Kernels

Kernels on Graphs
Diffusion kernels

Sriniwas G. Surampudi
Avinash Sharma
CVIT, IIIT-H

October 5, 2016

Outline

Outline

Regression formulation
primal to dual

- consider a set of points $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top{}_{1\times d} \\ \vdots \\ \mathbf{x}_n^\top{}_{1\times d} \end{bmatrix}_{n\times d}$ and corresponding target values $\mathbf{y} = [y_1, \cdots, y_n]^\top_{n\times 1}$.
- problem is to find a $\mathbf{g}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$ minimizing $\mathcal{L}(\mathbf{g}) := \sum_{i=1}^n (y_i - \mathbf{g}(\mathbf{x}_i))^2$ s.t. $\lambda \|\mathbf{w}\|^2$ is minimum.
- the *primal* solution: $\lambda \mathbf{w} = \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\mathbf{w})^{-1} \mathbf{y}$
- When $\mathbf{w}$ in terms of samples: $\mathbf{w} = \sum_{i=1}^n \alpha_i \mathbf{x}_i$
- the *dual* solution:

$$\mathbf{w} = \mathbf{X}^\top \boldsymbol{\alpha}$$
$$\lambda \boldsymbol{\alpha} = \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\mathbf{w})$$
$$\boldsymbol{\alpha} = (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_n)^{-1} \mathbf{y}$$

Solution in terms of inputs!
concept of a kernel

- The **Kernel**
  $\mathbf{G} = \mathbf{X}\mathbf{X}^\top$, $\mathbf{G}_{i,j} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$
  Capturing *similarity* between every sample pair.
  Information from the training examples is given by the inner products between pairs of training points in the matrix $\mathbf{G}$.
- For a new sample:
  $\mathbf{g}(\mathbf{x}) = \langle \sum_{i=1}^{n} \alpha_i \mathbf{x}_i, \mathbf{x} \rangle = \sum_{i=1}^{n} \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle = \boldsymbol{\alpha}^\top \mathbf{k}$
  $k_i = \langle \mathbf{x}_i, \mathbf{x} \rangle$
  Information about a novel example $\mathbf{x}$ required by the predictive function is just the inner products between the training points and the new example $\mathbf{x}$.

Non-linear transformations
kernels at help!

- Can use the same setting to find a non-linear regresser $\mathbf{g}$.
- Non-linear transformation:
  $\phi : \mathbf{x} \mapsto \phi(\mathbf{x}) \in F$
  $\mathbf{G}_{i,j} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$
  $k_i = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle$
- Do not need to know the transformation as...
  $\mathbf{g}(\mathbf{x}) = \langle \sum_{i=1}^{n} \alpha_i \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle = \sum_{i=1}^{n} \alpha_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle$

Defining a kernel
formal definition

Sriniwas G. Surampudi
Avinash Sharma
CVIT, IIIT-H

Definition

A *kernel* is a function κ that for all $\mathbf{x}, \mathbf{z} \in \mathbf{X}$ satisfies

$$\kappa(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$$

where ϕ is a mapping from $\mathbf{X}$ to (inner product) feature space F

$$\phi : \mathbf{x} \mapsto \phi(\mathbf{x}) \in F$$

Recollecting...

- Samples are embedded into a vector space called the feature space.
- Linear relations are sought among the images of samples in the feature space.
- The algorithms are implemented in such a way that the coordinates of the embedded samples are not needed, only their pairwise inner product are.
- The pairwise inner products can be computed efficiently directly from the original data items using a kernel function.

Outline

Mercer's theorm
positive semi-definiteness

Theorem

Let $\mathbf{K} : \mathcal{R}^d \times \mathcal{R}^d \mapsto \mathcal{R}$ be given. Then for $\mathbf{K}$ to be a valid (Mercer) kernel, it is necessary and sufficient that for any $\mathbf{z}$, the corresponding kernel matrix is symmetric positive semi-definite.

Proof.

$$\begin{aligned}
\mathbf{z}^\top \mathbf{K} \mathbf{z} &= \sum_i \sum_j z_i \mathbf{K}_{i,j} z_j \\
&= \sum_i \sum_j \sum_k z_i \phi_k(\mathbf{x}_i) \phi_k(\mathbf{x}_j) z_j \\
&= \sum_k (\sum_i z_i \phi_k(\mathbf{x}_i))^2 \\
&\geq 0
\end{aligned}$$


Eigen decomposition
frequency components

- spectrum of the kernel:

$$\begin{aligned}\mathbf{K} &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \\ &= \sum_{i=1}^{n} \mathbf{u}_i \mathbf{u}_i^\top \lambda_i\end{aligned}$$

- any function $\mathbf{g}$ on $\mathbf{K}$:

$$\mathbf{g}(\mathbf{K}) = \mathbf{U}\mathbf{g}(\mathbf{\Lambda})\mathbf{U}^\top$$

- cholesky decomposition :

$$\mathbf{g}(\mathbf{K}) = \mathbf{L}\mathbf{L}'$$

Outline

Laplacian of a graph

- An undirected unweighted graph G consists of a set of vertices V numbered 1 to n, and a set of edges E (i.e., pairs (i,j) where $i, j \in V$ and $(i,j) \in E \iff (j,i) \in E$).
- Laplacian matrix:
  $\mathbf{L} := \mathbf{D} - \mathbf{W}$
  Normalozed Laplacian :
  $\hat{\mathcal{L}} := \mathbf{D}^{-0.5}(\mathbf{D} - \mathbf{W})\mathbf{D}^{-0.5}$

Theorem

$\hat{\mathcal{L}}$ is a symmetric positive semi definite matrix with eigenvalues $\lambda_1, \cdots, \lambda_n$ satisfy $0 \leq \lambda_i \leq 2$.

