Lecture 21: Semi-supervised Learning

So far throughout lectures on ML we've been discussing supervised & unsupervised learning problems & techniques.

Semi-supervised learning combine features of both supervised & unsupervised learning. In a way it is half way in between.

In SSL we assume our data is available in the following form,

we have inputs

$$\mathbb{R}^{d \times N} \ni X = \{\underline{x}_0, \underline{x}_1, \dots, \underline{x}_{N-1}\}$$

along with some outputs,

$$\mathbb{R}^M \ni Y = \{y_0(\underline{x}_0), y_1(\underline{x}_1), \dots, y_{M-1}(\underline{x}_{M-1})\}$$

ie, outputs are only available for a subset $M \leq N$ of our inputs.

*Note ordering of $\underline{x}_j$ is innocuous so we can always re-order our data so that $\underline{x}_0, \dots, \underline{x}_{M-1}$ are the inputs for which the outputs are observed.

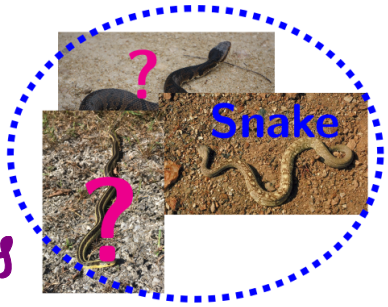
We call the pairs $\{(\underline{x}_0, y_0), \dots, (\underline{x}_{M-1}, y_{M-1})\}$ the labelled data (sub)set and the set $\{\underline{x}_M, \dots, \underline{x}_{N-1}\}$ the unlabelled data (sub) set.

Then the goal of SSL is to predict the outputs $\{y_M(\underline{x}_M), \dots, y_{N-1}(\underline{x}_{N-1})\}$ given the labelled & unlabelled data sets.

Idea: • If we only had the labelled dataset we would have used SL, such as kernel regression.

• If we only had the unlabelled data set we would have used some kind of spectral clustering.

• Combine the two approaches, by using the unlabelled set to create optimal feature maps that after SL on the labelled set help us predict the "unseen" labels in an optimal way!



Here we will focus on the particular case of SSL with graph Laplacian regularization.
a.k.a manifold regularization.

21.1 Semi-supervised regression with graph Laplacians

Semi-supervised regression (SSR) is a particular case of SSL where the outputs $y(\underline{x})$ are real valued then $\underline{y} = (y(\underline{x}_0), \ldots, y_{M-1}(\underline{x}_{M-1})) \in \mathbb{R}^M$.

We proceed analogously to kernel regression, ie, we wish to find a function

$$f(\underline{x}) = \sum_{j=0}^{J} c_j \cdot \psi_j(\underline{x})$$

so that $f(\underline{x}_j) \approx y(\underline{x}_j)$ for $M \leq j \leq N-1$. ie, we want f to approximate the output well only on the unlabelled set & not for any new input $\underline{x}$ as in regression!

so the question is, how do we pick the feature maps ψ_j? Graph Laplacian embedding!

Recall, from our discussion on spectral clustering that the eigenvectors of graph Laplacian graphs can serve as meaningful feature maps that encode geometry/clusters present in our data.

Then we propose to take the φ_j to be a finite number of the eigenvectors of a graph Laplacian on a similarity graph on X.

This leads to the following pseudo-algorithm

▷ Given $X \in \mathbb{R}^{d \times N}$ & $Y \in \mathbb{R}^M$.

- construct a similarity graph $G = \{X, W\}$ along with a graph Laplacian $L \in \mathbb{R}^{N \times N}$.
- Compute the first $K > 0$ eigenvectors $\{\underline{q}_k\}_{k=0}^{K-1}$ of L.
- Solve the Ridge regression problem

$$\hat{\underline{c}} = \underset{\underline{c} \in \mathbb{R}^K}{\operatorname{argmin}} \sum_{j=0}^{M-1} \left| \sum_{k=0}^{K-1} c_k q_{jk} - y_j \right|^2 + \lambda \|\underline{c}\|_2^2$$

