University of Glasgow | School of Computing Science

CSC3005 Laboratory/Tutorial 1 Solution: Data Exploration and Data Quality

3. Relationship cosine and correlation measures.
   a) What is the range of values that are possible for the cosine measure?

   Ans:
   For cosine measure, where $\cos(\mathbf{x}, \boldsymbol{y}) =< \mathbf{x}, \mathbf{y} >/||\mathbf{x}||||\boldsymbol{y}||$

   The sign of dot product $< \mathbf{x}, \mathbf{y} >$ can be negative as such the range can be from -1 to 1. There [-1,1].

   b) If two objects have a cosine measure of 1, are they identical? Explain.

   Ans:

   Not necessarily. We only know that their values of attributes differ by a constant factor

   c) What is the relationship of the cosine measure to correlation, if any?

   Ans:

   Cosine measure :

$$\cos(\mathbf{x}, \boldsymbol{y}) =< \mathbf{x}, \mathbf{y} >/||\mathbf{x}||||\boldsymbol{y}|| = \frac{\sum_{k=1}^{n} x_k y_k}{\sqrt{\sum_{k=1}^{n} x_k x_k)}\sqrt{\sum_{k=1}^{n} y_k y_k)}}$$

   Correlation:

$$\text{corr}(\mathbf{x}, \boldsymbol{y}) = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2}}$$

   Where n is number of features. If the mean is zero in which $\bar{x} = \bar{y} = 0$, $\cos(\mathbf{x}, \boldsymbol{y}) = corr(\mathbf{x}, \mathbf{y})$

d) Figure 3(a) shows the relationship of the cosine measure to Euclidean distance for 100,000 randomly generated points that have been normalized to have an L2 length of 1. What general observation can you make about the relationship between Euclidean distance and cosine similarity when vectors have an L2 norm of 1?

Ans:

Since all the 100,000 points fall on the curve, there is a functional relationship between Euclidean distance and cosine similarity for normalized data. More specifically, there is an inverse relationship between cosine similarity and Euclidean distance. For example, if two data points are identical, their cosine similarity is one and their Euclidean distance is zero, but if two data points have a high Euclidean distance, their cosine value is close to zero. Note that all the sample data points were from the positive quadrant, i.e., had only positive values. This means that all cosine values will be positive.

Cosine measure :

$$\cos(\mathbf{x}, \boldsymbol{y}) = < \mathbf{x}, \mathbf{y} > / ||\mathbf{x}|| ||\boldsymbol{y}|| = \frac{\sum_{k=1}^{n} x_k y_k}{\sqrt{\sum_{k=1}^{n} x_k x_k)} \sqrt{\sum_{k=1}^{n} y_k y_k)}}$$

Euclidean distance:

$$\mathrm{d}(\mathbf{x}, \boldsymbol{y}) = |\mathbf{x}, \mathbf{y}| = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

e) Figure 3(b) shows the relationship of correlation to Euclidean distance for 100,000 randomly generated points that have been standardized to have a mean of 0 and a standard deviation of 1. What general observation can you make about the relationship between Euclidean distance and correlation when the vectors have been standardized to have a mean of 0 and a standard deviation of 1?

Ans:

Same as in part (d).

Correlation:

$$\text{corr}(\mathbf{x}, \boldsymbol{y}) = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{\boldsymbol{k=1}}^{\boldsymbol{n}}(x_k - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{\boldsymbol{k=1}}^{\boldsymbol{n}}(y_k - \bar{y})^2}}$$

Euclidean distance:

$$\text{d}(\mathbf{x}, \boldsymbol{y}) = |\mathbf{x}, \mathbf{y}| = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

f) Derive the mathematical relationship between cosine similarity and Euclidean distance when each data object has an L2 length of 1.

Ans:

x and **y** are two vectors where each vector has an L2 length of 1. For such vectors, cosine between the two vectors is their dot product.

