

Dialogue Classifying

Stefan Hillmann

November 29, 2013

1 Measures

1.1 Cosine Distance

$$cs(P||Q) = \frac{\sum_{i=1}^n p_i q_i}{\sqrt{\sum_{i=1}^n p_i^2 \sum_{i=1}^n q_i^2}} \quad (1)$$

$$cd(P||Q) = 1 - cs(P, Q) \quad (2)$$

1.2 Kullback-Leibler Divergence

Pre-condition: $\sum_{i=1}^n p_i = 1$ and $\sum_{i=1}^n q_i = 1$

$$kl(P||Q) = \sum_{i=1}^n p_i \ln \left(\frac{p_i}{q_i} \right) \quad (3)$$

1.3 Mean Kullback-Leibler Distance

Pre-condition: $\sum_{i=1}^n p_i = 1$ and $\sum_{i=1}^n q_i = 1$

$$mkl(P||Q) = \frac{kl(P||Q) + kl(Q||P)}{2} \quad (4)$$

1.4 Symmetric Kullback-Leibler Distance

Pre-condition: $\sum_{i=1}^n p_i = 1$ and $\sum_{i=1}^n q_i = 1$

$$skl(P||Q) = \sum_{i=1}^n (p_i - q_i) \ln \left(\frac{p_i}{q_i} \right) \quad (5)$$

$$skl(P||Q) = skl(Q||P) \quad (6)$$

Proof.

$$\begin{aligned}
skl(P||Q) &= \sum_{i=1}^n (p_i - q_i) \ln \left(\frac{p_i}{q_i} \right) \\
&= \sum_{i=1}^n (p_i - q_i) (\ln(p_i) - \ln(q_i)) \\
&= \sum_{i=1}^n - (q_i - p_i) (\ln(p_i) - \ln(q_i)) \\
&= \sum_{i=1}^n (q_i - p_i) (\ln(q_i) - \ln(p_i)) \\
&= \sum_{i=1}^n (q_i - p_i) \ln \left(\frac{q_i}{p_i} \right) \\
&= skl(Q||P)
\end{aligned}$$

□

1.5 Jensen Difference Divergence

Pre-condition: $\sum_{i=1}^n p_i = 1$ and $\sum_{i=1}^n q_i = 1$

$$j(P||Q) = \sum_{i=1}^n \frac{p_i \ln(p_i) + q_i \ln(q_i)}{2} - \frac{p_i + q_i}{2} \ln \left(\frac{p_i + q_i}{2} \right) \quad (7)$$

2 N-gram model

2.1 Additive Smoothing

$$p_\lambda(x_i) = \frac{|x_i| + \lambda}{|X| + \lambda N} \quad (8)$$

In Equation 8 is $p_\lambda(x_i)$ the probability of n-gram x_i in model m . $|x_i|$ is the number of occurrences of x_i in m and $|X|$ the absolute number of all n-grams in m . Finally, N represents the number of *unique* n-grams in m .