1. Problem 1: Probability and Expectation (22 points)

- In a bag there are 50 balls, each having one of three colors: red (25 balls), blue (15 balls), and green (10 balls). When a ball is drawn from the bag, you get different points for every color:
-
- Red balls are worth 3 points each.
- Green balls are worth 5 points each.
- Blue balls are worth 10 points each.

1. Suppose that you draw a ball at random from the bag. What is the expected total point value of the ball you are drawing?
   $P(R) = 25/50 = 1/2 * 3 = 1.5$
   $P(G) = 10/50 = 1/5 * 5 = 1$
   $P(B) = 15/50 = 3/10 * 10 = 3$
   $E(x) = 1.5 + 1 + 3 = 5.5.$ E(x) = 5.5

2. I. X is the random variable that denotes the number of points you get by drawing a ball at random, what is the variance of X?
   $E(X^2) = (3^2 * 1/2) + (5^2 * 1/5) + (10^2 * 3/10)$
   $= 4.5 + 5 + 30 = 39.5$
   Var(X) = $E[X^2] - E^2[X]$
   = 39.5 - (5.5)^2 = 9.25 Var(X) = 9.25

3. Three balls were picked at random from the bag and it turned out to be two red balls and one green ball. Now, you are drawing another ball from the remaining ones. What is the expected total point value of the ball you are drawing?
   $P(R) = 23/47 * 3 \approx 1.468$
   $P(G) = 9/47 * 5 \approx 0.957$
   $P(B) = 15/47 * 10 \approx 3.19.$
   $E(x) = 1.468 + 0.957 + 3.191 \approx 5.616$ E(x) ≈ 5.616

4. Suppose that X is a random variable, with expectation $E(X) = 10$ and variance $Var(X) = 2$. What is the value of $E[X(X-1)]$?

$$E[X(X-1)] = E(X^2) - E(X)$$

Var(X) = $E[X^2] - E^2[X]$

$2 = E[X^2] - 10^2 \qquad E[X^2] = 102$

$E[X(X-1)] = 102 - 10 = 92$ E[X (X – 1)] = 92

2. Problem 2: Conditional Probability and Bayes Theorem (18 points)
Probability theory and the Bayes theorem serve as the basis of important machine learning techniques such as linear discriminant analysis (LDA), density estimation, and Naive Bayes (NB) classifiers, which are very effective in text classification settings.

A spam filter is designed by looking at commonly occurring phrases in spam. Suppose that 80% of email is spam. In 10% of the spam emails, the phrase "free money" is used, whereas this phrase is only used in 1% of non-spam emails.

1. What is the probability that an email has the phrase "free money"?
80/100 = P(S)
20/100 = P(N)
0.01 = P(F | N)
0.1 = P(F | S)
P(F) = (0.1 * 0.8) + (0.01 * 0.2) = 0.08 + 0.002 P(F) = 0.082 = 8.2%

2. A new email has just arrived, which does mention "free money". What is the probability that it is spam?
P(S | F) = $\frac{P(F|S)P(S)}{P(F)}$
= $\frac{0.1*0.8}{0.082}$ = 0.9756097561 P(S | F) ≈ 0.976 = 97.6%

3. Another email does not have the phrase "free money". What is the probability that it is not spam?
P("Not Spam" | "Not Free Money") = $\frac{P(\neg F|N)P(N)}{P(\neg F)}$
= $\frac{0.99*0.2}{0.918}$
= 0.2156862745 P(N |¬F) ≈ 0.216 = 21.6%

3. Problem 3: Matrices and vectors (20 points)
Since data is represented in matrices, many of the techniques that we will see, from linear regression to dimensionality reduction and neural networks, will boil down to matrix operations. Optimized matrix operations from modern packages allow us to train and test algorithms much faster than in an iterative approach. Define the following two matrices of size 4 × 4:

