LEVERAGING KULLBACK-DIVERGENCE LOSS FOR ENHANCED MACHINE LEARNING PERFORMANCE

MATH 451

Group V

Team member:
Apekshya Shrestha
Abhay Sharma
Bishownath Raut

February 12, 2024

Plan of Presentation

Problem Statement

1. To address the challenge by exploring the implementation of Kullback-Leibler Divergence loss, a metric traditionally used for measuring the dissimilarity between probability distributions.
2. KL divergence is asymmetric which limits its applicability in some cases. Developing a symmetric version of KL divergence could make it more broadly usable as a true statistical distance measure.

Objectives

The main objective of our projects is to compare two probability distributions - a true distribution P and an approximation Q [1].

1. Investigate the mathematical foundation of KL Divergence and its significance in measuring divergence between probability distributions.
2. Examine the diverse applications of KL Divergence in machine learning, including classification, generative modeling, regularization, and information retrieval.
3. Analyze the impact of incorporating KL Divergence loss in machine learning models and how it influences model convergence, generalization, and accuracy.
4. Explore methodologies for implementing KL Divergence loss in various machine learning frameworks and architectures.

Introduction

Historical Development

First introduced in 1951 by Solomon Kullback and Richard Leibler in their paper "On Information and Sufficiency"[5].It is also referred to as Kullback-Leibler divergence or relative entropy.

Some key related works:

1. In 1960, Fréchet further developed the mathematical foundations of KL divergence and statistical manifolds[2].
2. In 1975, Jeffreys divergence was introduced as a symmetric version of KL divergence[4].
3. In the 1990s and 2000s, KL divergence became widely used in machine learning for training generative models, clustering, dimensionality reduction, and more[3].

Kullback Divergence

It measures how one probability distribution diverges from a second, expected probability distribution.

Probability Distribution

a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment.

Entropy

entropy is a measure of the uncertainty in a random variable. The entropy of a discrete random variable X with probability mass function p(x) is defined as:

$$(X) = -\sum[p(x)log(p(x))$$

[5]where the sum is over all possible outcomes of X, and log is the natural logarithm.

The formula for KL Divergence for discrete probability distributions P and Q is defined as:

$$D_{\mathrm{KL}}(P||Q) = \sum P(X) \log(\frac{P(X)}{Q(X)})$$

[5] where the sum is over all possible outcomes.

Properties of KL divergence

1. Non-negativity:

   $$D_{\mathrm{KL}}(P||Q)$$

   is always greater than or equal to 0.
2. Not Symmetric:

   $D_{\mathrm{KL}}(P||Q)$ is not same as $D_{\mathrm{KL}}(Q||P)$

Methodology

We considered the dataset of 100 worms[6].

Model with a uniform distribution

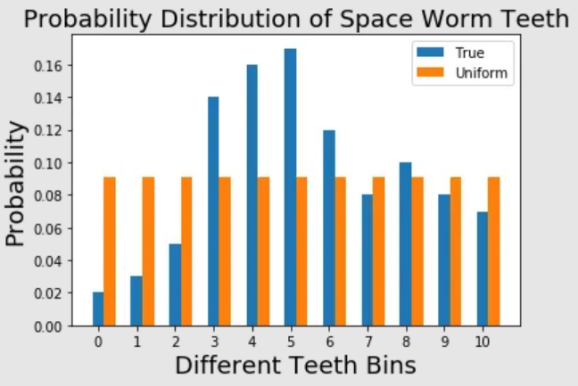

Figure 1: uniform distribution and true distribution

$$p_{uniform} = 1/totalevents = 1/11 = 0.0909$$

This is what the uniform distribution and the true distribution side-by-side looks like.

Model with a binomial distribution

This is what a comparison between the true distribution and the binomial distribution looks like

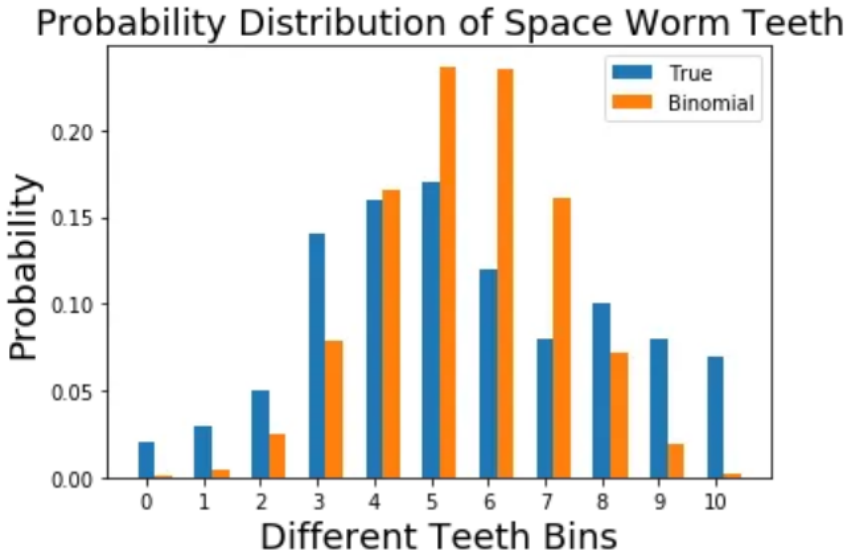

Figure 2: true distribution vs binomial distribution

Kullback Divergence

Computing the KL divergence for each of the approximate distributions
For the uniform distribution.

