Vinay Gaba
UNI: vhg2105

COMS 4771: Machine Learning Assignment 4

1. We are given the following training set:

$$S = \{(x1, y1), (x2, y2) \dots \dots \dots (xn, yn) \in \mathbb{R}^p \; x \; \mathbb{R}\}$$

and we are told that w_λ the ridge regression solution with regularization parameter λ.

$$\widehat{w}_\lambda = \arg\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle w, x_i \rangle)^2 + \lambda ||w||_2{}^2$$

Extra Credit (2%)

We are given the value of $\widehat{w}_\lambda \; as$:

$$\widehat{w}_\lambda = X^T (XX^T + n\lambda I_n)^{-1} y$$

This can be rewritten as:

$$(XX^T + n\lambda I_n)\widehat{w}_\lambda = X^T y$$

This can be rewritten as:

$$n\lambda \widehat{w}_\lambda = X^T (y - X\widehat{w}_\lambda) \qquad \text{--}(1)$$

We also know that:

$$\alpha \coloneqq y - X\widehat{w}_\lambda$$

$We\ will\ use\ equation\ 1\ and\ \alpha\ to\ show\ that$

$$\frac{1}{n\lambda} XX^T \alpha + \alpha = y$$

$$\frac{1}{n\lambda} XX^T (y - X\widehat{w}_\lambda) = y - (y - X\widehat{w}_\lambda) \text{ [Substituting } \alpha]$$

$$XX^T (y - X\widehat{w}_\lambda) = n\lambda \, X\widehat{w}_\lambda$$

$$X^{-1} XX^T (y - X\widehat{w}_\lambda) = n\lambda \, X^{-1} X\widehat{w}_\lambda$$

$$X^T (y - X\widehat{w}_\lambda) = n\lambda \, \widehat{w}_\lambda \quad (\dots X^{-1} X = I)$$

$$X^T (y - X\widehat{w}_\lambda) = n\lambda \widehat{w}_\lambda$$

Hence proved

Extra Credit (1%)

Solve equation (2) for α, and use your solution in equation (1) to get the following alternative expression for $\widehat{w_\lambda}$:

$$\widehat{w}_\lambda = \frac{1}{n\lambda} X^T \left(\frac{1}{n\lambda} XX^T + I_n\right)^{-1} y = X^T (XX^T + n\lambda I_n)^{-1} y$$

$$\frac{1}{n\lambda} XX^T \alpha + \alpha = y$$

$$\alpha(\frac{1}{n\lambda} XX^T + I) = y$$

$$\alpha = (\frac{1}{n\lambda} XX^T + I)^{-1} y$$

$$n\lambda \widehat{w}_\lambda = X^T (y - X\widehat{w}_\lambda) [\text{Using eq 1}]$$

$$n\lambda \widehat{w}_\lambda = X^T \alpha \; [\text{Substituting value of } \alpha]$$

$$\widehat{w}_\lambda = \frac{1}{n\lambda} X^T \alpha$$

Resubstituing α back in the equation above:

$$\widehat{w}_\lambda = \frac{1}{n\lambda} X^T (\frac{1}{n\lambda} XX^T + I)^{-1} y$$

$$= X^T (XX^T + n\lambda I_n)^{-1} y$$

This is the alternative value of $\widehat{w}_\lambda$

Part 1(weightage 4%)

As we solved in the extra credit part, $\widehat{w}_\lambda = X^T (XX^T + n\lambda I_n)^{-1} y$. We will be using this in our proof below.

We know that labeled points (x, y) are drawn from $X \ x \ \mathbb{R}$, where X is the feature space.

We also know that the kernel function is $k(x, x') = \langle \phi(x), \phi(x') \rangle$

Let $\widehat{w}_\lambda$ be the solution of the ridge regression problem with regularization parameter λ when data points are mapped to H using the feature mapping ϕ. It is given by the following formula:

$$\widehat{w}_\lambda = \arg\min_{w \in \mathbb{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle w, \phi(x_i) \rangle)^2 + \lambda ||W||_2{}^2$$

We will try to focus on the minimization aspect in the above equation. We can vectorize the above in the following way:

$$\begin{aligned}
R(W) &= \left(\frac{1}{n} \sum_{i=1}^{n} (y - \phi(x_i)w)^2 \right) + \lambda ||W||_2{}^2 \\
&= \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - \phi(x_i)w)^2 \right) + \lambda ||W||_2{}^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & \phi(x_i) \cdots \\ \vdots & \vdots \\ 1 & \phi(x_n) \ldots \end{bmatrix} \begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_n \end{bmatrix} \right)^2 \\
&\qquad + \lambda ||W||_2{}^2 \\
&= \frac{1}{n} \left\| \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & \phi(x_i) \cdots \\ \vdots & \vdots \\ 1 & \phi(x_n) \ldots \end{bmatrix} \begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_n \end{bmatrix} \right\|^2 + \lambda ||W||_2{}^2 \\
&= \frac{1}{n} \|Y - \phi(X)W\|^2 + \lambda ||W||_2{}^2
\end{aligned}$$

