Graphika - Formula Optimization Detail

Ruchit Desai

1. Pearson Correlation Coefficient

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}}$$

In this formula n is maximum column value of anonymized representations. For the smaller json file this number was 459316.

$\overline{x}$ was calculated by dividing the length of input set a (represents the number of 1's in the feature set) by n

$\overline{y}$ was calculated by dividing the length of input set b (represents the number of 1's in the feature set) by n.

2. Scenarios

Since the value of x_i and y_i can only be either 0 or 1. The above formula breaks down into the following scenarios which I have labeled a - d.

2.1. Scenario a

In this scenario $x_i = 1$ and $y_i = 1$. This yields the following formula:

$$\frac{\sum_{i=1}^{a}(1 - \overline{x})(1 - \overline{y})}{\sqrt{\sum_{i=1}^{a}(1 - \overline{x})^2(1 - \overline{y})^2}}$$

a = length of the intersection between the two input sets. The computation has an average time complexity of O(min(len(set_a), len(set_b)))

2.2. Scenario b

In this scenario $x_i = 1$ and $y_i = 0$. This yields the following formula:

$$\frac{\sum_{i=1}^{b}(1 - \overline{x})(-\overline{y})}{\sqrt{\sum_{i=1}^{b}(1 - \overline{x})^2(\overline{y})^2}}$$

b = length of the difference between set_a and set_b. The computation has an average time complexity of O(len(set_a))

2.3. Scenario c

In this scenario $x_i = 0$ and $y_i = 1$. This yields the following formula:

$$\frac{\sum_{i=1}^{c}(-\overline{x})(1 - \overline{y})}{\sqrt{\sum_{i=1}^{c}(\overline{x})^2(1 - \overline{y})^2}}$$

c = length of the difference between set_b and set_a. The computation has an average time complexity of O(len(set_b))

2.4. Scenario d

In this scenario $x_i = 0$ and $y_i = 0$. This yields the following formula:

$$\frac{\sum_{i=1}^{d} \overline{x}\overline{y}}{\sqrt{\sum_{i=1}^{d} \overline{x}^2\overline{y}^2}}$$

$a = n$ - length of the union between the two input sets. The computation has an average time complexity of O(len(set_a) + len(set_b))

3. Combine scenarios for calculation of PCC

$$n = a + b + c + d$$

$$r_{X,Y} = \frac{\sum_{i=1}^{a}(1-\overline{x})(1-\overline{y}) + \sum_{i=1}^{b}(1-\overline{x})(-\overline{y}) + \sum_{i=1}^{c}(-\overline{x})(1-\overline{y}) + \sum_{i=1}^{d} \overline{x}\overline{y}}{\sqrt{\sum_{i=1}^{a}(1-\overline{x})^2(1-\overline{y})^2} + \sqrt{\sum_{i=1}^{b}(1-\overline{x})^2(\overline{y})^2} + \sqrt{\sum_{i=1}^{c}(\overline{x})^2(1-\overline{y})^2} + \sqrt{\sum_{i=1}^{d} \overline{x}^2\overline{y}^2}}$$

4. Pearson Correlation Distance

Pearson correlation distance is defined as:

$$d_{X,Y} = 1 - r_{X,Y}$$

$$d_{X,Y} = 1 - \frac{\sum_{i=1}^{a}(1-\overline{x})(1-\overline{y}) + \sum_{i=1}^{b}(1-\overline{x})(-\overline{y}) + \sum_{i=1}^{c}(-\overline{x})(1-\overline{y}) + \sum_{i=1}^{d} \overline{x}\overline{y}}{\sqrt{\sum_{i=1}^{a}(1-\overline{x})^2(1-\overline{y})^2} + \sqrt{\sum_{i=1}^{b}(1-\overline{x})^2(\overline{y})^2} + \sqrt{\sum_{i=1}^{c}(\overline{x})^2(1-\overline{y})^2} + \sqrt{\sum_{i=1}^{d} \overline{x}^2\overline{y}^2}}$$