Kshitija Shelar | Text Analytics COMP4760 | Date: 03/11/20

20200155 | Practical 10

1. Kullback-Leibler (K-L) Divergence:

KL divergence calculates the measure of divergence of one probability distribution from another. It is used to measure dissimilarity between 2 probability distributions over the same variable x. [3]

The KL divergence between two distributions Q and P is often stated using the following notation:

- KL (P || Q) Here || is divergence

Calculating KL Divergence

$$KL_{divergence}(P||Q) = H(p,q) - H(p)$$
$$= -\sum_i p_i \log(q_i) + \sum_i p_i \log(p_i)$$
$$= \sum_i p_i \log(p_i/q_i)$$

"Here, H(p,q) is cross-entropy and H(p) is system entropy and pi is actual probability and qi is the estimated probability". [1]

Cross entropy: "Cross-entropy is a measure of the difference between two probability distributions for a given random variable or set of events" [2]
Entropy: For any variable x, entropy is the average uncertainty that is in x's all possible outcomes.

KL Divergence measures the average of extra information required to represent a message using Q instead of P. [2]

Calculating KL divergence on the following 2 documents.

```
d1 = """mary was fine john fell down harry fell as-well down by the stream the sun shone before it went down"""
d2 = """belinda was ill bill fell down jeff fell too down by the river the sun shone until it sunk down"""
```

Since these documents have many words in common the score of their probability distribution is also pretty similar. The common words are fell down the sun shone etc.

```
KL-divergence between d1 and d2: 3.2494321222681 56
KL-divergence between d2 and d1: 3.0478297683519773
```

The 3rd document is very different from the first 2 sentences. Doc 3 is as follows:

```
d3 = """A sentence is nothing but an amalgamation of words put together in
 any language to effectively communicate a message.."""
```

Therefore, the KL divergence score for Doc1, Doc2, and Doc3 would be very high. It is so because there are no common words between doc1 and doc2 with doc3.

```
KL-divergence between d1 and d2: 3.2494321222681 56
KL-divergence between d2 and d1: 3.0478297683519773
KL-divergence between d1 and d3: 7.177380625673269
KL-divergence between d2 and d3: 7.1946639445463 98
```

The higher the score, the more divergent are the probability distributions. Therefore, these score makes sense as Doc1 and Doc2 have similar words hence KL score is less and Doc1 and Doc3 are dissimilar and hence KL score is higher.

Parameters

Epsilon: when 2 documents have no words in common, the probability of it will be 0 and its log will be undefined/infinity. To avoid this scenario, we set a minimum threshold/value instead of 0. This value is usually never greater than the probability of a word that occurs the least number of times.

Gamma: To account for the epsilon, gamma is defined as a normalization coefficient so that the property of probability is satisfied. [4]

References:

[1] https://towardsdatascience.com/part-2-a-new-tool-to-your-toolkit-kl-divergence-736c134baa3d

[2] https://machinelearningmastery.com/cross-entropy-for-machine-learning/#:~:text=Cross%2Dentropy%20is%20a%20measure%20of%20the%20difference%20between%20two,encode%20and%20transmit%20an%20event.

[3] http://hanj.cs.illinois.edu/cs412/bk3/KL-divergence.pdf

[4] https://stackoverflow.com/questions/58895873/what-is-the-role-of-gamma-and-epsilon-in-the-calculation-of-k-l-divergence