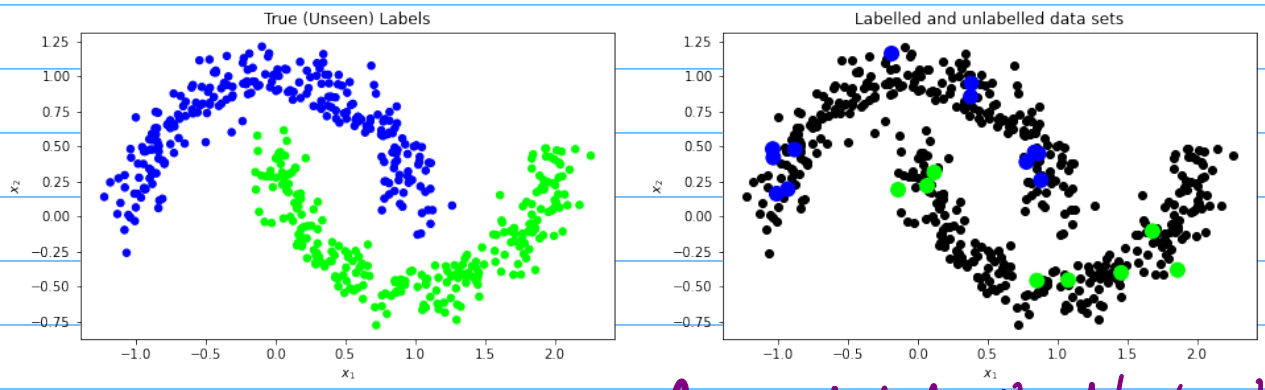
* Note you can use any normalization of L you like.

Observe that the above recipe leads to a function $\hat{f}$ of the form

$$\hat{f}(\underline{x}_j) = \sum_{k=0}^{K-1} \hat{c}_k q_{jk}, \quad \forall \underline{x}_j \in X.$$

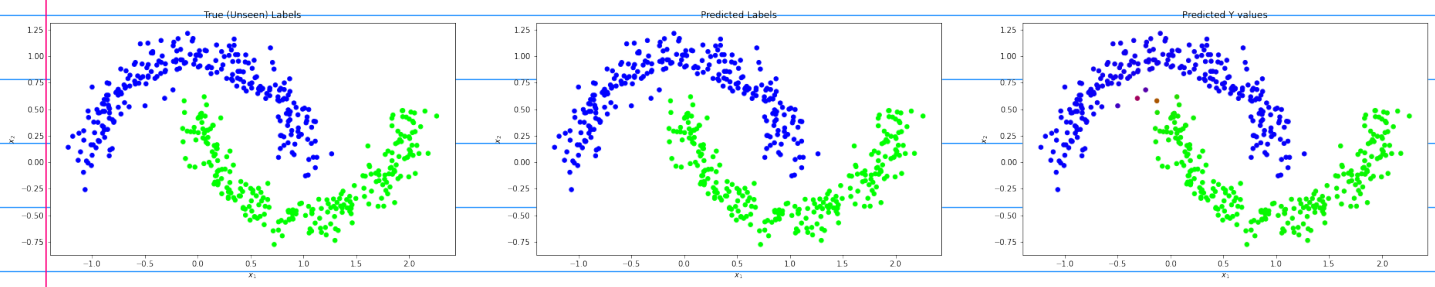
thus, $\hat{f}$ is only defined for points $\underline{x}_j \in X$!

ex two moons toy problem again



Given data on right frame label the black pts.

We use the same setup as Lecture 20 to construct our graph & the Laplacian. Our approach only differs in applying Ridge regression in the last step as opposed to K-means.



true labels

predicted labels $sign(\hat{f}(x_j))$

corresponding $\hat{f}(x_j)$ values.

Further reading:

Computer Sciences Department

Semi-Supervised Learning Literature Survey

Xiaojin Zhu

Technical Report #1530

September 2005

UNIVERSITY OF WISCONSIN MADISON

Laplacian Eigenmaps for Dimensionality Reduction and Data Representation

Mikhail Belkin* Partha Niyogi†

December 8, 2002

Journal of Machine Learning Research 7 (2006) 2399-2434 Submitted 4/05; Revised 5/06; Published 11/06

Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples

Mikhail Belkin MBELKIN@CSE.OHIO-STATE.EDU
Department of Computer Science and Engineering
The Ohio State University
2015 Neil Avenue, Dreese Labs 597
Columbus, OH 43210, USA

Partha Niyogi NIYOGI@CS.UCHICAGO.EDU
Departments of Computer Science and Statistics
University of Chicago
1100 E. 58th Street
Chicago, IL 60637, USA

Vikas Sindhwani VIKASS@CS.UCHICAGO.EDU
Department of Computer Science
University of Chicago
1100 E. 58th Street
Chicago, IL 60637, USA

On Manifold Regularization

Mikhail Belkin, Partha Niyogi, Vikas Sindhwani
{misha,niyogi,vikass}@cs.uchicago.edu
Department of Computer Science
University of Chicago