Similarity Measure

Ian Shaw LTJG, USN

January 10, 2019

1 Math

Consider the following definitions.

$$\vec{w} = \text{ weight vector, } \in \mathbb{R}^{q\times 1} \tag{1}$$
$$1 = \sum_{k=0}^{q} \vec{w}_k \tag{2}$$
$$Q = \text{ Quality Matrix, } \in \mathbb{R}^{q\times n} \tag{3}$$
$$0 \leq Q \leq 1 \tag{4}$$
$$S = \text{ Similarity Matrix, } \in \mathbb{R}^{n\times n} \tag{5}$$
$$0 \leq S \leq 1 \tag{6}$$
$$q + 1 = \text{number of qualities encoded} \tag{7}$$
$$m = \text{number of persons} \tag{8}$$
$$\tag{9}$$

The element-wise the similarity score between Person i and Person j is

$$S_{ij} = 1 - (Q_i - Q_j)^2 \bullet \vec{w} = 1 - \frac{1}{q}\sum_{k=0}^{q}(Q_{k,i} - Q_{k,j})^2\vec{w}_k$$

In words:

> *The similarity of two people is the square distance of their encoded quality vector, weighted by the importance of similarity for each quality.*

1.1 Possible Alteration

Suppose a quality q is considered to be more valuable as it increases (for example, competency in some skill), without any negative consequences. In order to not negatively impact a person for being different but better in a quantitative way (but still different), the definition of similarity would need to be altered to be non-equal ($S_{ij} \neq S_{ji}$), but could take the following formulation.

$$S_{ij} = 1 - \max(Q_i - Q_j, 0) \bullet \vec{w} = 1 - \frac{1}{q}\sum_{k=0}^{q}\max(Q_k, i - Q_k, j, 0)\vec{w}_k$$

In the case that Person i is better than or equal to Person j in some metric, yet absolutely greater in at least one quality

$$S_{ij} > S_{ji}$$

2 Explanation

For a given pair of people, Person i and Person j, their similarity score S_{ij}, is a number between 0 and 1. With 0 being completely dissimilar, 1 being completely identical in terms of the encoded qualities.

The Matrix Q is the quality matrix. The entry Q_{ki} is a number between 0 and 1 indicating the strength of quality k for Person i.

The vector $\vec{w}$ is the weight vector. The entry $\vec{w}_k$ is a number between 0 and 1 indicating the importance of the quality k to determine the value of a person for the position in question.

3 Implementation

1. Hiring official determines the qualities important for their position (past jobs, competence in certain skills, etc.). These are the qualities 1 through q.
2. Hiring official determines the importance of each quality to the position on a scale of 1 (absolutely necessary) to 0 (irrelevant). These values form the vector $\vec{w}$.
3. Position seekers (the n individuals) or some authority generate the quantitative measure for each quality q, these populate the matrix Q. These could be 1/0 for yes/no (ex: have or have not attended a certain training), or some continuous scale (ex: 0.7 for 70% qualified in a particular skill). This can be done by asking the position seekers, referencing their official records, or administering some sort of evaluation.
4. Evaluate Similarity scores across the set of people.
5. Leverage similarity score information.

3.1 Use Cases of Similarity Score

1. If hiring official wants a new employee most similar to a previous one, choose a position seeker with the highest similarity score.
2. If a hiring official wants a diverse group, can create an optimization function with the minimal similarity across the team.

4 Future State

After enough records and performance metrics developed, can instead use records as data and performance metrics as labels to develop a machine learning algorithm to predict people's performance in a specific job. This could be used in concert with, or replace the need for a similarity score.