Similarity Distance

Michael H. Coen and **M. Hidayath Ansari**
Dept. of Computer Sciences
University of Wisconsin-Madison

General normalization schemes

In this section we review some popular normalization schemes. By "normalization scheme" we mean a technique to transform features so as to make comparison of examples (or sets of examples) easier. (Note that normalization is different from feature selection: feature selection schemes are intended to make different features commensurable or orthogonal, while normalization schemes are intended to make different point sets commensurable.)

Below $\mathbf{x}$ denotes a single training or test example, x_i denotes its ith feature value, and T_i denotes a transformation of the ith feature.

Linear scaling. Fix a lower bound L_i and an upper bound U_i on feature i. For example, L_i and U_i can be set to the largest and smallest values of feature i in the corpus, or they can be set so as to capture some fraction of the variance of feature i in the corpus. Define $T_i(x_i) = (\min(\max(x_i, L_i), U_i) - L_i)/(U_i - L_i)$. Then $0 \leq T_i(x_i) \leq 1$. [1]

Sample mean and variance normalization. Given a corpus $\{\mathbf{x}_j\}_{j=1}^n$ of examples, let $m_i = n^{-1} \sum_{j=1}^n x_{ji}$ denote the sample mean of the ith feature and $s_i^2 = (n-1)^{-1} \sum_{j=1}^n (x_{ji} - m_i)^2$ denote the sample variance. Define $T_i(x_i) = (x_i - m_i)/s_i$. Then the collection $\{T_i(x_{ji})\}_{j=1}^n$ has sample mean zero and sample variance 1. [2]

Rank normalization. Given a corpus $\{\mathbf{x}_j\}_{j=1}^n$ of examples, define $T_i(x_i) = n^{-1} \sum_{j=1}^n I(x_{ji} \leq x_i)$, where $I(A)$ is the indicator function of A. Then $0 \leq T_i(x_i) \leq 1$, and the collection $\{T_i(x_{ji})\}_{j=1}^n$ becomes uniformly distributed as n increases,[3] since T_i is feature i's empirical distribution function.

Gaussianization. Given a corpus $\{\mathbf{x}_j\}_{j=1}^n$ of examples, let $R_i = \sum_{j=1}^n I(x_{ji} \leq x_i)$ denote the rank of x_i in the corpus. Let $f(y) = (\sqrt{2\pi})^{-1} \exp(-y^2/2)$ denote the standard normal density. Define $T_i(x_i)$ so that the equation $(n - R_i + 1/2)/n = \int_{y=-\infty}^{T_i(x_i)} f(y)\, dy$ holds.

Distribution matching. Given a corpus $\{\mathbf{x}_j\}_{j=1}^n$ of examples, with cdf F_i and desired cdf G_i, define $T_i(x_i) = G_i^{-1}(F_i(x_i))$. If F_i is the empirical cdf and G_i is the identity function $r \mapsto r$, then distribution matching reduces to rank normalization. Claim: Distribution matching also generalizes gaussianization.

Similarity distance isn't just EMD, normalized

It is clear that similarity distance is not a simple application of any of the above normalization schemes to earth mover distance. Perhaps similarity distance could be formulated as a kind of distribution matching, but certainly not in an obvious way.

A classical rescaling method in NLP is TF-IDF. This has the same purpose as in our setting, i.e., one wants to compare two sets of documents, e.g. {query} and {corpus doc 1, ..., corpus doc n} in the information retrieval setting, but these two sets have different scaling properties, since for example the query will typically be much shorter than any document in the corpus.

Thus even if in a sense similarity distance **is** a "normalization" of EMD, this does not preclude it being very useful. TF-IDF is a "normalization" of TF but is used much more frequently than TF.

Applications of similarity distance

Information retrieval. (I know mike doesn't like NLP.)

See also

See also my summaries of papers and notes on them, in `../lit/citations.txt`.

I need to add citations above. For now, here is a citation to keep bibtex happy: (Stolcke, Kajarekar, and Ferrer 2008).

References

Stolcke, A.; Kajarekar, S.; and Ferrer, L. 2008. Nonparametric feature normalization for svm-based speaker verification. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 1577–1580.

[1] Claim: Under some circumstances, $T_i(x_i)$ will be uniformly distributed.

[2] Claim: Assumes gaussian-distributed data. Note that it is possible to perform just sample mean normalization ($x_i - m_i$) or sample variance normalization (x_i/s_i).

[3] Right?