AI6102: Machine Learning Methodologies & Applications

Additional Notes on Kernelized Regularized Regression Models

Sinno Jialin Pan

Nanyang Technological University, Singapore

Homepage: http://www.ntu.edu.sg/home/sinnopan

NANYANG TECHNOLOGICAL UNIVERSITY

Kernelized Regularized Regression

- Regularized linear regression with kernels

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}\cdot\phi(\boldsymbol{x}_i) - y_i)^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$$

- It can be reformulated in terms of a dual form where kernel function arises naturally

$$\frac{\partial\left(\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}\cdot\phi(\boldsymbol{x}_i) - y_i)^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2\right)}{\partial\boldsymbol{w}} = \boldsymbol{0}$$

$$\Downarrow$$

$$\sum_{i=1}^{N}(\boldsymbol{w}\cdot\phi(\boldsymbol{x}_i) - y_i)\phi(\boldsymbol{x}_i) + \lambda\boldsymbol{w} = 0$$

Closed-form Solution

- By using kernel trick $k(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$, the regression prediction function $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x})$ can be written as

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{k}_{\boldsymbol{x}}^{T}(\mathbf{K} + \lambda\mathbf{I})^{-1}\boldsymbol{y}$$

$$\mathbf{K} = \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_N) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_N, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_N, \boldsymbol{x}_N) \end{pmatrix} \begin{pmatrix} \phi(\boldsymbol{x}_1) \cdot \phi(\boldsymbol{x}_1) & \cdots & \phi(\boldsymbol{x}_1) \cdot \phi(\boldsymbol{x}_N) \\ \vdots & \ddots & \vdots \\ \phi(\boldsymbol{x}_N) \cdot \phi(\boldsymbol{x}_1) & \cdots & \phi(\boldsymbol{x}_N) \cdot \phi(\boldsymbol{x}_N) \end{pmatrix}$$

$$\boldsymbol{k}_{\boldsymbol{x}} \quad \begin{array}{|c|} \hline k(\boldsymbol{x}_1, \boldsymbol{x}) \\ \hline k(\boldsymbol{x}_2, \boldsymbol{x}) \\ \hline \dots \\ \hline k(\boldsymbol{x}_N, \boldsymbol{x}) \\ \hline \end{array} \qquad \boldsymbol{y} \quad \begin{array}{|c|} \hline y_1 \\ \hline y_2 \\ \hline \dots \\ \hline y_N \\ \hline \end{array} \qquad \mathbf{I}\ (N \times N) \quad \begin{array}{|c|c|c|c|} \hline 1 & 0 & \dots & 0 \\ \hline 0 & 1 & \dots & 0 \\ \hline \dots & \dots & \dots & \dots \\ \hline 0 & 0 & \dots & 1 \\ \hline \end{array}$$

Derivation

$$\sum_{i=1}^{N}(\boldsymbol{w}\cdot\phi(\boldsymbol{x}_i)-y_i)\phi(\boldsymbol{x}_i)+\lambda\boldsymbol{w}=0$$

$$\Downarrow$$

$$\boldsymbol{w}=-\frac{1}{\lambda}\sum_{i=1}^{N}(\boldsymbol{w}\cdot\phi(\boldsymbol{x}_i)-y_i)\phi(\boldsymbol{x}_i)$$

By denoting $a_i=-\frac{1}{\lambda}(\boldsymbol{w}\cdot\phi(\boldsymbol{x}_i)-y_i)$

$$=\sum_{i=1}^{N}a_i\phi(\boldsymbol{x}_i)$$

a_i's are dual variables

Derivation (cont.)

