Similarity Measure Metrics

F1-Score: It considers both the **precision *p*** and the **recall *r*** of the test to compute the score: *p* is the number of correct positive results divided by the number of all positive results returned by the classifier, and *r* is the number of correct positive results divided by the number of all relevant samples.

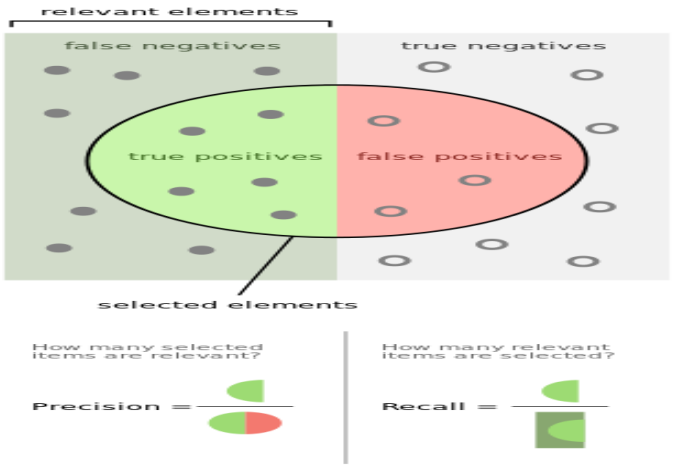
The F_1 score is the harmonic average of the precision and recall, where its best value at 1 (perfect precision and recall) and worst at 0.

Let A be the set of found items, and B the set of wanted items.

Prec=|AB|/|A|, Rec=|AB|/|B|.

Their harmonic mean, the F1-measure, is the same as the Dice coefficient:

$$F1(A,B) = \frac{2}{1/P+1/R} = \frac{2}{|A|/|AB|+|B|/|AB|}$$



$$Dice(A,B) = \frac{2|AB|}{|A|+|B|}$$

For example, to calculate the similarity between: **night and nacht**

We would find the set of bigrams in each word:

{ni,ig,gh,ht}

{na,ac,ch,ht}

Each set has four elements, and the intersection of these two sets has only one element: ht.

$s = (2 \cdot 1) / (4 + 4) = 0.25.$

The PWF measure is given by the relation between pair wise *precision* and *recall*. This relation is

$$PWF = \frac{(1+\beta^2)\, precision \times recall}{(\beta \times precision) + recall}$$

where $\beta > 0$ is a parameter used to favour either precision or recall. It is common to leave $\beta = 1$. To calculate precision and recall, the following expressions are used

$$precision = \frac{|S \cap T|}{|S|}$$

$$recall = \frac{|S \cap T|}{|T|}$$

where *S* is the set of node pairs that are assigned to the same community and *T* is the set of node pairs that have the same attribute.

Euclidean Distance

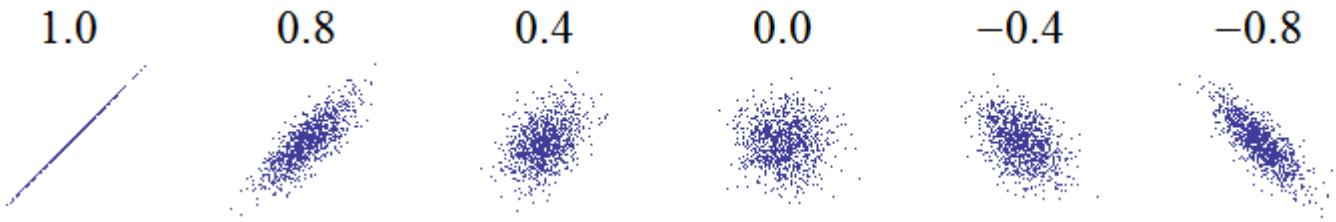
A simple yet powerful way to determine similarity is to calculate the Euclidean Distance between two data objects. To do this, we need the data objects to have numerical attributes. We also may need to normalize the attributes. For example, if we were comparing people's rankings of movies, we need to make sure that the ranking scale is the same across all people; it would be problematic to compare someone's rank of 5 on a 1-5 scale and another person's 5 on a 1-10 scale. The next step is to apply the Euclidean distance formula:

$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}.$$

We subtract each attribute in one data object from the other corresponding attribute and add them in quadrature. The result is the "distance" between the two data objects. The shorter the distance, the more similar the data objects are.

