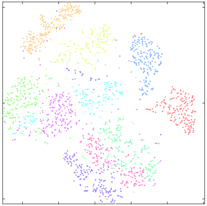
Learning Interpretable, Tree-Based Projection Mappings for Nonlinear Embeddings

Arman Zharmagambetov and Miguel Á. Carreira-Perpiñán

Dept. of Computer Science & Engineering
University of California, Merced

AISTATS 2022

Overview



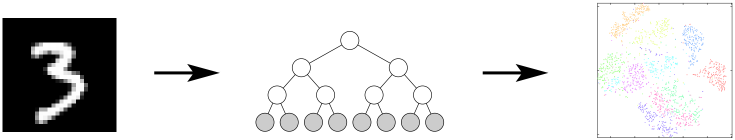
- Nonlinear embeddings (NLE), such as t-SNE, are widely used DR methods.
- Recall that such DR methods do not naturally define an out-of-sample mapping, rather they directly learn a low-dimensional projection for each training point.
- We consider the problem of learning interpretable out-of-sample mappings for NLE.

Overview

Why interpreting a projection mapping matters?

- Low-dimensional embeddings may not be a faithful projection of the original, high-dimensional data:
  1. The result depends in an obscure way on the objective function and on hyperparameters;
  2. The resulting embeddings may give a misleading view of the data, e.g. t-SNE has a strong tendency to find clusters where none exist [1];
- Augmenting the t-SNE embedding with an interpretable out-of-sample mapping allows one to understand how the high-dimensional input instances are projected to the embedding and understand whether that makes sense.

What mapping should we use?



- We argue for the use of sparse oblique decision trees as an out-of-sample mapping;
- Trees are considered to be interpretable models;
- Sparse oblique trees strike a good tradeoff between accuracy and interpretability which can be controlled via a hyperparameter.
- They can make full use of any and all features of an instance.

Jointly learning an optimal tree and embedding

Consider the elastic embedding objective function:

$$E(\mathbf{Z}) = \sum_{n,m=1}^{N} \left(w_{nm} \|\mathbf{z}_n - \mathbf{z}_m\|^2 + \alpha e^{-\|\mathbf{z}_n - \mathbf{z}_m\|^2} \right)$$

Call the resulting embeddings $\mathbf{z}$ the free embedding. If we want an out-of-sample mapping $\mathbf{F}$ so we can project new points, then $\mathbf{z} = \mathbf{F}(\mathbf{x})$ by definition and we have a parametric embedding objective function:

$$E(\mathbf{F}) = \sum_{n,m=1}^{N} \left(w_{nm} \|\mathbf{F}(\mathbf{x}_n) - \mathbf{F}(\mathbf{x}_m)\|^2 + \alpha e^{-\|\mathbf{F}(\mathbf{x}_n) - \mathbf{F}(\mathbf{x}_m)\|^2} \right) + \lambda\, \phi(\mathbf{F})$$

Not easy to optimize since $\mathbf{F}$ is non-differentiable and non-convex mapping (a tree)!

Jointly learning an optimal tree and embedding

- Solution: use the method of auxiliary coordinates (MAC) [2, 3]. Consider the following equivalent constrained problem with "auxiliary coordinates" $\mathbf{Z}$:

$$\min_{\mathbf{Z},\mathbf{F}} E(\mathbf{Z}) + \lambda\,\phi(\mathbf{F}) \quad \text{s.t.} \quad \mathbf{Z} = \mathbf{F}(\mathbf{X})$$

We solve this using a penalty method. We describe the quadratic penalty method for simplicity, but in the experiments we use the augmented Lagrangian. This defines a new, unconstrained objective function:

$$\min_{\mathbf{Z},\mathbf{F}} E(\mathbf{Z}) + \lambda\,\phi(\mathbf{F}) + \mu\|\mathbf{Z} - \mathbf{F}(\mathbf{X})\|^2. \tag{1}$$

Jointly learning an optimal tree and embedding

Finally, we optimize (1) by alternating optimization over $\mathbf{Z}$ and $\mathbf{F}$:

- Over $\mathbf{Z}$, eq. (1) is the original embedding objective E but with a quadratic regularization term on $\mathbf{Z}$:

$$\min_{\mathbf{Z}} E(\mathbf{Z}) + \mu \|\mathbf{Z} - \mathbf{F}(\mathbf{X})\|^2.$$

