SVR: Purpose of KKT & Dual

Optimization in example space

Consider: Φ^T $\xrightarrow{\text{Dual view}}$ $f(x) = \sum_{j=1}^{m} \alpha_j \phi^T(x)\phi(x_j)$

$+ y_j$: - - - - -

↓

for b

Primal view

↓

Optimizing in feature space

$$\left.\begin{array}{c|cccc|} & \alpha_1, \alpha_1^* & \alpha_2, \alpha_2^* & & \alpha_m, \alpha_m^* \\ \hline w_1 & \phi(x_1) & \phi(x_2) & \cdots & \phi(x_m) \\ w_2 & | & | & & | \\ \vdots & | & | & & | \\ w_n & & & & \end{array}\right.$$

$$f(x) = w^T\phi(x) + b = \sum_{i=1}^{n} w_i \phi_i(x) + b$$

Tutorial 5

Thursday 1st September, 2016

Importance of equivalence: Instead of tuning ξ on "tuning set", tune λ on "tuning set"

Problem 1. Relation between Penalized Ridge Regression (λ) and Constrained Ridge Regression (θ): (Solution also recorded as part of lecture 13 video)

Show that the solution to the *Constrained Ridge Regression* problem

$$\mathbf{w}_{Con} = \underset{\mathbf{w}}{\operatorname{argmin}} \ ||\phi\mathbf{w} - \mathbf{y}||_2^2 \rightarrow ①$$

such that $||\mathbf{w}||_2^2 \leq \xi$ $\Rightarrow \lambda(||w||_2^2 - \xi)$ penalty in Lagrange fn.

is the same as that to the solution to *Penalized Ridge Regression*.

$$\mathbf{w}_{Pen} = \underset{\mathbf{w}}{\operatorname{argmin}} \ ||\phi\mathbf{w} - \mathbf{y}||_2^2 + \lambda||\mathbf{w}||_2^2 \rightarrow ②$$

for some λ that is a function of ξ.

Hint: You can make convexity assumptions and use KKT conditions.

Solution Sketch:

'Regularized' Linear Regression

- Consider the formulation in which we limit the weights of the coefficients by putting a constraint on size of the L2 norm of the weight vector:

$$\textcircled{h} \qquad \left.\begin{matrix}\operatorname{argmin}_{\mathbf{w}}(\mathbf{\Phi w} - \mathbf{y})^T(\mathbf{\Phi w} - \mathbf{y}) \\ \|\mathbf{w}\|_\mathbf{2}^\mathbf{2} \leq \xi\end{matrix}\right\} \text{ Start with } ①$$

- The objective function, namely $f(\mathbf{w}) = (\mathbf{\Phi w} - \mathbf{y})^\mathbf{T}(\mathbf{\Phi w} - \mathbf{y})$ is strictly convex. The constraint function, $g(\mathbf{w}) = \|\mathbf{w}\|_\mathbf{2}^\mathbf{2} - \xi$, is also convex.
- For convex $g(\mathbf{w})$, the set $\{\mathbf{w}|\mathbf{g}(\mathbf{w}) \leq \mathbf{0}\}$, is also convex. (Why?)
- To minimize the error function subject to constraint $|\mathbf{w}| \leq \xi$, we apply KKT conditions at the point of optimality $\mathbf{w}^*$ $L(\omega) = f(\omega) + \lambda g(\omega) = ||\phi\omega - y||^2 + \lambda(||\omega||_2^2 - \xi)$

$$\nabla_{\mathbf{w}^*}(f(\mathbf{w}) + \lambda\mathbf{g}(\mathbf{w})) = \mathbf{0}$$

(the first KKT condition). Here, $f(\mathbf{w}) = (\Phi\mathbf{w} - \mathbf{y})^T(\Phi\mathbf{w} - \mathbf{y})$ and, $g(\mathbf{w}) = \|\mathbf{w}\|^2 - \xi$.

