T-61.5020 Statistical Natural Language Processing

Answers 2 — Similarity measures
Version 1.0

1. **Euclidean distance (L_2 norm)**

   Euclidean distance between the vectors $\mathbf{x} = [x_1\ x_2 \ldots x_n]$ and $\mathbf{y} = [y_1\ y_2 \ldots y_n]$ is defined as

$$Euc(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

   The distance between Tintus and Koskisen korvalääke is calculated as an example:

$$\begin{aligned} Euc(Ti, Ko) &= \sqrt{(0-10)^2 + (0-6)^2 + (5-2)^2 + (1-1)^2 + (4-0)^2} \\ &= 12.7 \\ Euc(Ko, Te) &= 9.9 \\ Euc(Ti, Te) &= 5.1 \end{aligned}$$

   L_1 norm

   The distance according to the L_1 norm is defined as

$$L_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i| \tag{2}$$

   So the distances are:

$$\begin{aligned} L_1(Ti, Ko) &= |0-10| + |0-6| + |5-2| + |1-1| + |4-0| \\ &= 23.0 \\ L_1(Ko, Te) &= 17.0 \\ L_1(Ti, Te) &= 10.0 \end{aligned}$$

   Cosine

   The cosine measure is a little different case. It can be defined as

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{3}$$

	fresh	acidic	sweet	fruity	soft
Tintus	0	0	0.50	0.10	0.40
Korvalääke	0.53	0.32	0.11	0.05	0
Termiitti	0.07	0.29	0.21	0.21	0.21

Table 1: ML estimates for the word probabilities

Let's calculate the distances:

$$\begin{aligned} cos(Ti, Ko) &= \frac{0 \cdot 10 + 0 \cdot 6 + 5 \cdot 2 + 1 \cdot 1 + 4 \cdot 0}{\sqrt{5^2 + 1 + 4^2}\sqrt{10^2 + 6^2 + 2^2 + 1^2}} \\ &= 0.14 \\ cos(Ko, Te) &= 0.55 \\ cos(Ti, Te) &= 0.70 \end{aligned}$$

Here a larger value corresponds to a larger similarity, so the distances are in the same order as before.

Information radius

For the information radius we formulate the maximum likelihood estimates for that the next word is generated by a source l_i (Tintus, Korvalääke, Termiitti) is w_i. This is done by dividing the each element of the table by the sum of its row (Table 1). Last we define that

$$0 \log \frac{0}{x} = 0, \ \forall x \in \Re.$$

The information radius is given by the formula

$$\begin{aligned} Irad(p, q) &= D(p || \frac{p+q}{2}) + D(q || \frac{p+q}{2}) \\ &= \sum_i p_i \log \frac{p_i}{\frac{p_i + q_i}{2}} + \sum_i q_i \log \frac{q_i}{\frac{p_i + q_i}{2}} \end{aligned}$$

Let's calculate it for the given sources:

$$\begin{aligned} Irad(Ti, Ko) &= 0 \cdot \log \frac{2 \cdot 0}{0.53} + 0 \cdot \log \frac{2 \cdot 0}{0.32} + 0.50 \cdot \log \frac{2 \cdot 0.50}{0.61} + 0.10 \cdot \log \frac{2 \cdot 0.10}{0.15} \\ &\quad + 0.40 \cdot \log \frac{2 \cdot 0.40}{0.40} + 0.53 \cdot \log \frac{2 \cdot 0.53}{0.53} + 0.32 \cdot \log \frac{2 \cdot 0.32}{0.32} \\ &\quad + 0.11 \cdot \log \frac{2 \cdot 0.11}{0.61} + 0.05 \cdot \log \frac{2 \cdot 0.05}{0.15} + 0 \cdot \log \frac{2 \cdot 0}{0.40} \\ &= 1.5 \\ Irad(Ko, Te) &= 0.6 \\ Irad(Ti, Te) &= 0.5 \end{aligned}$$

We see that all the measures set the medicines to a similar order: Tintus and Temiitti are the most similar ones, Tintus and Korvalääke are the most different.

Kullback-Leibler divergence

From the definition of the KL divergence we can directly see some of its problems:

$$D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$$

First, the KL divergence is not symmetric, so we should each time decide which one of the two drugs is the reference drug p. The second problem is that if the compared distribution has a zero probability in some dimension where the reference distribution has a non-zero probability, the divergence goes to infinity.

2. **Kullback-Leibler divergence**

The definition of the Kullback-Leibler divergence was

$$D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$$

Let's find a distribution that minimizes the KL divergence. We add a Lagrange coefficient λ_1 to make sure that p shall be a correct probability distribution (i.e. $\sum_i p_i = 1$) and λ_2 for q.

$$E = D(p||q) + \lambda(1 - \sum_i p_i) = \sum_i p_i \log \frac{p_i}{q_i} + \lambda_1(1 - \sum_i p_i) + \lambda_2(1 - \sum_i q_i)$$

Let's set the partial derivative with respect to the p_i to zero:

$$\begin{aligned}\frac{\partial E}{\partial p_i} &= p_i \cdot \frac{1}{\frac{p_i}{q_i}} \cdot \frac{1}{q_i} + \log \frac{p_i}{q_i} - \lambda_1 \\ &= \log p_i - \log q_i + 1 - \lambda_1 = 0\end{aligned}$$

