Graphical Semi-Supervised Learning

AMATH 563

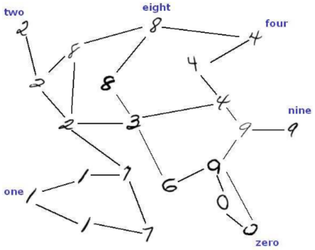
Lucas Cassin Cruz Burke

May 31, 2023

UNIVERSITY of WASHINGTON

Semi Supervised Learning

- Semi-supervised learning is the problem of predicting the labels of an unlabeled set of data $(x_i, y_i)_{i=M+1}^{N}$ based on a subset of labeled datapoints $(x_i, y_i)_{i=1}^{M}$.
- Graphical SSL algorithms build a graph on top of the x_i's and formulate the problem as a regularized regression problem on G using the Graph Laplacian matrix, which can be used to learn the structure of the underlying manifold.



Probit model & Laplacian based regularization

- We assume our generating process is of the form $y_j = \mathrm{sign}(f(\mathbf{x}_j) + \varepsilon_j)$ with normally distributed noise $\varepsilon_j \sim \psi$.
- We then formulate a binary classification task in the RKHS framework using the Probit loss function and the regularization function given by

$$\mathbf{f}^* = \arg\min_{f \in \mathbb{R}^M} \left\{ -\sum_{j=1}^{N} \log \Psi(f_j y_j) + \beta \mathbf{f}^T C \mathbf{f} \right\}$$

where $C = \left(\Delta + \tau^2 I\right)^{-\alpha/2}$ and $\Delta = D - W$ is the graph Laplacian.

- C is strictly PDS and defines an RKHS, hence by the reproducing property we have

$$\mathbf{f}^* = C(:, 1:N) C(1:N, 1:N)^{-1} \mathbf{z}^*$$

where $\mathbf{z}^* = \arg\min_{\mathbf{z} \in \mathbb{R}^N} \left\{ -\sum_{j=1}^{N} \log \Psi(y_j z_j) + \beta \mathbf{z}^T C(1:N, 1:N)^{-1} \mathbf{z} \right\}$.

A related problem

Let us briefly ponder the related inhomogeneous differential equation

$$C\mathbf{f} = \left(\Delta + \tau^2 I\right)^{-\alpha/2} \mathbf{f} = \mathbf{h}$$

- C is linear and elliptic, and so the solution can be written using the Green's function G as

$$\mathbf{f}(\mathbf{x}) = \int h(\mathbf{z}) G(\mathbf{x}, \mathbf{z}) d\mathbf{z}$$

  where $(CG)(\mathbf{x}, \mathbf{z}) = \delta(\mathbf{x} - \mathbf{z})$.
- It turns out that this Green's function is given by the Matérn kernel functional

$$C_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{d}{\rho}\right)^\nu K_\nu \left(\sqrt{2\nu}\frac{d}{\rho}\right)$$

- The Matérn family includes the exponential kernel ($C_{1/2}$) and the Gaussian kernel (C_∞) as particular cases.

SSL by diffusion: Laplace & Poisson methods

- **Laplace method**: Treat labeled points as Dirichlet boundary conditions.

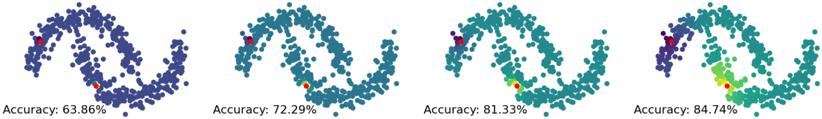


Figure: Heat diffusion with Dirichlet boundary conditions at labeled points.

- **Poisson method**: Treat labeled points as heat sources/sinks.

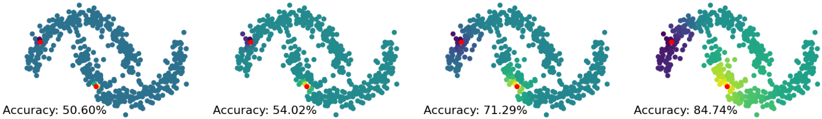


Figure: Heat diffusion with sources/sinks at labeled points.

