Active Heirarchical Metric Learning

Nicolas Beltran
Department of Computer Science
Columbia University
New York City, NY 10027
nb2838@columbia.edu

Ketan Jog
Department of Computer Science
Columbia University
New York City, NY 10027
kj2473@columbia.edu

Abstract

Many problems require a well defined notion of a distance between points in space. Constructing or finding such a measure falls into the field of metric learning. Although many algorithms exist in the field when a learner has access to a fixed dataset, there is room for improvement in terms of samples efficiency that the learner needs to know, imposition of desired structure, especially when the data appears in an *online* manner. We propose a project that reduces the problem of online/active metric learning to bandits. In case our plan turn out to be too ambitious, we have a fallback - an empirical investigation of some algorithms that have dealt with the problem in an online setting or in situations where the learner can selective query the points that it wants to know information about.

1 Introduction

Metric learning consists in smartly adjusting a distance or similarity function using using data. The resulting measure is well suited to the problem of interest and can lead to significant improvement in downstream tasks like classification, regression or ordering [1]. Lots of methods focus on learning a Mahalanobis distance. This can be seen as finding a linear projection of the data. There has been work done especially in online metric learning algorithms that can offer guarantees in the form of regret bounds [4]. Most metric learning algorithms however work with very granular levels of data. Online metric learning algorithms like POLA and LEGO develop online methods that work with similarity and dissimilarity sets. While these algorithms succeed in capturing local structure in the data embedding space, due to the nature of the provided context they fail to encapsulate the general global structure that the data is being sampled from. Many practical settings of metric learning problems allow active access to queryable data. Furthermore, the complexity of such problems requires that the metric imposed on the data not only be accurate on points chosen close by, but also to interpret the meaning of distance between datapoints that are chosen from very different localities. The goal for this project is to devise a metric learning algorithm in an online setting, which successfully preserves the *global* structure of the data in the embedding space. An online algorithm receives requests one by one and needs to "respond" each request immediately without knowing future requests. We further want to analyse the regret of our formulation.

In general, learning a nonlinear transformation is difficult — unlike linear transformations, which can be expressed as a matrix of parameters, the set of nonlinear transformations is difficult to parametrize. To begin with, we will restrict ourselves to the set of linear transformations, maybe further delving into kernelized linear transformations. We will begin by making some strong distributional assumptions, such as the data distribution being a mixture of Gaussians with each Gaussian as one of the clusters. Our notion of global structure will be encoded in some object like a tree, where ideas of intercluster and intracluster distance can be formalised.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

2 Proposal

We would like to focus on creating a new algorithm for hierarchical metric learning in an active setting. Our goals are defined precisely in section (2.1). Due to the difficulty of the problem we also have an additional proposal which focuses on an empirical analysis of active metric learning algorithms. We intend to work on both for most of the semester and focus on one during the last weeks for the final report and presentation.

2.1 Main proposal

Our more ambitious goal is to design an algorithm that can learn a mahalanobis metric in an online fashion using expert feed back. Formally, we can describe it as follows:

Assumptions We assume that:

1. There exists metric space $(\mathcal{X}, d^\star)$ where $d^\star \in M$ and M is the space of all Mahalanobis metrics.
2. There exists a tree $\mathcal{T} = (V, E)$ and a mapping $\tau : \mathcal{X} \to V$ such that each $x \in \mathcal{X}$ is mapped to exactly one of the leaves of the tree. Intuitively, this is meant to represent hierarchies where each node of the tree refers to grouping and subtrees represent subgroupings.
3. There is an oracle $\mu : \mathcal{X} \times \mathcal{X} \times M$ which provides feedback on how good a proposed metric is in representing the true distance of two points. This definition is vague because we haven't yet settled exactly on what type of feedback to use.
4. We have access to a finite set of elements $\mathcal{D} = \{x_i \in \mathcal{X} | i \in [n]\}$ but the metric $d^\star$ is unkown to us. Furthermore we assume that there are at least 2 elements for each leaf in the tree.
5. Let $f : \mathcal{X} \times \mathcal{V} \to \{0, 1\}$ be a function which assigns labe 1 if point $x \in \mathcal{X}$ is a child of node $v \in \mathcal{V}$ and label 0 otherwise. Then we assume that under $d^\star$.
$$0 = \sum_{x \in \mathcal{D}} \sum_{v \in V} \mathbb{1}\left\{f(x, v) \neq f(\text{argmin}_x^* d^*(x^*, x), v)\right\}$$
This is to say that a 1-KNN classifier would do a perfect job at classifying points in the hierarchies.

