Transductive Kernels for Gaussian Processes on Graphs

Oxford-Man Institute

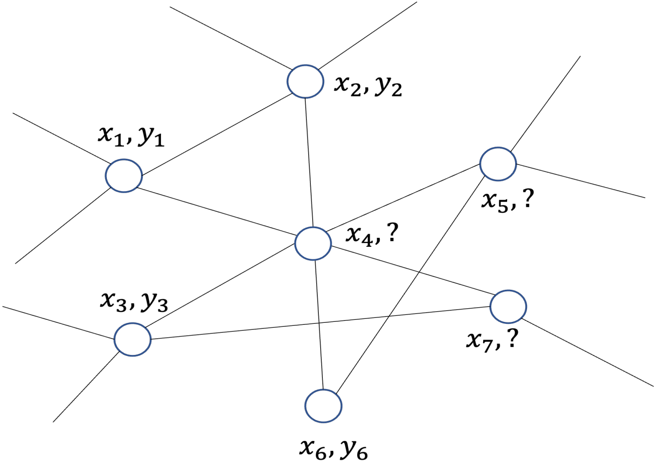
Department of Engineering

Yin-Cong Zhi

Felix L. Opolka, Yin Cheng Ng, Pietro Liò, Xiaowen Dong

March 9, 2023

Semi-Supervised Classification with Graph Gaussian Processes



What is $k(x_i, x_j)$?

Kernels and Regularization

Many machine learning problems can be presented as a minimization task

$$\min_f L(f, y) + \Omega(||f||^2).$$

Kernels and Regularization

Many machine learning problems can be presented as a minimization task

$$\min_f L(f, y) + \Omega(||f||^2).$$

From a kernel methods point of view, we are interested in the regularizer $\Omega(||f||^2)$:

$$\Omega(||f||^2) = \langle Pf, Pf \rangle = \langle f, P.Pf \rangle = \langle f, r(\Delta)f \rangle = \langle f, f \rangle_{\mathcal{H}}.$$

The Hilbert space $\mathcal{H}$ w.r.t. inner product with $r(\Delta)$.

Kernels and Regularization

Many machine learning problems can be presented as a minimization task

$$\min_f L(f, y) + \Omega(||f||^2).$$

From a kernel methods point of view, we are interested in the regularizer $\Omega(||f||^2)$:

$$\Omega(||f||^2) = \langle Pf, Pf \rangle = \langle f, P.Pf \rangle = \langle f, r(\Delta)f \rangle = \langle f, f \rangle_{\mathcal{H}}.$$

The Hilbert space $\mathcal{H}$ w.r.t. inner product with $r(\Delta)$.
Examples include $P = 1, \nabla, (1, \nabla)^\top, \ldots$; and $\nabla.\nabla = \Delta$

$$\nabla = \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}\right) \qquad \Delta = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}$$

Reproducing Kernel Hilbert Space

A function $k(\cdot,\cdot)$ is called a reproducing kernel of $\mathcal{H}$ if it satisfies

$$\forall x, f,\ \langle f, k(x,\cdot)\rangle_{\mathcal{H}} = f(x) \quad (= \langle f, r(\Delta)k(x,\cdot)\rangle),$$

$\mathcal{H}$ is then a RKHS. This is called the *reproducing property*.

The kernel function is derived by computing $K = r^{-1}(\Delta)$ (more details later).

Reproducing Kernel Hilbert Space

A function $k(\cdot,\cdot)$ is called a reproducing kernel of $\mathcal{H}$ if it satisfies

$$\forall x, f,\ \langle f, k(x,\cdot)\rangle_{\mathcal{H}} = f(x) \quad (= \langle f, r(\Delta)k(x,\cdot)\rangle),$$

$\mathcal{H}$ is then a RKHS. This is called the *reproducing property*.

The kernel function is derived by computing $K = r^{-1}(\Delta)$ (more details later).

Representer Theorem:
If $\mathcal{H}$ is a reproducing kernel hilbert space (RKHS), then there is a solution to

$$\min_f L(f,y) + \Omega(||f||^2_{\mathcal{H}})$$

that takes the form $f^*(\cdot) = \sum_i \alpha_i k(\cdot, x_i)$.