(B) We first prove that there exists $\lambda \geq 0$ that satisfies KKT conditions & $\therefore$ is a solution to constrained ridge reg

- Solving we get,

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

From the second KKT condition we get,

$$\|\mathbf{w}^*\|^2 \leq \xi$$

From the third KKT condition,

$$\lambda \geq 0$$

From the fourth condition

$$\lambda\|\mathbf{w}^*\|^{\mathbf{2}} = \lambda\xi$$

$\Big\}$ if $\|w^*\|^2 < \xi$ then $\lambda = 0$ OR if $\lambda > 0$ $\|w\|^2 = \xi$

(C)

- (1) Values of $\mathbf{w}$ and λ that satisfy all these equations would yield an optimal solution. That is, if

if $\|w^*\| = \operatorname{argmin} \|\Phi w - y\|^2$ is st $\|w^*\|^2 \leq \xi$

$$\|\mathbf{w}^*\|^2 = \|(\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}\|^2 \leq \xi \quad (\text{with } \lambda = 0)$$

then $\lambda = 0$ is the solution. (2) Else, for some sufficiently large value, λ will be the solution to

$$\|\mathbf{w}^*\| = \|(\Phi^T\Phi + \lambda I)^{-1}\Phi^T\mathbf{y}\| = \xi$$

$\Big\}$ Q: Is such λ (>0) guaranteed to exist?

- **Bound on λ in the regularized least square solution:** Consider,

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

We multiply $(\Phi^T\Phi + \lambda I)$ on both sides and obtain,

$$\underline{\|\phi^T\phi\, w^*\|} + \lambda\|w^*\| \geq \|(\Phi^T\Phi)\mathbf{w}^* + (\lambda\mathbf{I})\mathbf{w}^*\| = \|\mathbf{\Phi^T y}\|$$

$\|\phi^T\phi\|\,\|w^*\| \geq$

Using the triangle inequality we obtain,

$$\|(\Phi^T\Phi)\mathbf{w}^*\| + (\lambda)\|\mathbf{w}^*\| \geq \|(\mathbf{\Phi^T\Phi})\mathbf{w}^* + (\lambda\mathbf{I})\mathbf{w}^*\| = \|\mathbf{\Phi^T y}\|$$

- By the Cauchy Shwarz inequality, $\|(\Phi^T\Phi)\mathbf{w}^*\| \leq \alpha\|\mathbf{w}^*\|$ for some $\alpha = \|(\Phi^T\Phi)\|$. Substituting in the previous equation,

$$(\alpha + \lambda)\|\mathbf{w}^*\| \geq \|\mathbf{\Phi^T y}\|$$

i.e.

$$\left\{\text{soln to } \|(\phi^T\phi + \lambda I)^{-1}\phi^T y\| = \xi\right\} \Rightarrow \lambda \geq \frac{\|\Phi^T\mathbf{y}\|}{\|\mathbf{w}^*\|} - \alpha \equiv \|w^*\| \geq \frac{\|\phi^T y\|}{(\lambda + \alpha)}$$

Note that when $\|\mathbf{w}^*\| \to \mathbf{0}$, $\lambda \to \infty$. (Any intuition?) Using $\|\mathbf{w}^*\|^{\mathbf{2}} \leq \xi$ we get, $\longrightarrow$ $\therefore$ for some λ, $\|w^*\| = \xi$

Condition (B) is invoked only when $\|(\phi^T\phi)^{-1}\phi^T y\|^2 = \|w^*\|^2 > \xi$

$$\lambda \geq \frac{\|\Phi^T\mathbf{y}\|}{\sqrt{\xi}} - \alpha$$

. You can consult solution to Problem 5 of Tutorials 3 and 4 for the intuition[1]

[1] https://www.cse.iitb.ac.in/~cs725/notes/lecture-slides/tut3-solutions.pdf

- This is not the exact solution to λ but the bound proves the existence of λ for some ξ and ϕ.

The Resultant alternative objective function Substituting $g(\mathbf{w}) = \|\mathbf{w}\|^2 - \xi$, in the first KKT equation considered earlier:

If such λ exists as per KKT, then

$$\underbrace{\nabla_{\mathbf{w}^*}(f(\mathbf{w}) + \lambda \cdot (\|\mathbf{w}\|^2 - \xi)) = \mathbf{0}}$$

This is equivalent to solving

① Use that specific λ^* in $L(\omega, \lambda)$

② Drop other KKT conditions & min $L(\omega, \lambda^*)$

$$\min(\| \Phi \mathbf{w} - \mathbf{y} \|^2 + \lambda \| \mathbf{w} \|^2) \rightarrow \text{penalized ridge reg.}$$

without constraints

for the same choice of λ. This form of **regularized** ridge regression is the **penalized ridge regression**.

which satisfied KKT for constrained ridge reg.

Problem 2. Consider a data set in which each data point y_i is associated with a weighting factor r_i, so that the sum-square error function becomes

$$\frac{1}{2}\sum_{i=1}^{m} r_i(y_i - w^T\phi(x_i))^2$$

Useful for weighing down outliers!