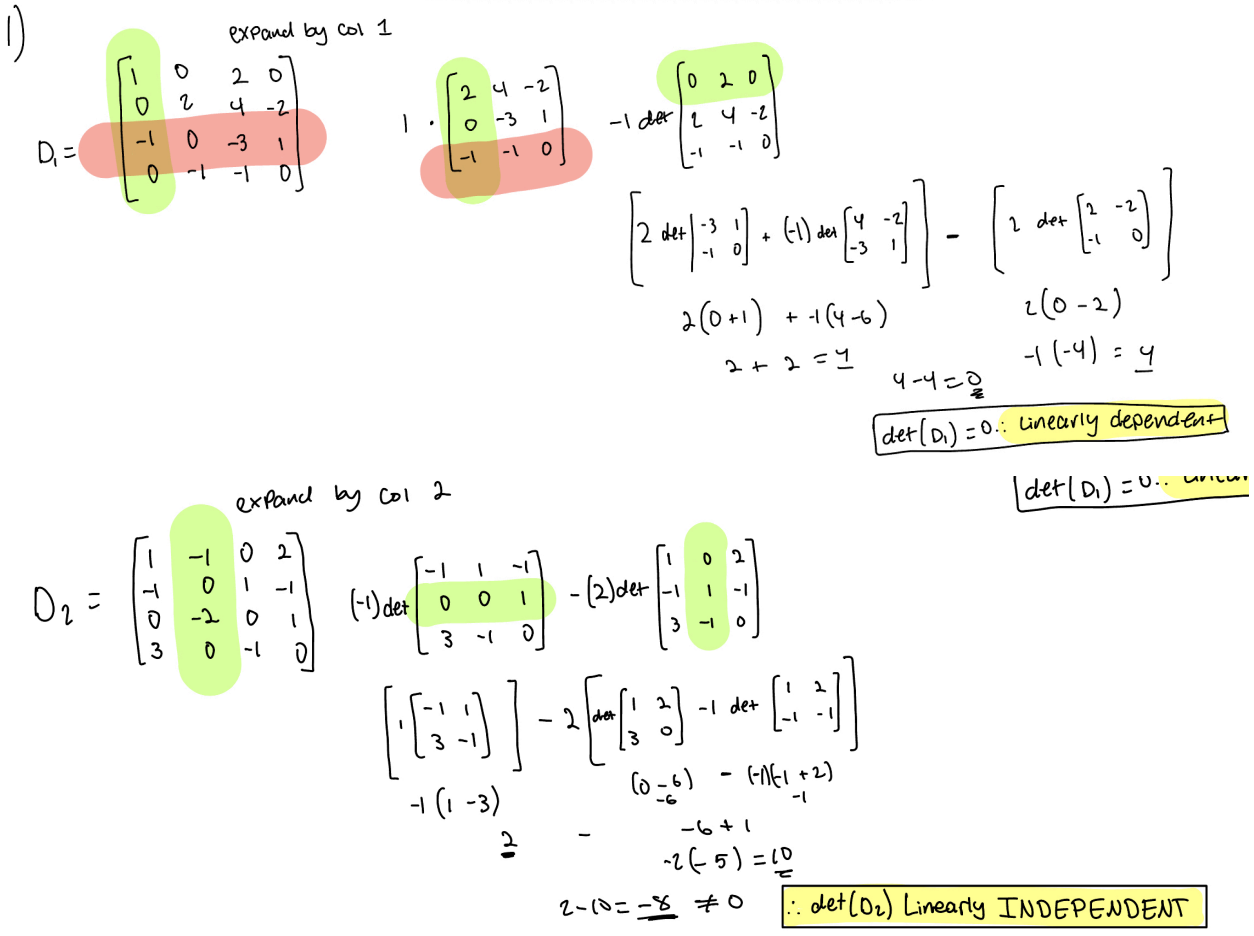
$$D_1 = \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 2 & 4 & -2 \\ -1 & 0 & -3 & 1 \\ 0 & -1 & -1 & 0 \end{bmatrix}$$

$$D_2 = \begin{bmatrix} 1 & -1 & 0 & 2 \\ -1 & 0 & 1 & -1 \\ 0 & -2 & 0 & 1 \\ 3 & 0 & -1 & 0 \end{bmatrix}$$

1. Are the 4 columns in each of the matrix linearly independent or linearly dependent?

   D1 = Linearly dependent

   D2 = Linearly Independent



2. For each of the matrices above, find the maximum number of columns that are linearly independent.

   $det(D_1) = 0$ and linearly dependent, therefore rank < 4,

   Since there exists a nonzero 3×3 subset, **the maximum number of columns that are linearly independent is 3**

. $det(D_2) = -8 \neq 0$

Nonzero determinant = **4 linearly independent columns**

3. What is the rank of each matrix? Use the procedure we discussed in class to compute the rank of each matrix, and show the steps you performed.

