Arlind Stafaj -- Data Mining -- 6930 Spring 2020

**Homework # 3**

**Question 1)**

A black and silver text on a screen

Description automatically generated

Classification on Salary: 10 training data instances. 4/10 = High, 6/10 = Low

**Instance 1:**

**P(Low | X)** where X = (x1 = High School, x2 = Service, x3 = <3)

*Without Laplace Smoothing:*

P(x1 | Low) \* P(x2 | Low) \* P(x3 | Low) \* P(Low) 🡪 (4/6) \* (4/6) \* (2/6) \* (6/10) = 8.88 %

*With Laplace Smoothing = 1:*

(4+1/6+2) \* (4+1/6+2) \* (2+1/6+3) \* (6/10) = (5/8) \* (5/8) \* (3/9) \* (6/10) = **7.8125 %**

**P(High | X)** where X = (x1 = High School, x2 = Service, x3 = <3)

*Without Laplace Smoothing:*

P(x1 | High) \* P(x2 | High) \* P(x3 | High) \* P(High) 🡪 (1/4) \* (1/4) \* (1/4) \* (4/10) = 0.625 %

*With Laplace Smoothing = 1:*

(1+1/4+2) \* (1+1/4+2) \* (1+1/4+3) \* (4/10) = (2/6) \* (2/6) \* (2/7) \* (4/10) = **1.2698 %**

**Instance 2:**

**P(Low | X)** where X = (x1 = College, x2 = Retail, x3 = <3)

*Without Laplace Smoothing:*

P(x1 | Low) \* P(x2 | Low) \* P(x3 | Low) \* P(Low) 🡪 (2/6) \* (**0**/6) \* (2/6) \* (6/10) = 0 %

*With Laplace Smoothing = 1:*

(2+1/6+2) \* (0+1/6+**3**) \* (2+1/6+3) \* (6/10) = (3/8) \* (1/9) \* (3/9) \* (6/10) = **0.83333 %**

**P(High | X)** where X = (x1 = College, x2 = Retail, x3 = <3)

*Without Laplace Smoothing:*

P(x1 | High) \* P(x2 | High) \* P(x3 | High) \* P(High) 🡪 (3/4) \* (**0**/4) \* (1/4) \* (4/10) = 0 %

*With Laplace Smoothing = 1:*

(3+1/4+2) \* (0+1/4+**3**) \* (1+1/4+3) \* (4/10) = (4/6) \* (1/7) \* (2/7) \* (4/10) = **1.088 %**

**Instance 3:**

**P(Low | X)** where X = (x1 = Graduate, x2 = Service, x3 = 3-10)

*Without Laplace Smoothing:*

P(x1 | Low) \* P(x2 | Low) \* P(x3 | Low) \* P(Low) 🡪 (0/6) \* (4/6) \* (2/6) \* (6/10) = 0 %

*With Laplace Smoothing = 1:*

(0+1/6+3) \* (4+1/6+2) \* (2+1/6+3) \* (6/10) = (1/9) \* (5/8) \* (3/9) \* (6/10) = **1.3888 %**

**P(High | X)** where X = (x1 = Graduate, x2 = Service, x3 = 3-10)

*Without Laplace Smoothing:*

P(x1 | High) \* P(x2 | High) \* P(x3 | High) \* P(High) 🡪 (0/4) \* (1/4) \* (1/4) \* (4/10) = 0 %

*With Laplace Smoothing = 1:*

(0+1/4+3) \* (1+1/4+2) \* (1+1/4+3) \* (4/10) = (1/7) \* (2/6) \* (2/7) \* (4/10) = **0.5442 %**

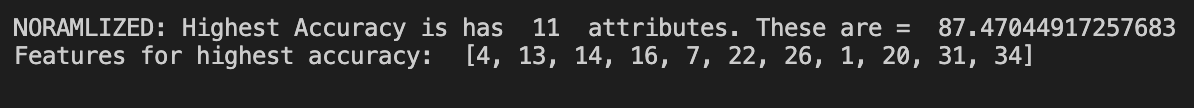
**Question 2)**

1. A close up of text on a black background

   Description automatically generatedResults: because we are interested in the magnitude of the correlation without considering direction, we take the absolute value of r. This takes the strength of the relationship, removing the issues of whether difference is in the negative or positive direction.
2. After running python3 q2.py you will get results displayed below which show accuracy results.

A close up of a screen

Description automatically generatedResults after normalizing data and results without normalizing data.

I also tried computing normalization in a different way, for which I commented the code out and you can see in the code under the z\_score function. The results I got were slightly different.

**Question 3)**

1. For question 3 run the program and the results of all selected features are printed. Also, I computed this with both calculations of z\_score.

A close up of text on a black background

Description automatically generated

A close up of text on a black background

Description automatically generatedSecond way (commented out code) of normalization code yields: