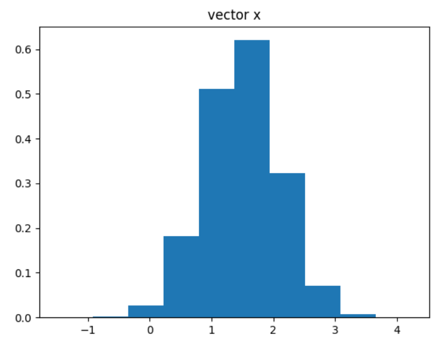
Intro to ML
PS1 Report

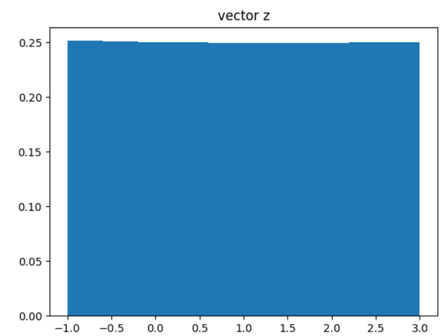
Name: Timothy Horrell
ID: 4475990

Questions:

1) A regression problem could involve predicting an SAT score increase for the next time a student takes the test.
    a. The features I would use would be practice problems answered.
    b. The label would be the increase of SAT score of the next test taken.
    c. I would collect by looking at existing college board data on student scores from test to test, along with survey data on time studied.
    d. The problem could be challenging because everyone will have varied performance on the SAT. Each student starts at a different baseline score also, which may limit ability to grow if a student achieves close to a maximum score.
2) A classification problem could involve determining what pitch a pitcher threw.
    a. The features I would use would be pitch speed and spin rate.
    b. The labels would be any pitch the pitcher has in their arsenal (ex. Fastball, changeup, curveball, slider). This varies from pitcher to pitcher, but I will assume the selection of a pitcher that only throws these four pitches.
    c. I would collect data from a baseball pitch database. The statcast baseball database has access to pitch speed and spin rate. All data will be from the specific pitcher whose pitch I wish to predict.
    d. Some challenges I could run into are that spin rate and speed has increased year to year, so curveball data from recent years (a medium, high spin pitch) may look like a slider from older years (a medium speed, medium spin pitch). Additionally, the MLB baseball changes year to year and results in league-wide spin rate changes based on the grip on the baseball, which could disrupt the ability to predict across years of this pitcher's career.
3) .
    a. .
    b. .
    c. The histogram for x does look like a Gaussian distribution, however slightly skewed right. The histogram for z does look like a uniform distribution.
    (Images shown below)



Ps1-3-c-1.png



ps1-3-c-2.png

d. Loop Add Time: 0.205s

e. Non-Loop Add Time: 0.002s. It is significantly more efficient to add a constant to a vector without using a loop method.

f. Elements 1st Time: 374,857
Elements 2nd Time: 375,267
Elements 3rd Time: 374,523
There is a small difference between each time the code is run. This can be attributed to vector z being initialized as a randomized uniform distribution. This likely uses some internal clock or seed to randomize data, which would update each time the code is ran.

4) .

a. .

b. X = 0.3, Y = 0.4, Z = 0

c. X1, L1 Norm = 0.5 + 0 + 1.5 = 2
X2, L1 Norm = 1 + 1 + 0 = 2
X1, L2 Norm = sqrt(0.5^2 + 0 + 1.5^2) = 1.58
X2, L2 Norm = sqrt(1^2 + 1^1 + 0) = 1.41

```
x1, l1 norm: [2.]
x2, l1 norm: [2.]
x1, l2 norm: 1.5811388300841898
x2, l2 norm: 1.4142135623730951
```

5) .

```
The input array is[[1 2 3]
 [4 5 6]]
The sum squared vector 1 is: [17. 29. 45.]
The input array is[[2 2]
 [3 3]
 [4 4]
 [5 5]]
The sum squared vector 2 is: [54. 54.]
```