Set $r_i = 0$ for outliers!

Find an expression for the solution w^* that minimizes this error function. The weights r_i's are known before hand. (Exercise 3.3 of Pattern Recognition and Machine Learning, Christopher Bishop).

Solution: Refer to the solution to problem 3, part 1. The solution to problem 2 is included therein.

Example of non-param learning: # params = f(# of data pts)

Problem 3. In problem 2, we discussed weighted regression. In this problem, we will deal with weighted regression, with the weights obtained using some kernel $K(.,.)$. Given a training set of points $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_n, y_n)\}$, we predict a regression function $f(x') = (\mathbf{w}^\top \phi(x') + b)$ for each test (or query point) x' as follows:

$$(\mathbf{w}', b') = \underset{\mathbf{w}, b}{\operatorname{argmin}} \sum_{i=1}^{n} \underbrace{K(x', x_i)}_{K(x', x_i) = r_i^{x'}} \left(y_i - (\mathbf{w}^\top \phi(x_i) + b)\right)^2$$

→ Every point matters in $\mathcal{D}$ but matters to different extent

ie WTD NEAREST NBR

1. If there is a closed form expression for $(\mathbf{w}', b')$ and therefore for $f(x')$ in terms of the known quantities, derive it.

2. How does this model compare with linear regression and $k-$nearest neighbor regression? What are the relative advantages and disadvantages of this model?

3. In the one dimensional case (that is when $\phi(x) \in \Re$), graphically try and interpret what this regression model would look like, say when $K(.,.)$ is the linear kernel[2].

Solution:

This problem is directly related to problem 2 and herein we present the solution to both the problems

[2] Hint: What would the regression function look like at each training data point?

1. The weighing factor $r_i^{x'}$ of each training data point $(\mathbf{x}_i, y_i)$ is now also a function of the query or test data point $(\mathbf{x}', ?)$, so that we write it as $r_i^{x'} = K(\mathbf{x}', \mathbf{x}_i)$ for $i = 1, \ldots, m$. Let $r_{m+1}^{x'} = 1$ and let R be an $(m+1) \times (m+1)$ diagonal matrix of $r_1^{x'}, r_2^{x'}, \ldots, r_{m+1}^{x'}$.

$$\overset{x'}{R} = \begin{bmatrix} r_1^{x'} & 0 & ... & 0 & \\ 0 & r_2^{x'} & ... & 0 & \\ ... & ... & ... & ... & 1 \\ 0 & 0 & 0 & ... & r_{m+1}^{x'} \end{bmatrix}$$

Further, let

$$\Phi = \begin{bmatrix} \phi_1(x_1) & ... & \phi_p(x_1) & 1 \\ ... & ... & ... & 1 \\ \phi_1(x_m) & ... & \phi_p(x_m) & 1 \end{bmatrix}$$

and

$$\widehat{\mathbf{w}}^{x'} = \begin{bmatrix} w_1 \\ ... \\ w_p \\ b \end{bmatrix}$$

and

$$\mathbf{y} = \begin{bmatrix} y_1 \\ ... \\ y_m \end{bmatrix}$$

The sum-square error function then becomes

$$\frac{1}{2}\sum_{i=1}^{m}\left(\sqrt{r_i}\left(y_i - \hat{w}^T\phi(x_i) - b\right)\right)^2 \longleftarrow \frac{1}{2}\sum_{i=1}^{m} \overset{x'}{r_i}(y_i - (\widehat{\mathbf{w}}^T\phi(x_i) + b))^2 = \underline{\frac{1}{2}||\sqrt{R}\mathbf{y} - \sqrt{R}\Phi\widehat{\mathbf{w}}||_2^2}$$

where $\sqrt{R}$ is a diagonal matrix such that each diagonal element of $\sqrt{R}$ is the square root of the corresponding element of R. This is a convex function being minimized (prove this using techniques similar to what we employed for least squares linear regression) and therefore has a global minimum at $\widehat{\mathbf{w}}_*^{x'}$ where the gradient must become 0. (again work out the steps using techniques similar to what we employed for least squares linear regression). The expression for the solution $\widehat{\mathbf{w}}^*$ that minimizes this error function is therefore

$$\widehat{\mathbf{w}}_*^{x'} = (\Phi^T \underline{R}\Phi)^{-1}\Phi^T \underline{R}\mathbf{y}$$