Kernel Functions

Examples of regularization functions $r(\Delta)$, and their kernels (Smola. A. 2003):

- RBF:

$$r(\Delta) = \sum_i \frac{\sigma^{2i}}{2^i i!} \Delta^i \implies k(x, x') = \exp\left\{ -\frac{1}{2\sigma^2} ||x - x'||^2 \right\}$$

- Laplacian kernel:

$$r(\Delta) = 1 + \sigma^2 \Delta \implies k(x, x') = \exp\left\{ -\frac{1}{\sigma} ||x - x'|| \right\}$$

Graphs Laplacians

The graph Laplacian is defined as

$$L = D - A$$

Graphs Laplacians

The graph Laplacian is defined as

$$L = D - A$$

This is the finite difference of Δ on a discrete space

$$\Delta \equiv L$$

Graphs Laplacians

The graph Laplacian is defined as

$$L = D - A$$

This is the finite difference of Δ on a discrete space

$$\Delta \equiv L$$

To define kernels on graphs, the regularization function $\Omega(||f||^2_{\mathcal{H}})$ becomes (Smola. A. 2003):

$$\langle f, r(\Delta) f \rangle \rightarrow \langle f, r(L) f \rangle$$

Kernels on Graphs

If f is a function on a graph G, reproducing property is

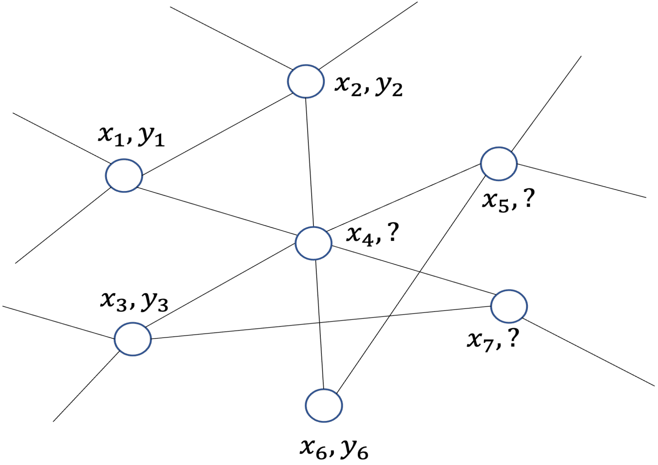
$$\begin{aligned}\langle f, r(L)K(i,:)\rangle_i &= f(i)\\ \implies f^\top r(L)K &= f^\top\\ \implies K &= r^{-1}(L)\end{aligned}$$

Kernel between two nodes is

$$k(v_i, v_j) = r^{-1}(L)_{i,j}$$

Note, this kernel depends on the graph only, and not any node data.

Kernels on Attributed Graphs



What is $k(x_i, x_j)$?

Kernels on Attributed Graphs

What if we have a graph with feature data on the nodes (node attributes)?

Start with the minimization task again

$$\min_f L(f, y) + \Omega(||f||^2)$$
$$\Omega(||f||^2) = \langle f, f \rangle_{\mathcal{H}}.$$

What $\mathcal{H}$ do we choose?

Kernels on Attributed Graphs

What if we have a graph with feature data on the nodes (node attributes)?

Start with the minimization task again

$$\min_f L(f, y) + \Omega(||f||^2)$$
$$\Omega(||f||^2) = \langle f, f \rangle_{\mathcal{H}}.$$

What $\mathcal{H}$ do we choose?

- $\langle f, r(\Delta) f \rangle$ leads to $k(x_i, x_j)$ - depends on feature only
- $\langle f, r(L) f \rangle$ leads to $k(v_i, v_j)$ - depends on graph only

Kernels on Attributed Graphs

What if we have a graph with feature data on the nodes (node attributes)?

Start with the minimization task again

$$\min_f L(f, y) + \Omega(||f||^2)$$
$$\Omega(||f||^2) = \langle f, f \rangle_{\mathcal{H}}.$$

What $\mathcal{H}$ do we choose?

- $\langle f, r(\Delta) f \rangle$ leads to $k(x_i, x_j)$ - depends on feature only
- $\langle f, r(L) f \rangle$ leads to $k(v_i, v_j)$ - depends on graph only

Solution: f depends on x and G, so use $r(\Delta)$ and $r(L)$!