Cosine measure :

$$\cos(\mathbf{x}, \boldsymbol{y}) = < \mathbf{x}, \mathbf{y} > / ||\mathbf{x}|| ||\boldsymbol{y}|| = \frac{\sum_{\boldsymbol{k=1}}^{\boldsymbol{n}} x_k y_k}{\sqrt{\sum_{k=1}^{n} x_k x_k)}\sqrt{\sum_{k=1}^{n} y_k y_k)}}$$

If L2 length =1, mean $\sqrt{\sum_{k=1}^{n} x_k x_k} = 1$ and $\sqrt{\sum_{k=1}^{n} y_k y_k} = 1$, therefore

$$\cos(\mathbf{x}, \boldsymbol{y}) = \sum_{k=1}^{n} x_k y_k$$

Euclidean distance:

$$\text{D}(\mathbf{x}, \boldsymbol{y}) = |\mathbf{x}, \mathbf{y}| = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

$$= \sqrt{\sum_{k=1}^{n}(x_k^2 - 2x_k y_k + y_k^2)}$$

$$= \sqrt{\sum_{k=1}^{n} x_k^2 - \sum_{k=1}^{n} 2x_k y_k + \sum_{k=1}^{n} y_k^2}$$

$$= \sqrt{1 - 2\cos(\mathbf{x}, \mathbf{y}) + 1}$$

$$= \sqrt{2 - 2\cos(\mathbf{x}, \mathbf{y})}$$

g) Derive the mathematical relationship between correlation and Euclidean distance when each data point has been been standardized by subtracting its mean and dividing by its standard deviation.

x and **y** be two vectors where each vector has an a mean of 0 and a standard deviation of 1. For such vectors, the variance (standard deviation squared) is just $n - 1$ times the sum of its squared attribute values and the correlation between the two vectors is their dot product divide by $n - 1$

Correlation:

$$\text{corr}(\mathbf{x}, \boldsymbol{y}) = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2}}$$

Euclidean distance:

$$\text{d}(\mathbf{x}, \boldsymbol{y}) = |\mathbf{x}, \mathbf{y}| = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

Now $x_k \rightarrow \frac{x_k - \bar{x}}{s_x}$ and $\boldsymbol{y_k} \rightarrow \frac{\boldsymbol{y_k} - \bar{y}}{\boldsymbol{s_y}}$ where $s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}$

and $s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2}$

Therefore,

$$\mathrm{d}(\mathbf{x},\mathbf{y}) = |\mathbf{x},\mathbf{y}| = \sqrt{\sum_{k=1}^{n}\left(\left[\frac{x_k-\bar{x}}{s_x}\right]-\left[\frac{y_k-\bar{y}}{s_y}\right]\right)^2}$$

$$= \sqrt{\sum_{k=1}^{n}\left(\frac{x_k-\bar{x}}{s_x}\right)^2 - 2\sum_{k=1}^{n}\left(\frac{x_k-\bar{x}}{s_x}\right)^2\left(\frac{y_k-\bar{y}}{s_y}\right)^2 + \sum_{k=1}^{n}\left(\frac{y_k-\bar{y}}{s_y}\right)^2}$$

$$= \sqrt{(n-1) - 2(n-1)corr(\mathbf{x},\mathbf{y}) + (n-1)}$$

$$= \sqrt{(2n-2) - (2n-2)corr(\mathbf{x},\mathbf{y})}$$

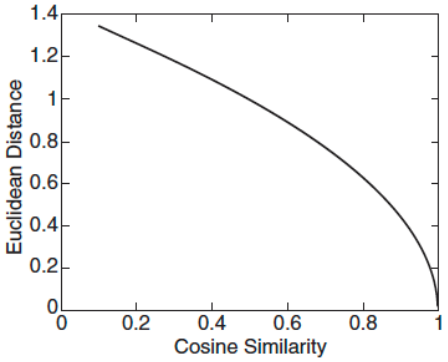

Fig 3(a) Relationship between Euclidean distance and Cosine Similarity

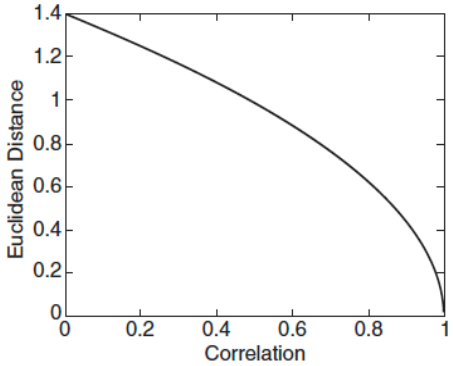

Fig 3(b) Relationship between Euclidean distance and Correlation