To minimize the equation wrt W, we solve to gradient = 0

$$\nabla_W R = 0$$

$$\nabla_w \left(\frac{1}{n} \|Y - \phi(X)W\|^2 + \lambda ||W||_2{}^2 \right) = 0$$

$$\frac{1}{n} \nabla_w ((Y - \phi(X)W)^T (Y - \phi(X)W)) + 2\lambda ||W|| = 0$$

$$\frac{1}{n} \nabla_w (Y^T Y - 2Y^T \phi(X)W + W^T \phi(X)^T \phi(X)W) + 2\lambda ||W|| = 0$$

$$\frac{1}{n} \left(-2Y^T \phi(X) + 2W^T \phi(X)^T \phi(X) \right) + 2\lambda ||W|| = 0$$

$$(-2Y^T \phi(X) + 2W^T \phi(X)^T \phi(X)) + 2n\lambda W = 0$$

$$2W(\phi(X)\phi(X)^T + n\lambda I) = 2Y^T \phi(X)$$

$$W = \phi(X)^T(\phi(X)\phi(X)^T + n\lambda I)^{-1}Y$$

We will use the above minimization of $\widehat{w}_\lambda$ and $\phi(X)$ to compute the prediction for any given point.

$$\langle \phi(X), \widehat{w}_\lambda \rangle = \phi(X)\phi(X)^T(\phi(X)\phi(X)^T + n\lambda I)^{-1}Y$$

The above equation can be rewritten using the kernel function as $\phi(X)\phi(X)^T$ can be kernalized. Thus, the prediction can be calculated without calculating the values of $\phi(X)$ & $\phi(X)^T$separately.

2. *Extra Credit:*

In order to solve this part, we use the connection between PCA and SVD to derive a kernelized form of PCA.

We know that X is the design matrix and $X = USV^T$ is the SVD of X.

Also, $X^TX = V\,S^2V^T$

We are also given the following facts:

- The columns of U are the eigenvectors of XX^T , with eigenvalues given by the squares of the diagonal entries in S.

Since we are told that X is a design matrix that is transformed from the training set S', it can be represented in the following form:

$$X := \begin{bmatrix} \cdots & \phi(x_1) & \cdots \\ \cdots & \phi(x_2) & \cdots \\ \cdots & \vdots & \cdots \\ \cdots & \phi(x_n) & \cdots \end{bmatrix}$$

$$X' := \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ \phi(x_1) & \phi(x_2) & \cdots & \phi(x_n) \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

We can compute X^TX in the following way:

$$XX^T = \begin{bmatrix} \cdots & \phi(x_1) & \cdots \\ \cdots & \phi(x_2) & \cdots \\ \cdots & \vdots & \cdots \\ \cdots & \phi(x_n) & \cdots \end{bmatrix} \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ \phi(x_1) & \phi(x_2) & \cdots & \phi(x_n) \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

$$XX^T = \begin{bmatrix} \cdots & < \phi(x_1), \phi(x'_1) > & \cdots \\ \cdots & < \phi(x_2), \phi(x'_2) > & \cdots \\ \cdots & \vdots & \cdots \\ \cdots & < \phi(x_n), \phi(x'_n) > & \cdots \end{bmatrix}$$

XX^T consists of rows of dot products of $\phi(x_1)$ and $\phi(x'_1)$. As we saw in previous classes, we can easily replace this using a kernel function $k\ (x, x')$. We can use this matrix to compute its eigenvectors by choosing the top l values. This gives us U_l. As you might have noticed, we were able to compute this value without computing the individual values of ϕ.

In order to compute the value of S, we will compute the eigenvectors of XX^T, take its square root and choose the top l values and fill the diagonals to matrix S to give us the value of S_l.

Part 2

Suppose you have computed S_l. and U_l. Specify the dimensions of these two matrices. Explain how you can compute the rank l PCA of the transformed training set using these matrices and the kernel function k without having to explicitly compute the mapping φ for any training points.

We know that the rank PCA representation of the training set is given by the columns of

$$S_l^{-1} U_l^{\ T} XX^T$$

From the above equation, we already have computed the values of S & U.

We know the following things about the dimension of each of the matrices:

Matrix	**Dimension**
S_l	$l * l$
U_l	$n * l$
XX^T	$n * n$

The value of XX^T is computed using the kernel function that we saw above. We then use all the values we computed in the equation $S_l^{-1} U_l^{\ T} XX^T$ to give us the rank of the PCA. We were able to achieve this without computing the value of ϕ for individual points.

Collaborators

Dikha Vanvari dhv2108
Varun Shetty vs2567