Kernels for Attributed Graphs

$$\Omega(||f||^2) = \langle f, [r_1(\Delta) + r_2(L)]f \rangle = \langle f, r_1(\Delta)f \rangle + \langle f, r_2(L)f \rangle$$

- $r_1(\Delta)$ deals with the feature data ($K_1 = r_1^{-1}(\Delta)$)
- $r_2(L)$ deals with the graph

$$K = [K_1^{-1} + r_2(L)]^{-1}$$

Kernels for Attributed Graphs

$$\Omega(||f||^2) = \langle f, [r_1(\Delta) + r_2(L)]f \rangle = \langle f, r_1(\Delta)f \rangle + \langle f, r_2(L)f \rangle$$

- $r_1(\Delta)$ deals with the feature data ($K_1 = r_1^{-1}(\Delta)$)
- $r_2(L)$ deals with the graph

$$K = [K_1^{-1} + r_2(L)]^{-1}$$

Now use Woodbury matrix identity:

$$[K_1^{-1} + r_2(L)]^{-1} = K_1 - K_1[K_1 + r_2^{-1}(L)]^{-1}K_1$$

or elementwise:

$$k_g(x_1, x_2) = k_1(x_1, x_2) - k_1(x_1, X)^\top [K_1 + r_2^{-1}(L)]^{-1} k_1(X, x_2)$$

Kernels for Attributed Graphs

$$\Omega(||f||^2) = \langle f, [r_1(\Delta) + r_2(L)]f \rangle = \langle f, r_1(\Delta)f \rangle + \langle f, r_2(L)f \rangle$$

- $r_1(\Delta)$ deals with the feature data ($K_1 = r_1^{-1}(\Delta)$)
- $r_2(L)$ deals with the graph

$$K = [K_1^{-1} + r_2(L)]^{-1}$$

Now use Woodbury matrix identity:

$$[K_1^{-1} + r_2(L)]^{-1} = K_1 - K_1[K_1 + r_2^{-1}(L)]^{-1}K_1$$

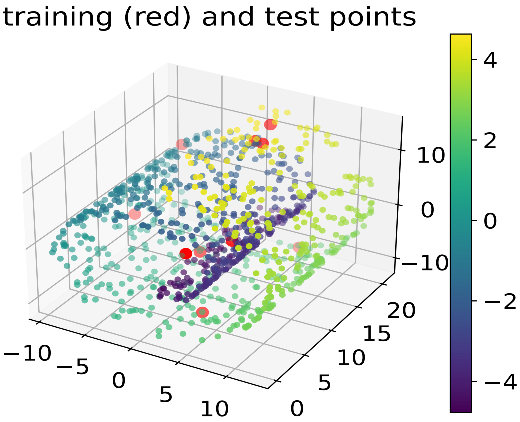
or elementwise:

$$k_g(x_1, x_2) = k_1(x_1, x_2) - k_1(x_1, X)^\top [K_1 + r_2^{-1}(L)]^{-1} k_1(X, x_2)$$

Note this depends on all the data - the transductive property

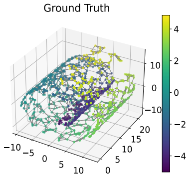
Synthetic Experiments

Swiss Roll regression, 1000 points, 10 training, 990 testing:

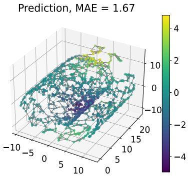


Synthetic Experiments

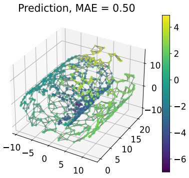
Predictions:



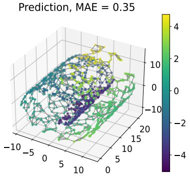
(a) Ground truth



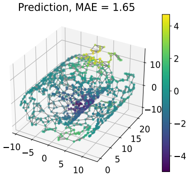
(b) GP with RBF



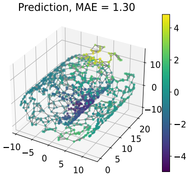
(c) Graph only



(d) TGGP (ours)