Now we solve p_i:

$$p_i = q_i \cdot e^{\lambda_1 - 1}$$

Let's calculate the partial derivative with respect to λ_1:

$$\begin{aligned}\frac{\partial E}{\partial \lambda_1} &= 1 - \sum_i p_i = 0 \\ \Rightarrow \sum_i p_i &= 1\end{aligned}$$

A similar condition is obtained for q_i when derivating with respect to λ_2 (which was exactly the purpose of the multipliers). The last condition is obtained by derivating with respect to q_i:

$$\begin{aligned}
\frac{\partial E}{\partial q_i} &= p_i \cdot \frac{1}{\frac{p_i}{q_i}} \cdot p_i \cdot (-\frac{1}{q_i^2}) - \lambda_2 = -\frac{p_i}{q_i} - \lambda_2 = 0 \\
\Leftrightarrow p_i &= -\lambda_2 q_i
\end{aligned}$$

Because both q and p should sum up to one, we get:

$$\begin{aligned}
1 &= \sum_i p_i = \sum_i (-\lambda_2 q_i) = -\lambda_2 \sum_i q_i = -\lambda_2 \\
\Rightarrow p_i &= -\lambda_2 q_i = q_i
\end{aligned}$$

Considering the second order derivates we can make sure that this is really the minimum and not maximum:

$$\begin{aligned}
\frac{\partial^2 E}{\partial p_i \partial p_i} &= \frac{1}{p_i} > 0 \\
\frac{\partial^2 E}{\partial q_i \partial q_i} &= \frac{p_i}{q_i^2} > 0 \\
\frac{\partial^2 E}{\partial p_i \partial p_j} = \frac{\partial^2 E}{\partial q_i \partial q_j} &= 0
\end{aligned}$$

If we set $q_i = p_i$ to the formula of KL divergence we get the divergence of zero. So *KL divergence is zero if and only if the distributions q and p are equal, otherwise greater than zero.*

Information radius

The definition of the information radius is

$$IRad(p,q) = D(p||\frac{p+q}{2}) + D(q||\frac{p+q}{2})$$

We just calculated that the KL divergence is zero if the distributions are same, and larger than zero if not. In the case of the information radius, the zero divergence is also obtained if and only if $q_i = p_i$:

$$IRad(p,q) = \sum_i p_i \log \frac{p_i}{\frac{p_i+p_i}{2}} + \sum_i p_i \log \frac{p_i}{\frac{p_i+p_i}{2}} = 0$$

So the condition is the same as before.

L_1 norm

Definition of the L_1 norm is

$$L_1(p,q) = \sum_i |p_i - q_i|$$

Clearly the smallest value is zero, which comes only if $q_i = p_i$.

To conclude, we notice that all the measures give zero distance with the same condition: The distributions must be equal.

3. **Kullback-Leibler -divergence**

Let's look at the definition once more:

$$D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$$

We can see that if $q_i = 0$ when $p_i \neq 0$ we get the distance ∞.

Information radius

Let's write the definition of information radius open:

$$IRad(p,q) = D(p||\frac{p+q}{2}) + D(q||\frac{p+q}{2}) = \sum_i p_i \log \frac{2p_i}{p_i+q_i} + \sum_i q_i \log \frac{2q_i}{p_i+q_i}$$

With intuition we might guess that a suitable distribution would be one where the distributions are in completely separate areas:

$$\begin{aligned} &\text{if } p_i > 0 \Rightarrow q_i = 0 \\ &\text{if } q_i > 0 \Rightarrow p_i = 0 \end{aligned}$$

Let's insert these to the equation:

$$\begin{aligned} IRad(p,q) &= \sum_i p_i \log \frac{2p_i}{p_i} + \sum_i q_i \log \frac{2q_i}{q_i} \\ &= \log 2 \sum_i p_i + \log 2 \sum_i q_i = 2 \log 2 \end{aligned}$$

We knew that this was the largest distance. To prove that it really is, and that the guessed conditions are required to get it, would be somewhat more diffcult.

L_1 norm

The definition for the L_1 norm was

$$L_1(p,q) = \sum_i |p_i - q_i|$$

With intuition we could say that the answer is the same as with information radius, but let's try to prove it more mathematically. We separate the elementary events I to two sets. In set $j \in I$ were have the cases where $p_j > q_j$ and in set $k \in I$ the cases where $q_k > p_k$. Using these,

$$\begin{aligned} L_1(p,q) &= \sum_j (p_j - q_j) + \sum_k (q_k - p_k) \\ &= \sum_j p_j - \sum_k p_k + \sum_k q_k - \sum_j q_j \end{aligned}$$

As the probabilities are positive and sum up to one, the largest distance is get when

$$\begin{aligned} &\text{if } p_i > 0 \Rightarrow q_i = 0 \\ &\text{if } q_i > 0 \Rightarrow p_i = 0 \end{aligned}$$

so the distance is

$$L_1(p,q) \;=\; \sum_i p_i + \sum_i q_i = 2$$

Conclusions

For both information radius and L_1 norm, the same conditions for the distributions are required to get the largest distance. The KL divergence, however, goes to infinity already when the distribution q is zero somewhere where the reference distribution p is not.