

1 Introduction

The task of semi-supervised learning learning assumes we have a small set of labeled examples and a much larger set of unlabeled examples and we want to use the distribution of the combined data to find a good embedding of the unlabeled examples to make it easier to classify them.

Cluster assumption: two points are likely to have the same class label if they are located in the same high density neighborhood.

Goal: Find an embedding based on the cluster assumption so that generalization classification accuracy of unlabeled points is as high as possible.

2 Some background

Recent advances in spectral clustering (particularly Ng, Jordan, and Weiss 2002) have shown how very good embeddings based on the cluster assumption can be used for machine learning. The basic idea used by Ng et al. and others is essentially to transform an $n \times m$ iid feature vector data set $\mathbf{X} = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n\}$ into a weighted adjacency (affinity) matrix A where $[A]_{ij} = \exp(-\|\vec{x}_i - \vec{x}_j\|^2/2\sigma^2)$. Using this transformation distances between points that are large (points are far apart) get mapped to edge weights (affinities) that are small and distances between points that are small get mapped to edge weights that are large. The matrix A however has many properties we would like to avoid. The two main ones are: 1) the matrix is not guaranteed to be positive semi-definite and 2) similar points are not as clustered together as they should be. However, there are a number of transformations that can be made on the matrix A which will guarantee that it is positive semi-definite (necessary for it to be a valid kernel matrix) and which will cluster similar points together. One example, not discussed in the paper, is called the Graph Laplacian, $L = D - A$, where D is the diagonal matrix where D_{ii} is the sum of row i in matrix A . Using this transformation, we get a matrix that is guaranteed to be positive semi-definite and whose spectral embedding (eigenvectors of the matrix correspond to dimensions in Euclidean space) optimally positions the points in the sense that the Euclidean distances between the points are minimized. We can see this as follows: << Show proof on white board if there is interest >>

Other transformations have been offered in the literature as well. They include Kernel PCA, random walk kernel, and others. The importance of this paper is that it shows how many of these transformations can all be united under a common framework. It also gives experimental results showing that some new transformations discovered using this framework give better classification accuracy than previously published semi-supervised techniques for data sets where the number of labels is relatively small.