$$D_{\mathrm{KL}}(True||Uniform) = 0.02(0.02/0.0909) + 0.03(0.03/0.0909) +$$

$$.. + 0.07(0.07/0.0909)$$

$$D_{\mathrm{KL}}(True||Uniform) = 0.136$$

Now for the binomial distribution we get,

$$D_{\mathrm{KL}}(True||Binomial) = 0.02 * log(0.02/0.0909) + 0.03 * log(0.03/0.0909) + .$$

$$.. + 0.07 * log(0.07/0.0909)$$

$$D_{\mathrm{KL}}(True||Binomial) = 0.427$$

Conclusion

Dataset: CIFAR10

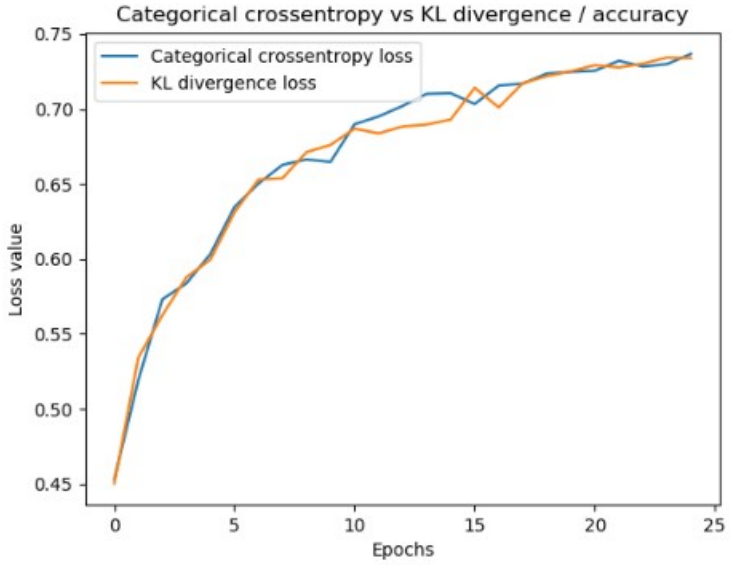

Figure 3: Categorical crossentropy vs KL divergence/accuracy

Conclusion

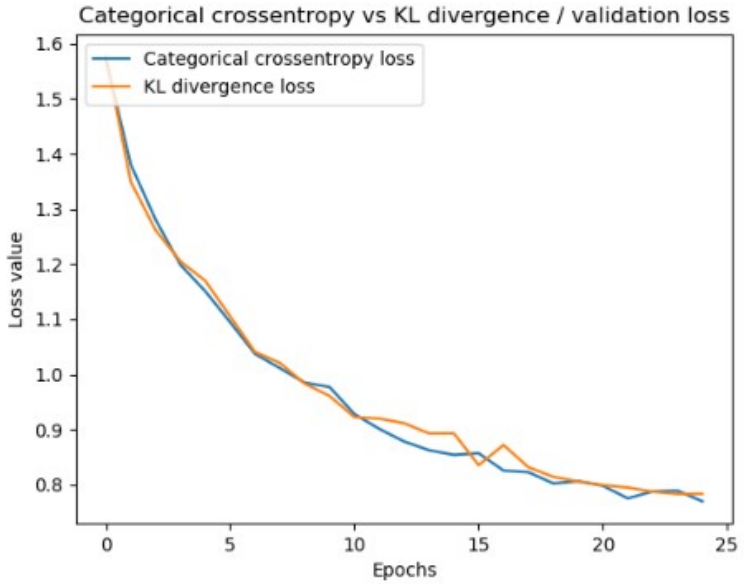

Figure 4: Categorical crossentropy vs KL divergence/loss

In 25 epochs, performance is very similar on CIFAR10 dataset in Keras.

Future Scope

1. Explore further mathematical properties and implications of KL divergence loss in different ML models, such as neural networks, decision trees, and ensemble methods.
2. Explore applications of KL divergence loss in semi-supervised and unsupervised learning scenarios, where it can encourage meaningful representations and clustering structures.

References I

- Kenneth P Burnham and David R Anderson, *Model selection and multimodel inference*, A practical information-theoretic approach **2** (2004).
- Maurice Fréchet, *Les éléments aléatoires de nature quelconque dans un espace distancié*, Annales de l'institut Henri Poincaré, vol. 10, 1948, pp. 215–310.
- John R Hershey and Peder A Olsen, *Approximating the kullback leibler divergence between gaussian mixture models*, 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, vol. 4, IEEE, 2007, pp. IV–317.
- Harold Jeffreys, *An invariant form for the prior probability in estimation problems*, Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences **186** (1946), no. 1007, 453–461.

References II

- Solomon Kullback and Richard A Leibler, *On information and sufficiency*, The annals of mathematical statistics **22** (1951), no. 1, 79–86.
- Jay Patel, *Light on math: Machine learning intuitive guide to understanding kl divergence*, 2019, Towards Data Science.

Thank you! Any query?