DS4400 - HW1

Rayn Anwar

January 2026

1 Problem 1: Probability and Expectation

1.
$$\mathbb{E}[X] = \sum_{i=1}^{n} x_i \cdot p_i$$
$$\mathbb{E}[X] = (3 * .5) + (5 * .2) + (10 * .3)$$

$$\mathbb{E}[X] = 1.5 + 3 + 1 = 5.5$$

2.
$$\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$
$$\mathbb{E}[X^2] = \sum_{i=1}^{n} x_i^2 \cdot p_i$$
$$\mathbb{E}[X^2] = (3^2 * .5) + (5^2 * .2) + (10^2 * .3) = 39.5$$
$$\mathrm{Var}[X] = 39.5 - (5.5)^2 = 9.25$$

3.
$$\mathbb{E}[X] = \sum_{i=1}^{n} x_i \cdot p_i$$
$$\mathbb{E}[X] = (3 * \frac{23}{47}) + (5 * \frac{15}{47}) + (10 * \frac{9}{47})$$

$$\mathbb{E}[X] = 5.62$$

4.
$$\mathbb{E}[X(X-1)] = \mathbb{E}[X^2 - X] = \mathbb{E}[X^2] - \mathbb{E}[X]$$

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = 2$$
$$\mathbb{E}[X^2] - 10^2 = 2$$

$$\mathbb{E}[X^2] = 102$$

Plug in to the original equation:
$$102 - 10 = 92$$

2 Problem 2: Conditional Probability and Bayes Theorem

1. F - Free Money
   S - Spam
   N - Not Spam

$$P(F) = P(F|S)\dot{P}(S) + P(F|N)\dot{P}(N)$$
$$P(F) = (0.10 \cdot 0.80) + (0.01 \cdot 0.2) = 0.08 + 0.002 = 0.082$$

2.

$$P(S|F) = \frac{P(F|S)P(S)}{P(F)}$$
$$\frac{0.1 * 0.8}{0.082} = 0.976$$

   97.6% chance of the email being spam.

3.

$$P(N|F^c) = \frac{P(F^c|N)P(N)}{P(F^c)}$$
$$P(F^c|N) = 1 - P(F|N) = 0.99$$
$$P(F^c) = 1 - P(F) = 0.918$$
$$\frac{0.99 \cdot 0.01}{0.918} = 0.216$$

3 Problem 3: Matrices and Vectors

1. No, the 4 columns aren't linearly independent in D_1. The determinant of the matrix is 0, therefore one of the columns is a linear combination of the other columns.

2. In matrix D_1, there are 3 columns that are linearly independent. In matrix D_2, all 4 columns are linearly independent.

3.

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

   Step 1: ($R_3 = R_3 + R_1$):

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

   Step 2: ($R_4 = R_4 + \frac{1}{2}R_2$):

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

   Step 3: ($R_4 = R_4 + R_3$):

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

Count nonzero rows: $\text{rank}(D_1) = 3$.

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

$(R_2 = R_2 + R_1,\ R_4 = R_4 - 3R_1)$:

$$\begin{bmatrix} 1 & -1 & 0 & 2 \\ 0 & -1 & 1 & 1 \\ 0 & -2 & 0 & 1 \\ 0 & 3 & -1 & -6 \end{bmatrix}$$

$(R_3 = R_3 - 2R_2,\ R_4 = R_4 + 3R_2)$:

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

$(R_4 = R_4 + R_3)$**:**

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

Count nonzero rows: All 4 rows are nonzero, therefore

$$\text{rank}(D_2) = 4$$

and all 4 columns are linearly independent.

4.

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

$$D_1 \cdot \theta = \begin{bmatrix} 1 \cdot (-1) + 0 \cdot 0 + 2 \cdot 1 + 0 \cdot 2 \\ 0 \cdot (-1) + 2 \cdot 0 + 4 \cdot 1 + (-2) \cdot 2 \\ -1 \cdot (-1) + 0 \cdot 0 + (-3) \cdot 1 + 1 \cdot 2 \\ 0 \cdot (-1) + (-1) \cdot 0 + (-1) \cdot 1 + 0 \cdot 2 \end{bmatrix} = \begin{bmatrix} -1 + 0 + 2 + 0 \\ 0 + 0 + 4 - 4 \\ 1 + 0 - 3 + 2 \\ 0 + 0 - 1 + 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix}$$

The dimensions is 4x1

5.

$$\theta^\top = \begin{bmatrix} -1 & 0 & 1 & 2 \end{bmatrix}, \quad D_1 = \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 2 & 4 & -2 \\ -1 & 0 & -3 & 1 \\ 0 & -1 & -1 & 0 \end{bmatrix}$$

$$\theta^\top \cdot D_1 = \begin{bmatrix} -1 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 2 & 4 & -2 \\ -1 & 0 & -3 & 1 \\ 0 & -1 & -1 & 0 \end{bmatrix} = \big[-1 \cdot 1 + 0 \cdot 0 + 1 \cdot (-1) + 2 \cdot 0 \quad -1 \cdot 0 + 0 \cdot 2 + 1 \cdot 0 + 2 \cdot (-1) \quad -1 \cdot$$

$$\theta^\top \cdot D_1 = \begin{bmatrix} -1 + 0 - 1 + 0 & 0 + 0 + 0 - 2 & -2 + 0 - 3 - 2 & 0 + 0 + 1 + 0 \end{bmatrix} = \begin{bmatrix} -2 & -2 & -7 & 1 \end{bmatrix}$$

The dimension is 1x4

4 Problem 4: Matrix transpose and inverse

1.