Pearson Coefficient

The Pearson Coefficient is a more complex and sophisticated approach to finding similarity. The method generates a "best fit" line between attributes in two data objects.



A line that runs through all the data points and has a positive slope represents a perfect correlation between the two objects. This best fit line is generated by the Pearson Coefficient which is the similarity score. The Pearson Coefficient is found using the following equation:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

The coefficient is found from dividing the covariance by the product of the standard deviations of the attributes of two data objects. The advantage of the Pearson Coefficient over the Euclidean Distance is that it is more robust against data that isn't normalized. For example, if one person ranked movies "a", "b", and "c" with scores of 1, 2, and 3 respectively, he would have a perfect correlation to someone who ranked the same movies with a 4, 5, and 6.

Jaccard Coefficient

When dealing with data objects that have binary attributes, it is more effective to calculate similarity using a Jaccard Coefficient. The equation to find the Jaccard Coefficient is as follows:

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}.$$

The M11 represents the total number of attributes where both data objects have a 1. The M10 and M01 represent the total number of attributes where one data object has a 1 and the other has a 0. The total matching attributes are then divided by the total non-matching attributes plus the matching ones. A perfect similarity score would then be a 1. For example, if object "A" had attributes of 1, 0, 1, and object "B" had attributes of 1, 1, 1, the Jaccard Coefficient would be 2/3. This method eliminates matching attributes that share a 0, which makes it great for sparse data sets, or ones where most of the attributes are 0's. For instance, if one were to record purchases in a grocery store by having a 1 for each item purchased and a 0 for items not purchased by a

particular customer, the customer would have a lot of 0's and only a few 1's if the whole inventory of the store was taken into account. So when one customer is compared to another, all those items that weren't purchased by either person are not factored into the Jaccard Coefficient when finding how similar the people are.

Cosine Similarity

Finding the cosine similarity between two data objects requires that both objects represent their attributes in a vector. Similarity is then measured as the angle between the two vectors.

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}.$$

This method is useful when finding the similarity between two text documents whose attributes are word frequencies. A perfect correlation will have a score of 1 (or an angle of 0) and no correlation will have a score of 0 (or an angle of 90 degrees).

Tanimoto Coefficient

The Tanimoto coefficient is an extended version of the Jaccard Coefficient and cosine similarity. It also assumes that each data object is a vector of attributes. The attributes may or may not be binary in this case. If they all are binary, the Tanimoto method reduces to the Jaccard method. The Tanimoto Coefficient is found from the following equation:

$$T(A,B) = \frac{A \cdot B}{\|A\|^2 + \|B\|^2 - A \cdot B}.$$

In the equation, A and B are data objects represented by vectors. The similarity score is the dot product of A and B divided by the squared magnitudes of A and B minus the dot product. Using the grocery store example, the Tanimoto Coefficient ensures that a customer who buys five apples and one orange will be different from a customer who buys five oranges and an apple.

Density and Entropy

Given a graph $G(V,E)$ and a partition $\mathbf{C} = \{C1, C2, \ldots, Ck\}$ of G, density is defined as:

$$d(\mathbf{C}) = \frac{1}{|E|} \sum_{Ci \in \mathbf{C}} |E(Ci)|$$

where $E(Ci)$ is the set of edges that start and finish in the ith community. That is, density represents the proportion of edges that lie within the communities and a higher density corresponds to a better clustering.

The term *entropy*, used in several different contexts to measure the degree of disorder of a complex system, indicates the heterogeneity of the elements inside a cluster according to their attribute values. It is given by

$$\mathscr{H}(\mathbf{C}) = \frac{1}{|V|} \sum_{C_i \in \mathbf{C}} H(C_i)$$

where $H(Ci)$ is the entropy of the ith community and is calculated as

$$H(C_i) = -\sum_{j=1}^{r} p_{ij} \ln p_{ij} + (1 - p_{ij}) \ln (1 - p_{ij})$$

where r is the number of attributes and $pi\ j$ is the proportion of elements in the community Ci with the same value on the attribute j. The objective of the clustering is to reduce the entropy which is equivalent to increasing the homogeneity of the partition.

In statistics, **latent variables** are variables that are not directly observed but are rather inferred (through a mathematical model) from other variables that are observed (directly measured). Mathematical models that aim to explain observed variables in terms of latent variables are called latent variable models.