Laplace operator Δ to the graph Laplacian $\mathbf{L}$

5-point stencil approximation

Sriniwas G. Surampudi
Avinash Sharma
CVIT, IIIT-H

- Assume an m-dimensional regular grid and a function $\mathbf{f}$ on every node $\mathbf{x} = \sum_{i=1}^{m} x_i \mathbf{e}_i$

$$\begin{aligned}\Delta \mathbf{f}(x) &= \sum_{i=1}^{m} \frac{\partial^2}{\partial x_i^2} \mathbf{f} \\ &\approx \sum_{i=1}^{m} \frac{\frac{\partial}{\partial x_i}\mathbf{f}(\mathbf{x} - \frac{1}{2}\mathbf{e}_i) - \frac{\partial}{\partial x_i}(\mathbf{x} + \frac{1}{2}\mathbf{e}_i)}{\delta} \\ &\approx \sum_{i=1}^{m} \frac{\mathbf{f}(\mathbf{x} + \mathbf{e}_i) + \mathbf{f}(\mathbf{x} + \mathbf{e}_i) - 2\mathbf{f}(\mathbf{x})}{\delta^2} = -\frac{1}{\delta^2}\mathbf{L}\mathbf{f}\end{aligned}$$

- Both the continuous and the discrete Laplacians canonical operators, i.e., they are invariant under certain natural transformations of the underlying space, and hence are essentially unique.

Outline

Kernel as a function of $\hat{\mathcal{L}}$

- The manifold is approximated by the adjacency graph computed from the data points (samples). The Laplace Beltrami operator is approximated by the weighted Laplacian of the adjacency graph with weights chosen appropriately.
- In this sense, even Laplacian is a kernel capturing similarity between every sample pair.
- For any polynomial function : $\hat{\mathcal{L}}^p = \mathbf{U}\mathbf{\Lambda}^p\mathbf{U}^\top$
- From Taylor series :

$$\begin{aligned}\mathbf{g}(\hat{\mathcal{L}}) &= e^{-\hat{\mathcal{L}}t} \\ &= \mathbf{I} - t\hat{\mathcal{L}} + \frac{t^2}{2!}\hat{\mathcal{L}}^2 - \cdots \\ &= \mathbf{U}e^{-\mathbf{\Lambda}t}\mathbf{U}^\top\end{aligned}$$

Outline

Heat equation
exponential solution

- Differential equation:

$$\frac{d}{dt}\mathbf{K}_t = -\hat{\mathcal{L}}\mathbf{K}_t$$

- Solution:

$$\mathbf{K}_t = e^{-\hat{\mathcal{L}}t}$$

- Eigen-decomposition:
$\sum_{i=1}^{n} \mathbf{u}_i \mathbf{u}_i^\top e^{-\lambda_i t}$

Outline

A stochastic model

- Consider the random field obtained by attaching independent, zero mean, variance σ^2, random variables $Z_i(0)$ to each vertex $i \in V$.
- Let each of these Z_i's *send* a fraction $\alpha << 1$ to their neighbors at discrete time steps $t = 0, 1, 2, \cdots$.

$$Z_i(t+1) = Z_i(t) + \alpha \sum_{j \in V : j \sim i} (Z_j(t) - Z_i(t))$$

- Let

$$T(t) = (1 - \alpha \mathbf{L})^t$$

Then

$$\mathbf{Z}(t) = T(t)\mathbf{Z}(0)$$

Covariance matrix

- Covaiance between node pair:

$$\begin{aligned}
\mathbf{E} &= [(Z_i(t) - \mathbf{E}[Z_i(t)])(Z_j(t) - \mathbf{E}[Z_j(t)])] \\
&= \mathbf{E}[Z_i(t)Z_j(t)] \\
&= \mathbf{E}[(\sum_{i'} T_{ii'}(t)Z_{i'}(0))(\sum_{j'} T_{jj'}(t)Z_{j'}(0))] \\
&= \sigma^2 \sum_k T_{i,k}(t)\,T_{k,j}(t) \\
&= \sigma^2[T(t)]^2_{i,j} = \sigma^2[T(2t)]_{i,j}
\end{aligned}$$

Limit as the kernel

$$\lim_{\delta t \to 0} T(t) = \lim_{\delta t \to 0} (1 - \frac{\alpha \mathbf{L}}{\frac{1}{\delta t}})^{\frac{t}{\delta t}}$$
$$= \sigma^2 e^{-2\alpha t \mathbf{L}}$$