KNN (K=5). Gaussian kernel. Time steps incrementing by powers of 10 from 10^2 *to* 10^5*.*

Matérn kernel parameters

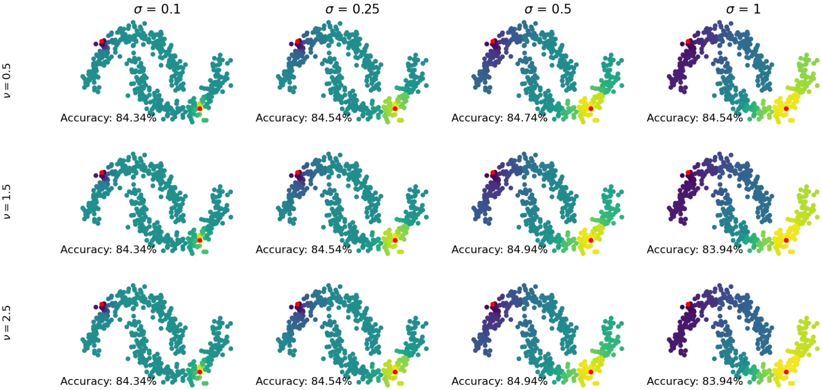


Figure: Laplace method solve for different values of ν and σ.

Comparing Graphical SSL methods

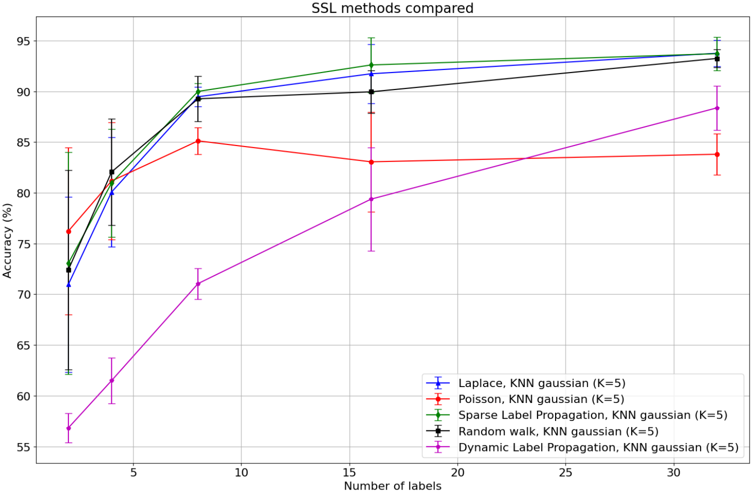


Figure: Comparison of graphical SSL method performance on Two Moons Probit problem.

Thank you :)

References

1. Mikhail Belkin and Partha Niyogi. Semi-supervised learning on riemannian manifolds. *Machine learning*, 56:209–239, 2004.
2. Viacheslav Borovitskiy, Iskander Azangulov, Alexander Terenin, Peter Mostowsky, Marc Deisenroth, and Nicolas Durrande. Matérn gaussian processes on graphs. In *International Conference on Artificial Intelligence and Statistics*, pages 2593–2601. PMLR, 2021.
3. Franca Hoffmann, Bamdad Hosseini, Zhi Ren, and Andrew M Stuart. Consistency of semi-supervised learning algorithms on graphs: Probit and one-hot methods. *The Journal of Machine Learning Research*, 21(1):7549–7603, 2020.
4. Daniel Sanz-Alonso and Ruiyi Yang. The spde approach to matérn fields: Graph representations. *Statistical Science*, 37(4):519–540, 2022.
5. Xiaojin Jerry Zhu. Semi-supervised learning literature survey. 2005.
6. Jeff Calder. GraphLearning Python Package. Zenodo. 2022.
7.