Nonparametric Divergence Estimation for Learning Manifolds of Distributions and Group Anomaly Detection

Barnabás Póczos, bapoczos@cs.cmu.edu
Liang Xiong, lxiong@cs.cmu.com
Jeff Schneider schneide@cs.cmu.edu

Carnegie Mellon University,
School of Computer Science,
5000 Forbes Ave, Pittsburgh, PA, USA, 15213

Low-dimensional embedding, manifold learning (Roweis and Saul, 2000), clustering, and anomaly detection (Chandola et al., 2009) are important problems in unsupervised learning and machine learning. The existing methods usually consider the case when each instance has a fixed, finite-dimensional feature representation, and the goal is to embed these feature vectors into a lower dimensional space, perform clustering, or detect outlier feature vectors. Here we consider a different setting. We assume that each instance corresponds to a continuous probability distribution. These distributions are unknown, but we are given some i.i.d. samples from each distribution. Our goal is to estimate the distances between these distributions and using these distances to perform low-dimensional embedding, clustering, or anomaly detection for these distributions.

The formal definition of the problem is as follows. We are given $\{X_{i,1},\dots,X_{i,T_i}\}$ i.i.d. samples from $\{f_i\}$ density functions ($i=1,\dots,I$, $X_{i,t}\in\mathbb{R}^D$). Using these samples, we want to estimate the distances between these $\{f_i\}$ density functions. Here, we use the L_2-divergence: $w_{i,j}\doteq\left(\int(f_i(x)-f_j(x))^2\mathrm{d}x\right)^{1/2}$. The problem, of course, is that we do not know these $\{f_i\}$ densities, and we want to compute the L_2-divergences without estimating them. For this purpose, we propose a consistent L_2-divergence estimator. The estimator is simple, can avoid the need for density estimation, and uses only certain k nearest neighbor statistics with a fixed k.

Let c denote the volume of a D-dimensional unit ball. For a fixed (i,j), let $\rho(t)$ be the Euclidean distance of the kth nearest neighbor of $X_{i,t}$ in the sample $X_{i,1:T_i\setminus t}$, and similarly let $\nu(t)$ denote the distance of the kth nearest neighbor of $X_{i,t}$ in the sample $X_{j,1:T_j}$. Under certain conditions, we can prove that the following expression is an L_2-consistent estimator for $w_{i,j}$.

$$\widehat{w}_{i,j}\doteq\frac{1}{T_i}\sum_{t=1}^{T_i}\left[\frac{k-1}{(T_i-1)c\rho^D(t)}+\frac{(T_i-1)c\rho^D(t)}{(T_jc\nu^D(t))^2}\frac{(k-2)(k-1)}{k}-\frac{2(k-1)}{T_jc\nu^D(t)}\right].$$

Having estimated the $\{w_{i,j}\}$ distances, we can analyze the distributions as if they were points in a finite-dimensional Euclidean space. For example, we can cluster the distributions, or embed them into a low-dimensional space while preserving proximity; distributions close to each other should be mapped into points that are also close to each other in the lower dimensional space. This can be done using multidimensional scaling (Borg and Groenen, 2005), Isomap (Tenenbaum et al., 2000), or other methods that require only the pairwise distances. This embedding provides a useful tool for visualization and unsupervised exploration of the data set.

Another interesting application of the proposed divergence estimator is the group anomaly detection problem. Anomaly detection is a widely studied research area. Most results, however, focus only on finding individual outlier points. Nonetheless, interesting larger scale phenomena can only be discovered when aggregated data is considered. For example, finding unusual galaxies in a sky survey is a standard anomaly detection problem, while finding unusual spatial clusters of galaxies is a group anomaly detection problem. Unlike traditional detection methods that focus on individual points, we are interested in finding *groups* of points that exhibit unusual behavior. We model each group as a bag of features, and assume that the ith group has a feature distribution f_i. Our goal is

to select those groups whose feature distributions are significantly different from the distributions of the other groups. This problem can be addressed using the proposed divergence estimator by first estimating the distances between the groups' feature distributions and then finding those groups that are far away from their neighbors.

References

I. Borg and P. Groenen. *Modern Multidimensional Scaling: theory and applications*. Springer-Verlag, New York, 2005.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41-3, 2009.

S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Topic: learning algorithms
Preference: oral presentation
Presenter: Barnabás Póczos