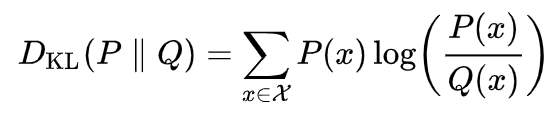
**Kullback-Leibler Divergence for Machine Learning**

If you’re diving deep into deep learning or machining a better grasp of machine learning then an understanding of the Kullback-Leibler divergence can be invaluable. Another name for this measure is relative entropy, which might give those with a physics or statistics background a hint about what is actually being measured.

In layman’s terms, the K-L divergence is a measure of how different a specific probability distribution is from a reference distribution. However, it is not a true statistical metric like variation of information which measures the distance between two clusterings.

Please don’t be put off by the following mathematics, for those who understand the notation, the following is the definition of the K-L divergence for two discrete probability distributions *P* and *Q* within the space *χ.*



formula for the K-L divergence

where D(K-L) is the divergence of *Q* from *P.* Again in layman’s terms, this can be seen as the expectation of the logarithmic difference between *P* and *Q.* If *P* and *Q* are absolutely continuously distributed over the entire space *χ,* then the above formula’s summation is replaced with an integral over the entire space. It must be noted that the log is set to base 2 when the information is measured in bits or to base *e* when measured in nats.

The K-L divergence is an important feature in a variety of machine learning models. One in particular is the Variational Autoencoder (VAE). This article assumes some background in machine learning. Briefly, a VAE is a generative model that uses a similar architecture to a standard auto encoder but adds a sampled vector which is decoded and is then analyzed by the loss function.

In a VAE, there are two components to the loss function: the reconstruction term and the regularization term. The reconstruction term serves as a measure of the efficacy of the encoder-decoder with respect to the initial data and takes place in the output layer. With 0 reconstruction loss, an autoencoder would perfectly reconstruct the input data. This would indicate extreme overfitting and a lack of interpretable latent features.

VAEs encode their inputs as normal (Gaussian) distributions rather than points. This is where the K-L divergence comes in. It is optimal for the distributions of the VAE to be regularized to increase the amount of overlap within the latent space. K-L divergence measures this and is added into the loss function. There is a tradeoff between reconstruction and regularization. If we want to reduce our reconstruction error, this comes at the expense of K-L divergence or regularization.

Another example of the application of K-L divergence in machine learning is in t-distributed Stochastic Neighbor Embedding, which is a dimensionality reduction technique which uses probabilistic methods to model complex non-linearities in data whilst also reducing the number of dimensions.