Thatcher Rickertsen

tor0002

09/23/18

Data Mining HW 02

1.
    a. Write the definition of Minkowski distance measure and 3 special cases of it.
        - **$d(i, j) = SQRT_h(|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + ... + |x_{ip} - x_{jp}|^h)$**
        - **Case 1: Manhatten Distance**
            - The difference in bits between two binary numbers.
        - **Case 2: Euclidean Distance**
            - "normal distance" / "straight line" distance
        - **Case 3: Supremum Distance**
            - The maximum distance between components of vectors.
    b. Compute the cosine similarity measurement between the following two vectors:
        - $x_1 = (4, 0, 1, 3, 5)$ | $x_2 = (-1, 2, 3, 0, 1)$
        - Formula: $\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1||*|d_2||$
        - $(x_1 \bullet x_2) = 4 * -1 + 0 * 2 + 1 * 3 + 3 * 0 + 5 * 1 = 4$
        - $||x_1|| = SQRT(4^2 + 0^2 + 1^2 + 3^2 + 5^2) = 7.14$
        - $||x_2|| = SQRT(-1^2 + 2^2 + 3^2 + 0^2 + 1^2) = 3.87$
        - $\cos(d_1, d_2) = 4 / (7.14 * 3.87) =$ **0.145**
2. Name and explain four methods to handle noisy data:
    - **Binning**
        - Partition data after sorting it according to different categories (means, medians, boundaries, etc.). Then you can smooth the data after everything is sorted to get more generalized statistics.
    - **Regression**
        - Put data in regression functions to smooth it over.
    - **Clustering**
        - Take a data set and determing outliers that will be removed.
    - **Combined Computer and Human Inspection**
        - Have an actual human look at the data as well to make sure it looks normal.
3. Name three methods for handling redundant attributes in data integration and explain how to interpret the results of them.
    - Correlation Analysis
        - **Nominal**
            - Perform a chi squared independence test, and the value you get for $x^2$ is going to represent how likely two variables are related. Even a few cells

can greatly change the result of $x^2$, which means it is often easy to test when the expected and actual values are differing.

- **Numeric**
  - Get the correlation (Pearson's product moment) coefficient using the formula. When this coeficient is positive, then the two values A and B used to calculate the coeficient are positively correlated, and vice-versa. The higher or lower the number, the more strongly or more weakly the data corresponds, respectively.

- **Covariance Analysis**
  - This is similar to numeric correlation analysis, and if Cov(A, B) is positive, it is larger than the expected value. If Cov(A, B) is negative, then A is larger than expected and B is less than expected. If Cov(A, B) = 0, then the data is independent.

4.

a. Compute the chi-squared calculation and justify it:

| | Male | Female | sum(row) |
|---|---|---|---|
| Like cats | 120(182) | 310(247) | 430 |
| Like dogs | 330(267) | 300(362) | 630 |
| sum(col.) | 450 | 610 | 1060 |

- $X^2 = (120 – 182)^2/182 + (330 – 267)^2/267 + (310 – 247)^2/247 + (300 – 362)^2/362$
- **$x^2 = 62.67$**
- **Relatively, the chi-squared value is quite low. This is because for the most part, the expected values are quite similar to the actual values.**

b. Compute the covariance matrix of the following vectors:

x1=(2.5 , 0.5 , 2.2 , 1.9 , 3.1 , 2.3)
x2=(2.3 , 0.8 , 3.0 , 2.2 , 2.5 , 2.8)

- Covariance Matrix = $\begin{matrix} Var(x1) & Cov(x2,x1) \\ Cov(x1,x2) & Var(x2) \end{matrix}$
- $Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$
- $E(x_1) = (2.5 + 0.5 + 2.2 + 1.9 + 3.1 + 2.3) / 6 = 2.08$
- $E(x_2) = (2.3 + 0.8 + 3.0 + 2.2 + 2.5 + 2.8) / 6 = 2.27$
- $Var(x_1) = (2.5^2 + 0.5^2 + 2.2^2 + 1.9^2 + 3.1^2 + 2.3^2)/6 - 2.08^2 = 0.65$
- $Var(x_2) = (2.3^2 + 0.8^2 + 3.0^2 + 2.2^2 + 2.5^2 + 2.8^2)/6 – 2.27^2 = 0.49$
- $Cov(x_1,x_2) = Cov(x_2,x_1) = (2.5 * 2.3 + 0.5 * 0.8 + 2.2 * 3.0 + 1.9 * 2.2 + 3.1 * 2.5 + 2.3 * 2.8)/ 6 – 2.08 * 2.27 = 0.47$
- **Answer =** $\begin{matrix} \mathbf{0.65} & \mathbf{0.47} \\ \mathbf{0.47} & \mathbf{0.49} \end{matrix}$

5. No answer required