Machine Learning 1 Homework 13 Theory Part

01.02.2016

Kernel Ridge Regression

a) The cost function of Ridge Regression is defined as:

$$\mathcal{E}_{RR}(\mathbf{w}) = (\mathbf{y} - \mathbf{w^T X})^2 + \lambda ||\mathbf{w}||^2$$

$\mathbf{X}$: Matrix of input data points from training dataset, each column of it is a data point $\mathbf{x_i} \in \mathbb{R}^d$
$\mathbf{y}$: Vector of labels from training dataset
Take the derivative w.r.t $\mathbf{w}$ yields:

$$\frac{\partial \mathcal{E}_{RR}(\mathbf{w})}{\partial \mathbf{w}} = -2\mathbf{Xy} + 2\mathbf{XX^T w} + 2\lambda \mathbf{w}$$

We want to optimize this cost function, so set the derivative to 0, we get:

$$\begin{aligned}
\mathbf{XX^T w} + \lambda \mathbf{w} &= \mathbf{Xy} \\
(\mathbf{XX^T} + \lambda \mathbf{I})\mathbf{w} &= \mathbf{Xy} \\
\mathbf{w} &= (\mathbf{XX^T} + \lambda \mathbf{I})^{-1} \mathbf{Xy}
\end{aligned}$$

Now we get the solution for Ridge regression.

b) To kernelize this, we introduce kernel function: $k(\mathbf{x_i}, \mathbf{x_j})$ which maps the innerproduct of two datapoints to another feature space.
The mapping can be denoted as:

$$\begin{aligned}
\mathbf{x} &\rightarrow \mathbf{\Phi}(\mathbf{x}) \\
< \mathbf{x_i}, \mathbf{x_j} > &\rightarrow k(\mathbf{x_i}, \mathbf{x_j}) = \mathbf{\Phi}(\mathbf{x_i})^T \mathbf{\Phi}(\mathbf{x_j})
\end{aligned}$$

Thus we get the kernelized Ridge regression cost function:

$$\mathcal{E}_{RR}(\mathbf{w}) = (\mathbf{y} - \mathbf{w^T \Phi(X)})^2 + \lambda ||\mathbf{w}||^2$$

Take the derivative w.r.t $\mathbf{w}$ and set it to 0, we get the solution:

$$\mathbf{w} = \left(\mathbf{\Phi(X)\Phi(X)^T} + \lambda \mathbf{I}\right)^{-1} \mathbf{\Phi(X)y}$$

Combined with the linear assumption, we can get the estimated value given $\mathbf{x}$ as input:

$$\begin{aligned}
\hat{y} &= \Phi(\mathbf{x})^T \mathbf{w} \\
&= \Phi(\mathbf{x})^T \left(\mathbf{\Phi(X)\Phi(X)^T} + \lambda \mathbf{I}\right)^{-1} \mathbf{\Phi(X)y} \\
&= \Phi(\mathbf{x})^T \left(\mathbf{\Phi(X)\Phi(X)^T} + \lambda \mathbf{I}\right)^{-1} \mathbf{\Phi(X)} \left(\mathbf{\Phi(X)\Phi(X)^T} + \lambda \mathbf{I}\right) \left(\mathbf{\Phi(X)\Phi(X)^T} + \lambda \mathbf{I}\right)^{-1} \mathbf{y} \\
&= \Phi(\mathbf{x})^T \left(\mathbf{\Phi(X)\Phi(X)^T} + \lambda \mathbf{I}\right)^{-1} \left(\mathbf{\Phi(X)\Phi(X)\Phi(X)^T} + \lambda \mathbf{I\Phi(X)}\right) \left(\mathbf{\Phi(X)\Phi(X)^T} + \lambda \mathbf{I}\right)^{-1} \mathbf{y} \\
&= \Phi(\mathbf{x})^T \left(\mathbf{\Phi(X)\Phi(X)^T} + \lambda \mathbf{I}\right)^{-1} \left(\mathbf{\Phi(X)\Phi(X)^T} + \lambda \mathbf{I}\right) \mathbf{\Phi(X)} \left(\mathbf{\Phi(X)\Phi(X)^T} + \lambda \mathbf{I}\right)^{-1} \mathbf{y} \\
&= \Phi(\mathbf{x})^T \mathbf{\Phi(X)} \left(\mathbf{\Phi(X)\Phi(X)^T} + \lambda \mathbf{I}\right)^{-1} \mathbf{y} \\
&= \mathbf{k}^* \left(\mathbf{K} + \lambda \mathbf{I}\right)^{-1} \mathbf{y} \\
\mathbf{k}^* &= \Phi(\mathbf{x})^T \mathbf{\Phi(X)} \\
\mathbf{K} &= \mathbf{\Phi(X)\Phi(X)^T}
\end{aligned}$$

Since $(\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{y}$ can be calculated using training dataset, this part can be seen as coefficients α. $\mathbf{k}^*$ is a vector which contains the mapped innerproduct of input data point $\mathbf{x}$ with all the training data points. So this result can be written as the inner product of this vector $\mathbf{k}^*$ with coefficients α. Thus we get the formula:

$$\hat{y} = \sum_{i=2}^{n} k(\mathbf{x}, \mathbf{x_i})\alpha_i$$

c) Consider the new prime problem:

$$\begin{aligned} &\min_{\xi,\mathbf{w}} \sum_{i=1}^{n} \xi_i^2 \\ subject \quad to \quad &\xi_i = \mathbf{w^T x_i} - y_i \\ &||\mathbf{w}||^2 \leq C \end{aligned}$$

Apply Lagrangian multiplier method:

$$\mathcal{L} = \sum_{i=1}^{n} \xi_i^2 + \sum_{i=1}^{n} \lambda_i(\mathbf{w^T x_i} - y_i - \xi_i) - \alpha(||\mathbf{w}||^2 - C)$$

Calculate derivatives w.r.t $\mathbf{w}$ and ξ, we get:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \xi_i} &= 2\xi_i - \lambda_i = 0 \\ \frac{\partial \mathcal{L}}{\partial \mathbf{w}} &= \sum_{i=1}^{n} \lambda_i \mathbf{x_i} - 2\alpha\mathbf{w} = 0 \\ \Rightarrow \qquad \xi_i &= \frac{1}{2}\lambda_i \\ \mathbf{w} &= \frac{1}{2\alpha}\sum_{i=1}^{n} \lambda_i \mathbf{x_i} \end{aligned}$$

insert the result to the primal problem, we get a dual problem:

$$\begin{aligned} &\max_{\lambda,\alpha} \sum_{i=1}^{n} \lambda_i^2 \\ subject \quad to \quad &\frac{1}{4\alpha^2}(\sum_{i=1}^{n} \lambda_i \mathbf{x_i}) \leq C \\ &\alpha \geq 0 \end{aligned}$$

This problem has an additional constraints than the original quadratic problem but different from ridge regression. It can use kernel trick as the solution contains inner product.