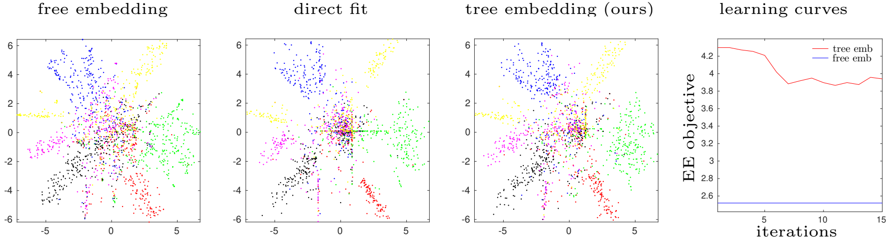
  Solution: off-the-shelf algorithm to optimize the original embedding (e.g. t-SNE) with a minor modification to handle the additional quadratic term.

- Over $\mathbf{F}$, eq. (1) reduces to a regression fit of a tree which we solve using the Tree Alternating Optimization (TAO) [4]:

$$\min_{\mathbf{F}} \|\mathbf{Z} - \mathbf{F}(\mathbf{X})\|^2 + \frac{\lambda}{\mu} \phi(\mathbf{F})$$

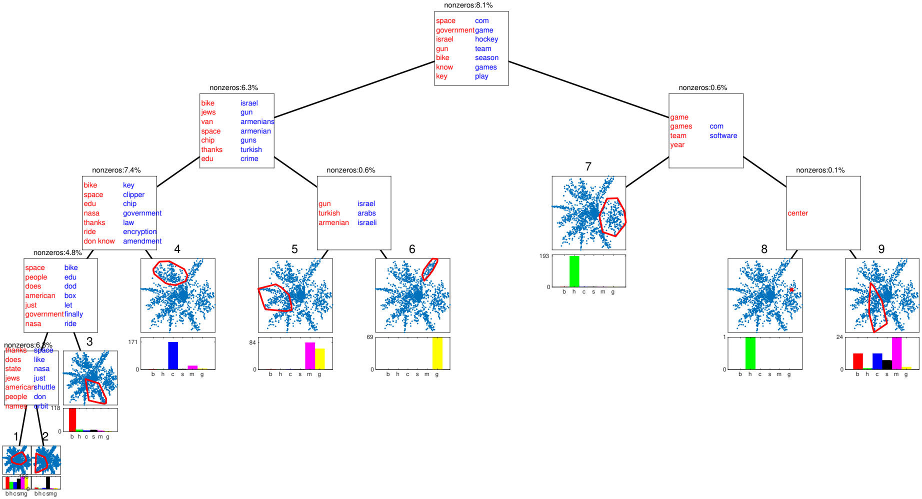
  The ability of the TAO algorithm to take an initial tree and improve over it is essential here to make sure that the step over $\mathbf{F}$ improves over the previous iteration, and to be able to use warm-start to speed up the computation.

Experiments



- Results on 20-newsgroups dataset: 6 classes, tf-idf statistics on unigrams and bigrams as features (1000 features in total).
- We used elastic embedding to produce the free embedding.
- Direct fit trains an oblique tree (using TAO) directly to a free embedding, i.e. it uses free embedding as a label.
- The first iteration ($\mu = 0$) in learning curves (left plot) represents a direct fit. Our proposed approach (tree embedding) improves over this baseline (see iterations).

Experiments



References

[1] M. Á. Carreira-Perpiñán. The elastic embedding algorithm for dimensionality reduction. In J. Fürnkranz and T. Joachims, editors, *Proc. of the 27th Int. Conf. Machine Learning (ICML 2010)*, pages 167–174, Haifa, Israel, June 21–25 2010.

[2] M. Á. Carreira-Perpiñán and W. Wang. Distributed optimization of deeply nested systems. arXiv:1212.5921, Dec. 24 2012.

[3] M. Á. Carreira-Perpiñán and W. Wang. Distributed optimization of deeply nested systems. In S. Kaski and J. Corander, editors, *Proc. of the 17th Int. Conf. Artificial Intelligence and Statistics (AISTATS 2014)*, pages 10–19, Reykjavik, Iceland, Apr. 22–25 2014.

[4] A. Zharmagambetov and M. Á. Carreira-Perpiñán. Smaller, more accurate regression forests using tree alternating optimization. In H. Daumé III and A. Singh, editors, *Proc. of the 37th Int. Conf. Machine Learning (ICML 2020)*, pages 11398–11408, Online, July 13–18 2020.