Lecture 08: Ridge Regression, Equivalent Formulations and KKT Conditions

Instructor: Prof. Ganesh Ramakrishnan

Recap: Duality and KKT conditions

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:

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

Bound on λ in the regularized least square solution

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

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

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

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

From the second KKT condition we get,

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

From the third KKT condition,

$$\lambda \geq 0$$

From the fourth condition

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

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,

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

i.e.

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

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

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

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

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

This is equivalent to solving

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

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

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:**
  - $\min_{w,b} \sum_{i=1}^{n} (y_i - w^\top \phi(x_i) - b)^2 + \lambda \|w\|^2$
  - 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,
  $$w^* = (\phi^\top \phi)^{-1}\phi^\top y$$
  - 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
- *Goal:*
  Do well on unseen (test) data as well. Therefore, high training error might be acceptable if test error can be lower