Name: Michael Belmer

Homework 2: CSCI 347: Data Mining

Show your work. Include any code snippets you used to generate an answer, using comments in the code to clearly indicate which problem corresponds to which code.

Consider the following data matrix:

	X_1	X_2	X_3.
x_1	red	yes	North
x_2	blue	no	South
x_3	yellow	no	East
x_4	yellow	no	West
x_5	red	yes	North
x_6	yellow	yes	North
x_7	blue	no	West

1. [4 points] Use one-hot encoding to transform **all** the categorical attributes to numerical values. Write down the transformed data matrix. Call this new matrix Y.

$$
y = \begin{array}{cccccccccc}
 & X_{1R} & X_{1B} & X_{1Y} & X_{2Y} & X_{2N} & X_{3N} & X_{3S} & X_{3E} & X_{3W} \\
x_1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
x_2 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
x_3 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\
x_4 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\
x_5 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
x_6 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\
x_7 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1
\end{array}
$$

2. [2 points] What is the Euclidean distance between data instance x_2 (second row) and data instance x_7 (seventh row) after applying one-hot encoding?

$$\|x_2 - x_7\|_2 = \sqrt{(0-0)^2 + (1-1)^2 + (0-0)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2 + (1-0)^2 - (0-0)^2 + (0-1)^2}$$

$$= \sqrt{2}$$

3. [2 points] What is the cosine similarity (cosine of the angle) between data instance x_2 and data instance x_7 after applying one-hot encoding?

$$\cos(\theta) = \frac{0(0) + 1(1) + 0(0) + 0(0) + 1(1) + 0(0) + 1(0) + 0(0) + 0(1)}{\|x_2\|_2 \|x_7\|_2}$$

$$= \frac{2}{\sqrt{3}\sqrt{3}} = \frac{2}{3}$$

4. [2 points] What is the Hamming distance between data instance x_2 and data instance x_7?

$$= d - s = 3 - 2 = 1$$

5. [2 points] What is the Jaccard coefficient between data instance x_2 and data instance x_7 after applying one-hot encoding?

$$= \frac{s}{2d - s} = \frac{2}{6 - 2} = \frac{1}{2}$$

6. [2 points] What is the (multivariate) mean of Y?

$$\left(\frac{2}{7}\ \frac{2}{7}\ \frac{3}{7}\ \frac{4}{7}\ \frac{4}{7}\ \frac{3}{7}\ \frac{1}{7}\ \frac{1}{7}\ \frac{2}{7}\right)$$

7. [2 points] What is the sample variance of the first column of Y (using the matrix written in the answer to (1)) ?

$$\left(\frac{5}{7}\right)^2+\left(\frac{-2}{7}\right)^2+\left(\frac{-2}{7}\right)^2+\left(\frac{-2}{7}\right)^2+\left(\frac{5}{7}\right)^2+\left(\frac{-2}{7}\right)^2+\left(\frac{-2}{7}\right)^2$$

$$=1.429$$

8. [4 points] Write down the resulting matrix after applying standard (z-score) normalization to the matrix Y. Call this matrix Z.

Python Code:

```
from sklearn import preprocessing
import numpy as np
D=np.matrix('100101000; 010010100;
001010010; 001010001; 100101000; 00110
1000; 010010001')
standard_scalar = preprocessing.Standard scalar( )
standard_normalized_D = standard_scalar.fit_transform(D)
```

$$Z=\begin{array}{c|ccccccccc} & X_{1R} & X_{1B} & X_{1Y} & X_{2Y} & X_{2N} & X_{3N} & X_{3S} & X_{3E} & X_{3W} \\ \hline x_1 & 1.58 & -0.63 & -.87 & 1.15 & -1.15 & 1.15 & -.41 & -.41 & -.63 \\ x_2 & -0.63 & 1.58 & -.87 & -.87 & .87 & -.86 & 2.45 & -.41 & -.63 \\ x_3 & -0.63 & -0.63 & 1.15 & -.87 & .87 & -.87 & -.41 & 2.45 & -.63 \\ x_4 & -0.63 & -0.63 & 1.15 & -.87 & .87 & -.87 & -.41 & -.41 & 1.58 \\ x_5 & 1.58 & -0.63 & -.87 & 1.15 & -1.15 & 1.15 & -.41 & -.41 & -.63 \\ x_6 & -0.63 & -0.63 & 1.15 & 1.15 & -1.15 & 1.15 & -.41 & -.41 & -.63 \\ x_7 & -0.63 & 1.58 & -.87 & -.87 & .87 & -.87 & -.41 & -.41 & 1.58 \end{array}$$

9. [2 points] What is the (multivariate) mean of Z?

(0 0 0 0 0 0 0 0 0)

10. [2 points] Let z_i be the ith row of Z. What is the Euclidean distance between z_2 and z_7?

Python code:

```
import numpy.lnalg as LA
z2 = standard_normalized_D[1,:]
z7 = standard_normalized_D[6;:]
LA.norm(z2-z7)
```

= 3.16