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$AA^\top = \begin{bmatrix} 2 & 1 & 0 & 1 \\ 1 & 3 & 2 & 2 \\ 0 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

$$A^\top A = \begin{bmatrix} 2 & 1 & 1 & 1 & 0 \\ 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 3 & 3 & 2 \\ 1 & 1 & 3 & 3 & 2 \\ 0 & 1 & 2 & 2 & 2 \end{bmatrix}$$

AA^T shows how many languages each pair of people share (diagonal = languages known, off-diagonal = shared languages), while $A^T A$ shows how many people speak each pair of languages (diagonal = speakers per language, off-diagonal = co-occurring languages).

```
Matrix 1:
[[-4  9  4]
 [ 0 -3 10]
 [-4  8  0]]

Inverse of Matrix 1:
[[ 0.90909091 -0.36363636 -1.15909091]
 [ 0.45454545 -0.18181818 -0.45454545]
 [ 0.13636364  0.04545455 -0.13636364]]

Product: [[ 1.  0. -0.]
 [ 0.  1. -0.]
 [ 0.  0.  1.]]

Matrix 2:
[[ 0 10 -7]
 [-3 -8 10]
 [-9  1 -5]]

Inverse of Matrix 2:
[[-0.05714286 -0.08190476 -0.08380952]
 [ 0.2         0.12       -0.04      ]
 [ 0.14285714  0.17142857 -0.05714286]]
```

2.

5 Problem 5: Average, variance, and correlation

1.

2.

Link to notebook Jupiter Notebook

```
Inverse of Matrix 2:
[[-0.05714286 -0.08190476 -0.08380952]
 [ 0.2         0.12       -0.04      ]
 [ 0.14285714  0.17142857 -0.05714286]]

Product: [[ 1. -0. -0.]
 [ 0.  1. -0.]
 [ 0.  0.  1.]]

Matrix 3:
[[ -9  10 -10]
 [  1   1   6]
 [ -1   5   4]]

Inverse of Matrix 3:
[[-0.35135135 -1.21621622  0.94594595]
 [-0.13513514 -0.62162162  0.59459459]
 [ 0.08108108  0.47297297 -0.25675676]]

Product: [[ 1. -0.  0.]
 [ 0.  1. -0.]
 [-0. -0.  1.]]
```

```
...   Feature statistics:
                         Mean         Min           Max      Variance
      price         540088.141767  75000.0000  7.700000e+06  1.347824e+11
      bedrooms           3.370842      0.0000  3.300000e+01  8.650150e-01
      bathrooms          2.114757      0.0000  8.000000e+00  5.931513e-01
      sqft_living     2079.899736    290.0000  1.354000e+04  8.435337e+05
      sqft_lot       15106.967566    520.0000  1.651359e+06  1.715659e+09
      floors             1.494309      1.0000  3.500000e+00  2.915880e-01
      waterfront         0.007542      0.0000  1.000000e+00  7.485226e-03
      view               0.234303      0.0000  4.000000e+00  5.872426e-01
      condition          3.409430      1.0000  5.000000e+00  4.234665e-01
      grade              7.656873      1.0000  1.300000e+01  1.381703e+00
      sqft_above      1788.390691    290.0000  9.410000e+03  6.857347e+05
      sqft_basement    291.509045      0.0000  4.820000e+03  1.958727e+05
      yr_built        1971.005136   1900.0000  2.015000e+03  8.627973e+02
      yr_renovated      84.402258      0.0000  2.015000e+03  1.613462e+05
      lat               47.560053     47.1559  4.777760e+01  1.919990e-02
      long            -122.213896   -122.5190 -1.213150e+02  1.983262e-02
      sqft_living15   1986.552492    399.0000  6.210000e+03  4.697612e+05
      sqft_lot15     12768.455652    651.0000  8.712000e+05  7.455182e+08

      Lowest average feature: long -122.21389640494147
      Highest average feature: price 540088.1417665294

      Lowest variance feature: waterfront 0.007485225502686407
      Highest variance feature: price 134782378397.24687
```

```
...
  Correlation with price:
   price            1.000000
  sqft_living      0.702035
  grade            0.667434
  sqft_above       0.605567
  sqft_living15    0.585379
  bathrooms        0.525138
  view             0.397293
  sqft_basement    0.323816
  bedrooms         0.308350
  lat              0.307003
  waterfront       0.266369
  floors           0.256794
  yr_renovated     0.126434
  sqft_lot         0.089661
  sqft_lot15       0.082447
  yr_built         0.054012
  condition        0.036362
  long             0.021626
  Name: price, dtype: float64

  Features positively correlated with price: ['price', 'sqft_living', 'grade', 'sqft_above', 'sqft_living15', 'bathrooms', 'view', 'sqft_basement', 'bedrooms', 'lat', 'waterfront', 'floors', 'yr_renovated', 'sqft_lot', 'sqft_lot15', 'yr_built', 'condition',
  Feature with highest positive correlation: price 1.0

  Features negatively correlated with price:
   Series([], Name: price, dtype: float64)
```