Row 1: $1 \cdot c_1 + 0 \cdot c_2 + 2 \cdot c_3 + 0 \cdot c_4 = 0 \implies c_1 + 2c_3 = 0$

Row 2: $0 \cdot c_1 + 2 \cdot c_2 + 4 \cdot c_3 - 2 \cdot c_4 = 0 \implies 2c_2 + 4c_3 - 2c_4 = 0$

Row 3: $-1 \cdot c_1 + 0 \cdot c_2 - 3 \cdot c_3 + 1 \cdot c_4 = 0 \implies -c_1 - 3c_3 + c_4 = 0$

Row 4: $0 \cdot c_1 - 1 \cdot c_2 - 1 \cdot c_3 + 0 \cdot c_4 = 0 \implies -c_2 - c_3 = 0$

$c_1 + 2c_3 \mid 2c_2 + 4c_3 - 2c_4 \mid -c_1 - 3c_3 + c.4 \mid -c_2 - c_3$

Row 1: $c1 = -2c_3$

Row 4: $c2 = -c3$ Row 2: $0 = -2(c_3) + 4c_3 - 2c_4 => 2c_3 - 2c_4 = 0 => \underline{c3 = c4}$

Row 3: $-c_1 - 3c_3 + c_4 = 0 => 2c_3 - 3c_3 + c_3 = 0 => 3c_3 - 3c_3 = 0 => 0 = 0$

$c_3 = x$

$c_4 = x$

$c_1 = -2x$, $c_2 = -x$

Since there is 1 free unknown, 4 columns - 1 free = 3, therefore, Rank(D1) = 3

Row 1: $(c_1 - c_2 + 2c_4 = 0)$

Row 2: $(-c_1 + c_3 - c_4 = 0)$

Row 3: $(-2c_2 + c_4 = 0)$

Row 4: $(3c_1 - c_3 = 0)$

Row 4: $c_3 = 3c_1$ Row 3: $c_4 = 2c_2$

Row 2: $-c_1 + c_3 - c_4 = 0 \implies -c_1 + 3c_1 - 2c_2 = 0 \implies 2c_1 - 2c_2 = 0 \implies c_2 = c_1$

Row 1: $c_1 - c_2 + 2c_4 = c_1 - c_1 + 2(2c_2) = 4c_1 = 0 \implies c_1 = 0$

$c_1 = 0 \implies c_2 = 0, \quad c_3 = 0, \quad c_4 = 0$

The solution is trivial, this means there are no free unknown,

therefore Rank(D2) = 4 - 0 = 4

4. Assume that we define a column vector θ. Which of the products θD_1, $D_1\theta$, $\theta^T D_1$, and $D_1\theta^T$ can be computed? For the ones that can be computed, please write the result and its dimension. Please perform these computations by hand and show your work. Do not use a calculator for this problem.

$$\theta = \begin{bmatrix} -1 \\ 0 \\ 1 \\ 2 \end{bmatrix}$$

4.

$\theta = (4 \times 1) \quad D_1 = (4 \times 4)$

a) $\theta D_1 = (4\times1)(4\times4)$

does NOT match

Cannot be Computed

b) $D_1\theta = (4\times4)(4\times1)$

matches ✓ can be computed

$$\begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 2 & 4 & -2 \\ -1 & 0 & -3 & 1 \\ 0 & -1 & -1 & 0 \end{bmatrix} \times \begin{bmatrix} -1 \\ 0 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -1+0+2+0 \\ 0+0+4-4 \\ 1+0-3+2 \\ 0+0-1+0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix}$$

c) $\theta^T D_1 = (1\times4)(4\times4)$

matches ✓ can be computed

$$[-1,0,1,2] \times \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 2 & 4 & -2 \\ -1 & 0 & -3 & 1 \\ 0 & -1 & -1 & 0 \end{bmatrix} = \begin{matrix} (-1+0-1+0), \\ (0+0+0-2), \\ (-2+0-3-2), \\ (0+0+1+0) \end{matrix} = [-2,-2,-7,1]$$

