begin{gathered}
P(N|D) = \frac{P(N \cap D)}{P(D)} \\
P(N|D) = \frac{0.114}{0.307} = 0.3713
\end{gathered}$$