Lecture 08: Support Vector Regression

Instructor: Prof. Ganesh Ramakrishnan

Recap: Duality and KKT conditions

min $f(x)$
s.t $g_i(x) \leq 0$ & $h_j(x) = 0$

For the previously mentioned formulation of the problem, KKT conditions for all differentiable functions (i.e. f, g_i, h_j) with $\hat{\mathbf{w}}$ primal optimal and $(\hat{\lambda}, \hat{\mu})$ dual optimal point are:

- $\nabla f(\hat{\mathbf{w}}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{\mathbf{w}}) + \sum_{j=1}^{p} \hat{\mu}_j \nabla h_j(\hat{\mathbf{w}}) = 0$
- $g_i(\hat{\mathbf{w}}) \leq 0; 1 \leq i \leq m$
- $\hat{\lambda}_i \geq 0; 1 \leq i \leq m$
- $\hat{\lambda}_i g_i(\hat{\mathbf{w}}) = 0; 1 \leq i \leq m$
- $h_j(\hat{\mathbf{w}}) = 0; 1 \leq j \leq p$

$\exists \hat{\lambda}_i$ & $\hat{\mu}_j$
s.t ....

(For proving existence of λ for L_2 constrained least sq lin reg)

Bound on λ in the regularized least square solution

To minimize the error function subject to constraint $|\mathbf{w}| \le \xi$, we apply KKT conditions at the point of optimality $\mathbf{w}^*$

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

(the first KKT condition). Here, $f(\mathbf{w}) = (\phi\mathbf{w} - \mathbf{Y})^{\mathbf{T}}(\phi\mathbf{w} - \mathbf{Y})$ and, $g(\mathbf{w}) = \|\mathbf{w}\|_2^2 - \xi$. $(\|w\|_2^2 \le \xi)$

Solving we get,

$$\mathbf{w}^* = (\phi^{\mathbf{T}}\phi + \lambda\mathbf{I})^{-1}\phi^{\mathbf{T}}\mathbf{y} \longrightarrow (\text{Solving } (*))$$

From the second KKT condition we get,

$$\|\mathbf{w}^*\|_2^2 \le \xi$$

(if norm is left unspecified, assume L_2)

From the third KKT condition,

$$\lambda \ge 0$$

From the fourth condition

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

(complementary slackness)

Bound on λ in the regularized least square solution

Values of $\mathbf{w}_*$ and λ that satisfy all these equations would yield an optimal 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,

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

Using the triangle inequality we obtain,

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

Bound on λ in the regularized least square solution

$\|(\phi^T\phi)\mathbf{w}^*\| \leq \alpha\|\mathbf{w}^*\|$ for some α for finite $|(\phi^T\phi)\mathbf{w}^*\|$. Substituting in the previous equation,

Chebychev inequality for induced matrix norm:

$\alpha = \|\phi^T\phi\|_2$

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

$\|A\|_2 = \sup_{\|x\|=1} \|Ax\|_2$

i.e.

$$\lambda \geq \frac{\|\phi^T\mathbf{y}\|}{\|\mathbf{w}^*\|} - \alpha$$

$\|A\|_2$ = also called Frobenius norm

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

https://en.wikipedia.org/wiki/Matrix_norm#Frobenius_norm

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

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

Alternative objective function

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

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

This is equivalent to solving

$$\min(\| \Phi\mathbf{w} - \mathbf{y} \|^2 + \lambda \| \mathbf{w} \|^2)$$

Penalty is embedded into objective

for the same choice of λ. This form of **regularized** regression is often referred to as **Ridge regression**.

H/W: Argue that given

$$\omega^* = \underset{\omega}{\operatorname{argmin}} \|\phi\omega - y\|_2^2$$

$$\text{s.t.} \quad \|\omega\|_2^2 \leq \varepsilon_1$$

$\exists$ Dual λ s.t KKT conditions hold & in particular

$$\nabla_{\omega^*}\left(\|\phi\omega - y\|_2^2 + \lambda\|\omega\|_2^2\right) = 0$$

Next argue that, if given λ

$$\hat{\omega} = \underset{\omega}{\operatorname{argmin}} \|\phi\omega - y\|_2^2 + \lambda\|\omega\|_2^2$$

Then $\exists$ ε_1 for which $\|\hat{\omega}\|_2^2 \leq \varepsilon_1$

Regression so far

- **Linear Regression:**
  - $y_i = w^\top \phi(x_i) + b + \epsilon_i$, where:
    $y_i \in \mathbb{R}$, and ϵ_i is the error term
  - *Objective:* $\min_{w,b} \sum_{i=1}^{n} (y_i - w^\top \phi(x_i) - b)^2$
- **Ridge Regression:** [Hereafter penalty term will sit in the objective]
  - $\min_{w,b} \sum_{i=1}^{n} (y_i - w^\top \phi(x_i) - b)^2 + \lambda \|w\|^2$
  - Here, regularization is applied on the linear regression objective to reduce overfitting on the training examples (we penalize model complexity)

Closed-form solutions to regression

- Linear regression and Ridge regression both have closed-form solutions
  - For linear regression,

$$w^* = (\phi^\top \phi)^{-1} \phi^\top y$$

special case with $\lambda = 0$

  - For ridge regression,

$$w^* = (\phi^\top \phi + \lambda I)^{-1} \phi^\top y$$

(for linear regression, $\lambda = 0$)

- *Claim:*
  Error obtained on training data after minimizing ridge regression $\geq$ error obtained on training data after minimizing linear regression (larger search space)
- *Goal:*
  Do well on unseen (test) data as well. Therefore, high training error might be acceptable if test error can be lower