Denote by $\boldsymbol{\Phi} = \left(\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2), \dots, \phi(\boldsymbol{x}_N)\right)^T$ and $\boldsymbol{a} = (a_1, \dots, a_N)^T$

$\boldsymbol{\Phi}$: N-by-F matrix (suppose $\phi(\cdot)$ maps $\boldsymbol{x}_i$ to a F-dimensional space)

$\boldsymbol{a}$: a column vector

$$\boldsymbol{w} = \sum_{i=1}^{N} a_i \phi(\boldsymbol{x}_i) = \boldsymbol{\Phi}^T \boldsymbol{a}$$

$$a_i = -\frac{1}{\lambda}(\boldsymbol{w} \cdot \phi(\boldsymbol{x}_i) - y_i) \Longrightarrow \boldsymbol{a} = \frac{1}{\lambda}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w})$$

Derivation (cont.)

- By substituting $\boldsymbol{w} = \boldsymbol{\Phi}^T \boldsymbol{a}$ into $\boldsymbol{a} = \frac{1}{\lambda}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w})$, we have

$$\boldsymbol{a} = \frac{1}{\lambda}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\Phi}^T \boldsymbol{a})$$

$$\Downarrow$$

$$\lambda \mathbf{I} \boldsymbol{a} = \boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\Phi}^T \boldsymbol{a}$$

$$\Downarrow$$

$$(\lambda \mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Phi}^T)\boldsymbol{a} = \boldsymbol{y}$$

$$\Downarrow$$

$$\boldsymbol{a} = (\lambda \mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Phi}^T)^{-1} \boldsymbol{y}$$

Closed-form solution of $\boldsymbol{a}$

Derivation (cont.)

- By denoting the Kernel matrix (Gram matrix) as follows,

$$\mathbf{K} = \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_N) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_N, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_N, \boldsymbol{x}_N) \end{pmatrix} = \begin{pmatrix} \phi(\boldsymbol{x}_1) \cdot \phi(\boldsymbol{x}_1) & \cdots & \phi(\boldsymbol{x}_1) \cdot \phi(\boldsymbol{x}_N) \\ \vdots & \ddots & \vdots \\ \phi(\boldsymbol{x}_N) \cdot \phi(\boldsymbol{x}_1) & \cdots & \phi(\boldsymbol{x}_N) \cdot \phi(\boldsymbol{x}_N) \end{pmatrix}$$

$$\boldsymbol{a} = (\lambda\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\Phi}^T)^{-1}\boldsymbol{y} = (\lambda\mathbf{I} + \mathbf{K})^{-1}\boldsymbol{y}$$

$$\boldsymbol{w} = \boldsymbol{\Phi}^T\boldsymbol{a} = \boxed{\boldsymbol{\Phi}^T}(\lambda\mathbf{I} + \mathbf{K})^{-1}\boldsymbol{y}$$

Unknown as $\phi(\cdot)$ is unknown

Derivation (cont.)

- Our goal is to construct an explicit form of $f(\cdot)$ not $\boldsymbol{w}$

$$f(\boldsymbol{x}) = \boldsymbol{w} \cdot \phi(\boldsymbol{x}) = \phi(\boldsymbol{x})^T \boldsymbol{w}$$

- By substituting $\boldsymbol{w} = \boldsymbol{\Phi}^T(\lambda \mathbf{I} + \mathbf{K})^{-1}\boldsymbol{y}$ into $f(\boldsymbol{x})$, we have

$$f(\boldsymbol{x}) = \boxed{\phi(\boldsymbol{x})^T \boldsymbol{\Phi}^T}(\mathbf{K} + \lambda \mathbf{I})^{-1}\boldsymbol{y}$$

Denote $\boldsymbol{k}_{\boldsymbol{x}} = \boldsymbol{\Phi}\phi(\boldsymbol{x})$, which is an N-dimensional column vector, where the i-th element of $\boldsymbol{k}_{\boldsymbol{x}}$ is $k(x_i, \boldsymbol{x})$

- Therefore

$$f(\boldsymbol{x}) = \boldsymbol{k}_{\boldsymbol{x}}^T(\mathbf{K} + \lambda \mathbf{I})^{-1}\boldsymbol{y}$$

Thank you!