Goal To find an algorithm that can learn mahalanobis metric in an online fashion by using feedback in the form of the oracle μ assuing that $\mathcal{T}$ is knonw. Furthermore, this metric d should satisfy
$$0 = \sum_{x \in \mathcal{D}} \sum_{v \in V} \mathbb{1}\left\{f(x, v) \neq f(\text{argmin}_x^* d^*(x^*, x), v)\right\}$$

Idea We propose to use the following skeleton for the algorithm. Let B refer to a bandit algorithm which recieves two points x, y, previous rewards and returns a metric. Let $A(\mathcal{D})$ be a an algorithm which samples points from $\mathcal{D}$ in some way. Then the algorithm can be stated as follows:

Algorithm 1 Algorithm skeleton

while Stop criterion not met **do**
  $x, y \leftarrow$ sampled from $A(\mathcal{D})$
  $d \leftarrow B(x, y)$
  $r \leftarrow \mu(d)$
  B.update(r)
  $A(\mathcal{D})$.update(r,d)
end while

In words, our idea is to use an existing contextual bandit algorithm that recieves as context two points provided by A and then returns a metric which recieves some feedback from a reward function. Said differently, our goal is to reduce metric learning to bandits. Most of our work would be in designing both $\mathcal{A}$ and μ but we would need some work to decide what contextual bandit algorithm to use. In particular, it is not clear if linear bandits are the right approach because of the positive definite constraint on the metric and the shape of the reward function.

2.2 Fall back proposal

As a fallback project we intend to provide a literature survey and empirical evaluation of various algorithms which share similarities with ours or have desirable properties which we believe we could take inspiration from. We believe that the evaluation of these algorithms should be done on a common dataset of gaussian clusters, CIFAR-10, MNIST when possible and on synthetic datasets relevant to the specific algorithms. Below we describe the algorithms we intend to evaluate.

Structural query-by-committee Query-by-committee is a popular active learning algorithm that has been well studied for data labeling problems [3]. In [5] an extension is proposed which handles settings on which there is strcuture. In our case this structure is a class $\mathcal{F}$ of metrics on a space $\mathcal{X}$. Although the paper has a special focus on clustering, the framework presented can be adapted to metric learning as we intend to do. Moreover, we would think this would be an interesting contribution as there were no empirical experiments in such a setting in the original paper.

Bayesian Active Metric Learning In [7] the authors propose an algorithm for active metric learning in a setting where feedback from the experts exists via equivalence and inequivalence constraints (should the two points be together or not). This algorithm uses variational inference for updates, and a laplacian approximation to compute entropies used to determine which points to query. This would be helpful for our setting because it provides a framework which we could use to expand Structural Query by Committee to handle metrics, and it would provide inspiration for query selection (i.e designing A).

Stochastic Triplet Embedding Stochastic selection rules have had great success in dimensionality reduction. In [6] the authors use a stochastic neighbor approach as implemented in t-sne on triplet data of the form, "A is more similar to B than C". The authors show preservation of local structure in lower dimensional embeddings via this formulation. We want to mimic the way partial ordering information is used to design an embedding, to use sexpert signal on similar lines to construct a metric learning algorithm. It might be possible to adapt T-ste to an online setting, and that might serve as an exploratory direction for our project.

Low-dimensional embedding using adaptively selected ordinal data In [2], the authors study the problem of learning an embedding of n objects into d-dimensional Euclidean space. Like in [6], they focus on comparisons of the type "A is similar to B than C." This paper explores the lower bound on the number of comparisons that are needed to create such an embedding. They further create an algorithm that tries to achieve that bound by smart query selection. An empirical analysis of this algorithm will serve us well as part of our fallback, while we plan to build on the query selection algorithm used in this paper to formulate $A(\mathcal{D})$ that samples datapoints for our bandit formulation.

References

[1] Aurélien Bellet and Amaury Habrard. Robustness and generalization for metric learning. *CoRR*, abs/1209.1086, 2012.

[2] Kevin G. Jamieson and Robert D. Nowak. Low-dimensional embedding using adaptively selected ordinal data. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1077–1084, 2011.

[3] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 287–294, New York, NY, USA, 1992. Association for Computing Machinery.

[4] Shai Shalev-Shwartz, Yoram Singer, and Andrew Y. Ng. Online and batch learning of pseudo-metrics. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 94, New York, NY, USA, 2004. Association for Computing Machinery.

[5] Christopher Tosh and Sanjoy Dasgupta. Structural query-by-committee. *CoRR*, abs/1803.06586, 2018.

[6] Laurens van der Maaten and Kilian Weinberger. Stochastic triplet embedding. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, 2012.

[7] Liu Yang, Rong Jin, and Rahul Sukthankar. Bayesian active distance metric learning, 2012.