A short note on cosine preprocessing

Nathalie Villa-Vialaneix

March 26, 2014

1 Notations

In the following, Δ will denote a $n \times n$ dissimilarity matrix with $\delta_{ij} = \delta_{ji}$ and $\delta_{ii} = 0$ and $\delta(x_i, x_j) = \delta_{ij}$ for individuals x_i and x_j living in an abstract space $\mathcal{G}$.

2 Original cosine pre-processing

For the original data, the cosine preprocessing is used as follow:

1. at first, the dissimilarity matrix is doubled centered:

$$s_{ij} = -\frac{1}{2}\left[\delta_{ij} - \frac{1}{n}\sum_k (\delta_{ik} + \delta_{kj}) + \frac{1}{n^2}\sum_{k,k'} \delta_{kk'}\right],$$

   (see [Lee and Verleysen, 2007]). When the dissimilarity matrix is Euclidean, this produces a positive definite and symmetric similarity matrix which is thus a kernel;

2. then, standard cosine preprocessing [Ben-Hur and Weston, 2010] can thus be applied to (s_{ij}):

$$\tilde{s}_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}};$$

3. and finally, the scaled similarity matrix is turned back into a dissimilarity using the standard distance computation:

$$\tilde{\delta}_{ij} = \tilde{s}_{ii} + \tilde{s}_{jj} - 2\tilde{s}_{ij} = 2 - 2\tilde{s}_{ij}.$$

A few things has to be noted:

- in this case, the dissimilarity δ is assimilated with a squared distance, as shown by the last equation which corresponds to the computation in the implicite reproducing kernel Hilbert space associated with $\tilde{s}$ to $\|\phi(x_i) - \phi(x_j)\|^2$ (where ϕ is the feature map);
- when $\tilde{s}_{ij} = s_{ij}$, $\tilde{\delta}_{ij} = \delta_{ij}$.

3 Propagating cosine pre-processing to new data

Suppose now that new data are to be processed by the algorithm. In the following, the new data will be denoted by x_{n+1} and only $\delta_{n+1,i}$, for $i = 1, \ldots, n$ are known (and necessary) to predict the class of the new x_{n+1}. However, the cosine pre-process is thus a bit more complicated to define. The steps described above are propagated to the new data:

1. the dissimilarity is turned into a similarity using

$$s_{n+1,i} = -\frac{1}{2}\left[\delta_{n+1,i} - \frac{1}{n}\sum_{k=1}^{n} \delta_{ik} - \frac{1}{n}\sum_{k=1}^{n} \delta_{n+1,k} + \frac{1}{n^2}\sum_{k,k'=1}^{n} \delta_{kk'}\right].$$

   Note that when using this transformation with one row of the original matrix, the same similarity as in the previous section is recovered;

2. similarly, the data are applied a cosine process. As $\delta(x_{n+1}, x_{n+1}) = 0$ and thus, following the previous case, the auto-similarity of x_{n+1} can be estimated by:

$$s_{n+1,n+1} = \frac{1}{n}\sum_{k=1}^{n}\delta_{n+1,k} - \frac{1}{2n^2}\sum_{k,k'=1}^{n}\delta_{kk'}.$$

Note that, once again, this transformation provides the same auto-similarity that original cosine pre-processing when used with one of the original sample. The cosine preprocess thus gives

$$\tilde{s}_{n+1,i} = \frac{s_{n+1,i}}{\sqrt{s_{ii}}\sqrt{s_{n+1,n+1}}}.$$

3. finally, the data is turned back into a similarity using the standard

$$\tilde{\delta}_{n+1,i} = 2 - 2\tilde{s}_{n+1,i}.$$

References

[Ben-Hur and Weston, 2010] Ben-Hur, A. and Weston, J. (2010). *Data Mining Techniques for the Life Sciences*, volume 609 of *Methods in Molecular Biology*, chapter A user's guide to support vector machine, pages 223–239. Springer-Verlag.

[Lee and Verleysen, 2007] Lee, J. and Verleysen, M. (2007). *Nonlinear Dimensionality Reduction*. Information Science and Statistics. Springer, New York; London.