d) $D_1\theta^T = (4\times4)(1\times4)$

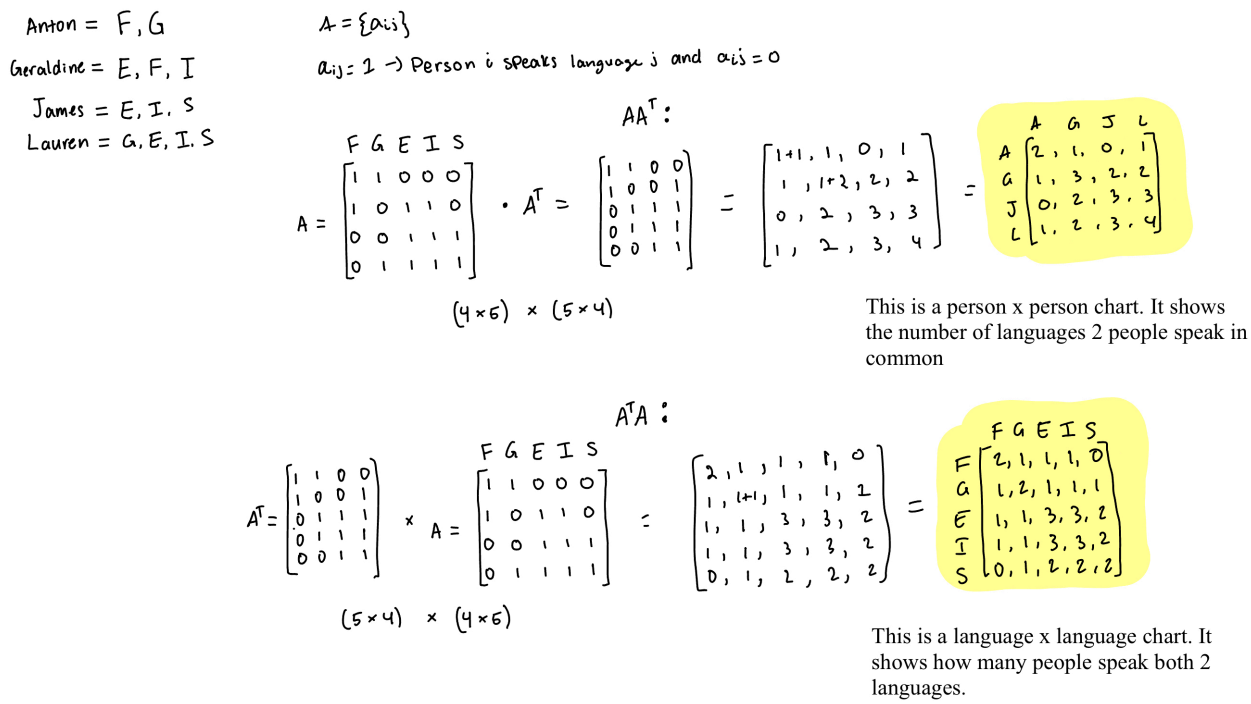
does NOT Match

Cannot be Computed

4. Problem 4: Matrix transpose and inverse (20 points)

1. Anton speaks French and German; Geraldine speaks English, French and Italian; James speaks English, Italian, and Spanish; Lauren speaks all the languages the others speak except French; and no one speaks any other language. Make a matrix $A = \{a_{ij}\}$ with rows representing the four people mentioned and columns representing the languages they speak. Put $a_{ij} = 1$ if person i speaks language j and $a_{ij} = 0$ otherwise.

   Compute the matrices AA^T and A^TA . Explain the significance of the matrices AA^T and A^TA. What does each entry in these two matrices represent?



2. Generate at random 3 matrices of size 3×3 and fill each entry with a random integer chosen from -10 to 10. Use an existing package to compute the inverse of each matrix if it exists. Include in your report:

   - Each of the 3 matrices
   - The inverses computed with the package
   - Compute the product of each matrix with its inverse to check that you obtain the identity matrix.

     Answer:

     https://github.com/haderie/DS4400/blob/main/HW1/Problem4.ipynb

5. Problem 5: Average, variance, and correlation (20 points)

   The dataset for this assignment is available here. The prediction task is to predict the price of a house (column **price**) given the other features. Please ignore the columns **id** and **date**, as well as the categorical column **zipcode**. File **kc_house_data.csv** includes all the records in the dataset. You should use the entire dataset for the assignment. You can also find a Word document including the feature description in the same folder. In this problem, we will perform some exploratory data analysis using the house price dataset.

   1. For each feature, write code to compute the average value, the min and max values, as well as its variance.

      Which features have the lowest and the highest average? Include the feature name and their average values for the features with the lowest and highest average.

Which features have the lowest and the highest variance? Include the feature name and their variance values for the features with the lowest and highest variance.

2. Compute the correlation coefficient of each feature with the response. Include a table with the correlation coefficient of each feature with the response. Which features are positively correlated (i.e., have positive correlation coefficient) with the response?
3. Which feature has the highest positive correlation with the response? 3. Were you able to find any features with a negative correlation coefficient with the response?

Answer:

https://github.com/haderie/DS4400/blob/main/HW1/Problem5.ipynb