(e) GGP



(f) WGGP

Semi-Supervised Classification

Method	Texas	Cornell	Wisconsin	Chameleon	Cora	Citeseer	Squirrel	Actor
# Nodes	183	183	251	2,277	2,708	3,327	5,201	7,600
Homo. Rat.	0.11	0.30	0.21	0.23	0.81	0.74	0.22	0.22
GCN	59.5 ±5.3	57.0 ±4.7	59.8 ±7.0	59.8 ±2.6	80.5 ±0.8	68.1 ±1.3	36.9 ±1.3	30.3 ±0.8
GAT	58.4 ±4.5	58.9 ±3.3	55.3 ±8.7	54.7 ±2.0	**82.6 ±0.7**	**72.2 ±0.9**	30.6 ±2.1	26.3 ±1.7
ChebNet	77.3 ±4.1	**74.3 ±7.5**	79.4 ±4.5	55.2 ±2.8	78.0 ±1.2	70.1 ±0.8	43.9 ±1.6	34.1 ±1.1
LP	37.8	21.6	23.5	44.5	71.3	49.9	32.7	22.4
GP	78.4	73.0	78.4	46.1	60.8	54.7	34.4	**34.9**
GGP	78.4	62.1	60.8	**73.5**	80.9	69.7	**64.8**	26.3
ChebGP	**81.1**	64.9	**82.4**	**69.1**	79.7	66.5	28.8	31.8
WGGP	78.4	67.6	**84.3**	64.5	**84.7**	**70.8**	**58.3**	32.6
TGGP (ours)	**81.1**	**75.7**	**82.4**	63.2	80.3	70.5	53.8	**34.9**
GGP-X	78.4	56.8	60.8	**77.6**	84.7	75.6	**71.9**	OOM
WGGP-X	81.1	75.7	84.3	65.6	**87.5**	**76.8**	61.3	OOM
TGGP-X (ours)	**86.5**	**81.1**	**86.3**	63.4	83.8	76.7	54.2	**36.9**

Supplementary

Solving Kernels

For finding $K = r^{-1}(\Delta)$:

- Firstly, note
  $\Delta e^{-i\omega x} = -\omega^2 e^{-i\omega x} \implies r(\Delta)e^{-i\omega x} = \hat{g}(\omega)e^{-i\omega x}$.
- By Plancherel theorem,

$$\langle h, h\rangle_{\mathcal{H}} = \int |h(x)^2 g(x)^2| dx = \int |\hat{h}(\omega)^2 \hat{g}(\omega)^2| d\omega.$$

- The inverse is

$$(r^{-1}(\Delta)_{x^*,x}) = k(x^*, x) = \int \frac{1}{\hat{g}(\omega)} e^{i\omega(x^*-x)}\, d\omega.$$

Example

Gaussian regularizer:

$$r(\Delta) = \sum_{i=0}^{\infty} \frac{1}{i!} \left(\frac{\sigma^2 \Delta}{2} \right)^i \equiv \exp \left\{ \frac{\sigma^2 \Delta}{2} \right\}.$$

This leads to $\hat{g}(\omega) = \exp\{\sigma^2 \omega^2 / 2\}$.

Compute the inverse Fourier transform of the reciprocal:

$$\int_{-\infty}^{\infty} \exp \left\{ -\frac{\sigma^2 \omega^2}{2} \right\} e^{i\omega(\mathbf{x}-\mathbf{x}')} \, d\omega \propto \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{x} - \mathbf{x}')^2 \right\}.$$

Last term corresponds to the popular Gaussian RBF kernel.

Generalization

How are kernels, graph kernels, and graph GPs defined so far?

Model	**Name**	$r_1(\Delta)$	$r_2(\mathbf{L})$
Label Propagation	Label Propagation	0	$\frac{1}{1-\alpha}(\mathbf{I}+\alpha\mathbf{L})$
Kernels on graphs	Regularized Laplacian	0	$(\mathbf{I}+\sigma^2\mathbf{L})$
	Diffusion	0	$\exp\{\frac{\sigma^2}{2}\mathbf{L}\}$
	p-step random walk	0	$(\alpha\mathbf{I}-\mathbf{L})^{-p}$
	Cosine	0	$(\cos(\mathbf{L}\pi/4))^{-1}$
GP kernels	Matérn kernel on graphs	0	$(\frac{2\nu}{\kappa^2}-\mathbf{L})^{\nu/2+d/4}$
	Laplacian kernel	$1+\|\|\Delta\|\|^2$	0
	Gaussian kernel	$e^{\frac{\sigma^2}{2}\|\|\Delta\|\|^2}$	0
	Matérn kernel on manifolds	$(\frac{2\nu}{\kappa^2}-\Delta)^{\nu/2+d/4}$	0
Graph GP	GGP	$(\mathbf{P}^\top)^{-1}r(\Delta)\mathbf{P}^{-1}$	0
Wavelet Graph GP	WGGP	$(\mathbf{W}^\top)^{-1}r(\Delta)\mathbf{W}^{-1}$	0
Transductive kernel (ours)	TGGP (ours)	$(\frac{2\nu}{\kappa^2}-\Delta)^{\nu/2+d/4}$	$\mathbf{U}[\text{softplus}(\sum_i(\beta_i\mathbf{\Lambda}^i))]\mathbf{U}^\top$

(In our experiments, we chose the Matern kernel w.r.t. $r_1(\Delta)$ and a softplus polynomial for the graph kernel $r_2(L)$)