Optimization task for kernel ridge regression

Victor Kitov

1 Usual solution

Ridge regression criterion

$$\sum_{n=1}^{N}\left(x_n^T\beta - y_n\right)^2 + \lambda\beta^T\beta \to \min_{\beta}$$

Stationarity condition can be written as:

$$2\sum_{n=1}^{N}x_n\left(x_n^T\beta - y_n\right) + 2\lambda\beta = 0$$

$$2X^T(X\beta - Y) + \lambda\beta = 0$$

$$\left(X^TX + \lambda I\right)\beta = X^TY$$

so

$$\widehat{\beta} = (X^TX + \lambda I)^{-1}X^TY$$

Complexity:

- training:

operation	**complexity**
X^TX	$O(D^2N)$
$+\lambda I$	$O(D)$
$(X^TX + \lambda I)^{-1}$	$O(D^3)$
X^TY	$O(DN)$
$(X^TX + \lambda I)^{-1}X^TY$	$O(D^2)$
total	$O(D^2N + D^3)$

- prediction: $\widehat{y}(x) = x^T \beta$, complexity $O(D)$.

2 Task for alternative solution in terms of scalar products

Derive solution for ridge regression: $\widehat{y}(x) = x^T w$ that would allow kernel trick. To do this rewrite the standard optimization task

$$\sum_{n=1}^{N} \left(x_n^T w - y_n\right)^2 + \lambda w^T w \to \min_w \tag{1}$$

in equivalent way:

$$\begin{cases} \frac{1}{2} \left\| z \right\|^2 + \frac{1}{2} \lambda \left\| w \right\|^2 \to \min_{w,z} \\ z_i = x_i^T w - y_i & n = \overline{1, N} \end{cases}$$

1. Write out Largrangian optimization (using the method of Lagrange multipliers).
2. From stationarity condition of Lagrangian w.r.t z, w find z, w in terms of dual variables and substitute them back into Lagrangian optimization task to obtain so called *dual optimization problem*.
3. Solve dual optimization problem in matrix form (you will need to introduce matrix $\{M\}_{i,j} = x_i^T x_j$) and explain, why it's solution depends only on scalar products.
4. Assuming dual variables are found, write out how prediction $\widehat{y}(x)$ depends only on scalar products.
5. Apply kernel trick:
   (a) rewrite solution for dual variables in terms of arbitrary kernels
   (b) assuming dual variables are found, rewrite prediction in terms of arbitrary kernels
6. Compare complexity of making prediction for single object (assuming model is already fitted) for
   (a) standard approach (direct solution to 1 from the lectures)
   (b) proposed approach, depending only on scalar products.

3 Solution derivation

Lagrangian becomes

$$L = \frac{1}{2}z^T z + \frac{1}{2}\lambda w^T w + \sum_i \alpha_i \left(x_i^T w - y_i - z_i\right)$$

$$\begin{aligned}
\frac{\partial L}{\partial w} &= \lambda w + \sum_i \alpha_i x_i = 0 \\
\frac{\partial L}{\partial z_i} &= z_i - \alpha_i = 0
\end{aligned}$$

It follows that $z_i = \alpha_i$ and $w = -\frac{1}{\lambda}\sum_i \alpha_i x_i$. Substituting these equation into Lagrangian we obtain a dual task (in terms of dual variables α):

$$\begin{aligned}
L &= \frac{1}{2}\alpha^T\alpha + \frac{1}{2}\lambda\frac{1}{\lambda^2}\sum_{i,j}\alpha_i\alpha_j x_i^T x_j + \sum_i \alpha_i \sum_j \alpha_j x_j^T x_i \left(-\frac{1}{\lambda}\right) - \sum_i \alpha_i \left(y_i + \alpha_i\right) \\
&= \frac{1}{2\lambda}\sum_{i,j}\alpha_i\alpha_j x_i^T x_j - \frac{1}{\lambda}\sum_{i,j}\alpha_i\alpha_j x_i^T x_j - \frac{1}{2}\alpha^T\alpha - \sum_i \alpha_i y_i \\
&= -\frac{1}{2\lambda}\sum_{i,j}\alpha_i\alpha_j x_i^T x_j - \sum_i \alpha_i y_i - \frac{1}{2}\alpha^T\alpha \to extr_\alpha
\end{aligned}$$

By changing sign we obtain

$$\frac{1}{2\lambda}\sum_{i,j}\alpha_i\alpha_j x_i^T x_j + \frac{1}{2}\alpha^T\alpha + \sum_i \alpha_i y_i \to extr_\alpha$$