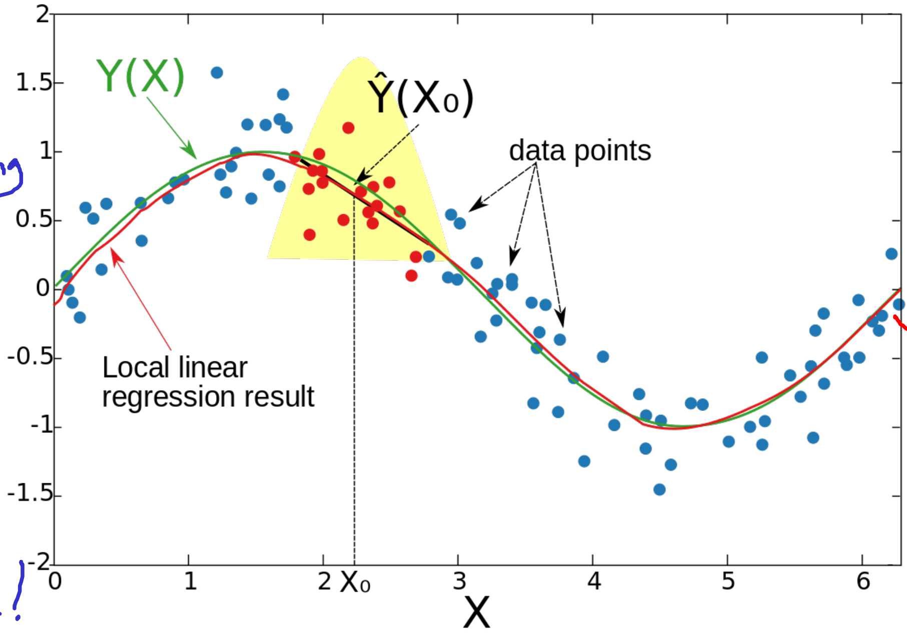
2. Let us refer to this model as local linear regression (Section 6.1.1 of Tibshi's book).

   As compared to linear regression, local linear regression gives more importance to points in $\mathcal{D}$ that are closer/similar to $\mathbf{x}'$ and less importance to points that are less similar. Thus, this method can be important if the regression curve is supposed to take different shapes or different parameters in different parts of the space. For example, in two different regions, the ideal regression curve might be linear in each but with different parameters. In this sense, local linear regression comes close to k-nearest neighbor. But unlike k-nearest neighbor, local linear regression gives you a smooth solution since contribution for regression at a point comes from all data points (in proportion to their closeness) and not just the k closest points.

Nearest nbr $\underset{\text{(more local)}}{>}$ Local linear regression $\underset{\text{(more local)}}{>}$ Least squares

3. Taking clue from the discussion above, one can try and plot this regression curve.

Note: r_i in Q2 can account for noise/outliers by giving them low weightage & hence prevent some overfitting but same $r_i^{x'}$ in Q3 is specific to each test/query pt & therefore can cause overfitting!!



can overfit but less than nearest nbr regression.

Problem 4. Put together the entire story of Support Vector Regression (SVR) one place. You can structure your story along the following lines

1. The motivation behind the basic formualation(s) of SVR with justification for each component of the objective and constraints.
2. The Langrangian function and KKT conditions.
3. Solutions to the KKT conditions and the concept of Support Vectors.
4. The Dual Optimization problem for SVR and the Kernel function
5. Valid kernel functions.

Problem 5. Optional, but could help in ML Project:

For each of the following functions, determine whether a local minimum (maximum) will correspond to a global minimum (maximum). You do not have to prove anything rigorously. You just need to understand the intuitive reason and can consult any web or textual resources.

1. $f(x) = e^x - 1$ on $\mathbb{R}$.
2. $f(x_1, x_2) = x_1 x_2$ on $\mathbb{R}^2_{++}$.
3. $f(x_1, x_2) = 1/(x_1 x_2)$ on $\mathbb{R}^2_{++}$.
4. $f(x_1, x_2) = x_1/x_2$ on $\mathbb{R}^2_{++}$.

5. $f(x_1, x_2) = x_1^2/x_2$ on $\mathbb{R} \times \mathbb{R}_{++}$.

6. $f(x_1, x_2) = x_1^{\alpha} x_2^{1-\alpha}),$ where $0 \leq \alpha \leq 1,$ on $\mathbb{R}_{++}^2$.

7. Suppose $p \geq 1$ and

$$f(x) = \left(\sum_{i=1}^{n} x_i^P \right)^{1/p}$$

with domain dom $f = \mathbb{R}_{++}^n$ What if $p < 1, p \neq 0$?