PML Wildcard

Milana M. Wolff

May 08, 2024

1 tSNE

t-distributed stochastic neighbor embedding enables users to visualize high-dimensional datasets by reducing the dimensionality of the underlying dataset. t-SNE accomplishes this by converting the similarities between data points to joint probabilities and attempting to minimizing the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. However, different initializations of t-SNE yield different results. Since t-SNE relies on a number of hyperparameters, including perplexity, exaggeration, and specification of distance metrics, and since t-SNE appears to be formulated as any other optimization or minimization problem, it seemed a suitable candidate for hyperparameter optimization.

In the wildcard project, I used the wine quality dataset, which contains 11 features and therefore cannot be immediately visualized in lower dimensions without the application of some sort of dimensionality reduction. I attempted to employ a Bayesian optimization approach. However, tSNE is implemented in `scikit-learn` as a manifold, not an estimator, and cannot be used directly within the standard `BayesSearchCV()` function. To account for this, I wrote a custom scoring function directly calculating the Kullback-Leibler divergence.

However, since t-SNE does not directly predict anything, using many of the functions designed for optimizing estimators resulted in errors despite my attempts to account for the differences in the applications of different algorithm types using custom scoring.

2 Other Projects

I also attempted to write a neural netwrok from scratch to familiarize myself with the details of doing so and included the overfitting exercise from the previous semester in the repository.