By introducing Gramm matrix $M \in \mathbb{R}^{NxN}$, defined as $\{M\}_{i,j} = x_i^T x_j$ we can rewrite the problem in matrix form:

$$Q = \frac{1}{2\lambda}\alpha^T M\alpha + \frac{1}{2}\alpha^T\alpha + \alpha^T y \to extr_\alpha$$

$$\frac{dQ}{d\alpha} = \frac{1}{\lambda}M\alpha + \alpha + y = 0$$

This is equivalent to

$$\left(\frac{1}{\lambda}M + I\right)\alpha = -y \quad \Longrightarrow \quad \alpha = -\left(\frac{1}{\lambda}M + I\right)^{-1} y$$

Complexity:

- training

operation	complexity
M	$O(N^2D)$
$\frac{1}{\lambda}M$	$O(N^2)$
$\frac{1}{\lambda}M+I$	$O(N)$
$\left(\frac{1}{\lambda}M+I\right)^{-1}$	$O(N^3)$
$-\left(\frac{1}{\lambda}M+I\right)^{-1}y$	$O(N^2)$
total	$O(N^2D+N^3)$

- prediction $\widehat{y}(x) = x^T w = -\frac{1}{\lambda}\sum_i \alpha_i x_i^T x$, complexity $O(DN)$.

Advantages:

- We have analytic solution for $\alpha =>$ fast training of the method.

- Solution always exists because Gramm matrix is positive-semi definite, because

$$\alpha^T M \alpha = \sum_{i,j} \alpha_i \alpha_j x_i^T x_j = \left(\sum_i \alpha_i x_i\right)^T \left(\sum_j \alpha_j x_j\right) = \left\|\sum_i \alpha_i x_i\right\|^2 \geq 0 \, \forall \alpha \in \mathbb{R}^N$$

  $\lambda > 0$, so $\frac{1}{\lambda}M + I$ is positive definite, thus non-degenerate.

Disadvantage:

Prediction becomes $\widehat{y}(x) = w^T x = -\frac{1}{\lambda}\sum_i \alpha_i x_i^T x$. Vector α is non-sparse, so it takes $O(ND)$ time to make a prediction.

4 Kernel trick

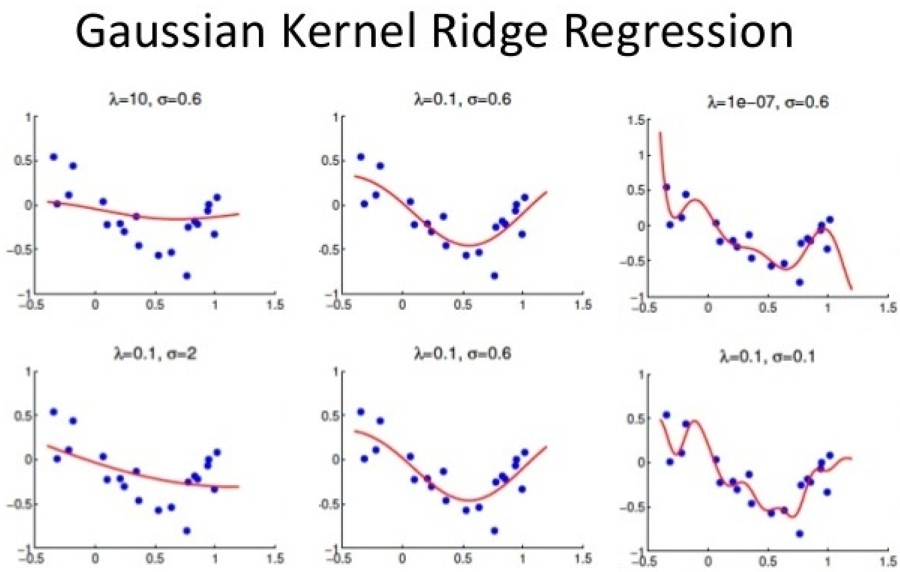
Both α and predition depend only on scalar products. So we may apply kernel trick. Let $x \to \phi(x)$. Scalar product $\langle x, x' \rangle$ corresponds to standard scalar product in transformed space $\langle \phi(x), \phi(x') \rangle = K(x, x')$.

Gramm matrix becomes $\{M\}_{i,j} = K(x_i, x_j)$, α is determined with new Gramm matrix $\alpha = \left(\frac{1}{\lambda}M + I\right)^{-1} y$ and prediction is made with

$$\widehat{y}(x) = \langle w, x \rangle = -\frac{1}{\lambda}\sum_i \alpha_i \langle x_i, x \rangle = -\frac{1}{\lambda}\sum_i \alpha_i K(x_i, x)$$

Consider Gaussian kernel

$$K(x, x') = e^{-\frac{\|x - x'\|^2}{2\sigma^2}}$$



Decreasing λ or decreasing σ leads to more complex model in ridge regression with Gaussian (RBF) kernel.