Coverage Expansions for Rejective Negatives via Concentration Feedback for Word Vectors.

Abstract

[TA] Malumat biriktirelim burada.

1. Preliminaries

2. Method

Art 1. Implicit Modeling. The learning of the high dimensional word embeddings using negative sampling is based on the idea that positive samples generated by model density should be discriminated from negative samples generated by a sampling distribution. For learning high dimensional densities, there exists a lot of degrees of freedom to determine this negative sampling distribution. When explicit modelling of the distribution is employed criterion such as Maximum Entropy(**?**) or Mixture Modeling ().

Art 2. Sampling Coverage. One must ensure that the generated negative samples provides enough informativeness to discriminate between data and negative samples. Here, we introduce the concept of *sampling coverage*, which is a measure of the outer coverage of the data generating density **[TA] Outer coverage who cares?**. Consider a toy scenario where data is being generated from a $\mathbb{R}^2$ Gaussian density as illustrated in Figure **??**. In the idealistic case in which we are allowed to sample infinite number of negative samples, the sampling coverage is noticeably large. In Figure **??**b, we illustrate what happens in typical word embedding negative sampling. Generated negative samples usually only covers a limited region of the sample space. This happens frequently when we sample only few negatives, and also when the representation of the context vector is not very-well established [1].

[TA] Polyhedron Volume – Inner Product Density.

Art 3. Negative Concentration. While sampling coverage helps us to understand when negative sampling is not informative, it is difficult to precisely measure the sampling coverage since embeddings usually lie in high dimensional space. We rather switch from the joint $(\boldsymbol{w}, \boldsymbol{c})$ space to the cosine angle space which is special inner product space, where each sample is equipped with a cosine product $\boldsymbol{w}^T\boldsymbol{c}$.

We then take embeddings for multiple random contexts and randomly sample a large collection of negative samples using a standard negative sampling technique in (**?**). We then depict the distribution of the cosine products in Figure **??**. This distribution exhibits a Gaussian concentration for the inner products of negative samples. The amount of concentration is denser as we increase the negative sample size. Although we can not directly measure the sampling coverage, operating in the inner product space can give us a clue on the dispersion of these negative sample subsets.

Art 4. Small Size Concentration Bias. When the sample size is small, we are much more likely to lose this well-concentration property, the negative subset do not necessarily concentrates around the particular mean. We show some problematic subsets in Figure **??**. For all these particular sampling instances, the subset has a significant mean discrepancy to the large sample sizes. This is a problematic case for achieving a large sampling coverage with negative samples.

[1]

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.