Lecture 09: Lasso and Support Vector Regression

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:

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

Equivalence of the two formulations of regularized least square

Formulation 1:

$$\mathit{min}_{\mathbf{w}}\ f_{\lambda}(\mathbf{w}) = \frac{\mathbf{1}}{\mathbf{2}}||\mathbf{y}-\phi\mathbf{w}||^{\mathbf{2}} + \lambda||\mathbf{w}||^{\mathbf{2}}$$

Formulation 2:

$$\mathit{min}_{\mathbf{w}}\ \frac{1}{2}||y-\phi\mathbf{w}||^{\mathbf{2}}\ \mathbf{s.t.}\ ||\mathbf{w}||^{\mathbf{2}} - \eta \leq \mathbf{0}$$

The Lagrangian for Formulation 2 is:

$$L(\mathbf{w}, \lambda) = \frac{\mathbf{1}}{\mathbf{2}}||\mathbf{y}-\phi\mathbf{w}||^{\mathbf{2}} + \alpha(||\mathbf{w}||^{\mathbf{2}} - \eta)$$

Necessary conditions for optimality for Formulation 1 are:

$$\nabla_{\mathbf{w}}\ f_{\lambda}(\mathbf{w}^{*}(\lambda)) = \mathbf{0}$$

where $\mathbf{w}^{*}(\lambda)$ is the optimal solution for a given λ.

Equivalence of the two formulations of regularized least square

- For Formulation 2, the KKT conditions imply that we have: $\nabla_{\mathbf{w}}\ f_{\alpha}(\mathbf{w}^*) = \mathbf{0}$ and $\alpha^*(||\mathbf{w}^*||^{\mathbf{2}} - \eta) = \mathbf{0}$, $\alpha^* \geq 0$.
- If formulation 1 is solved for a given λ and its solution is $\mathbf{w}^*(\lambda)$ then
  - by setting $\eta = ||\mathbf{w}^*(\lambda)||^{\mathbf{2}}$, you get that $\alpha^* = \lambda$ and $\mathbf{w}^* = \mathbf{w}^*(\lambda)$ satisfy the KKT conditions for formulation 2, showing that both formulations have the same solution.
  - if you solved formulation 2 and set $\lambda = \alpha^*$, you attain the same solution as attained by solving Problem 1.

Choice of regularizer and regularization parameter

- How do we decide which value of λ to choose for the regularizer? How about choosing the regularization parameter λ through cross-validation?
- Recall the polynomial curve fitting problem we considered earlier. As we increased the degree of the polynomial how will the training error wary? What about the test error? And what is the effect of varying λ on train and test errors?
- How about a different regularizer?
  - **Lasso:** When the L_1 norm is used (instead of L_2 as in ridge regression).
- How about a different error function?
  - **Support Vector Regression**.

Lasso: Continuing from Quiz 1, Problem 3

- $$\mathbf{w}^* = \underset{\mathbf{w}}{\textbf{argmin}} \|\phi \mathbf{w} - \mathbf{y}\|^{\mathbf{2}} \text{ s.t. } \|\mathbf{w}\|_{\mathbf{1}} \leq \eta, \tag{1}$$

  where

  $$\|\mathbf{w}\|_1 = \left(\sum_{i=1}^{n} |w_i| \right) \tag{2}$$

- Since $\|\mathbf{w}\|_1$ is not differentiable, one can express (2) as a set of constraints

  $$\sum_{i=1}^{n} \xi_i \leq \eta, \;\; w_i \leq \xi_i, \;\; -w_i \leq \xi_i$$

- The resulting problem is a linearly constrained Quadratic optimization problem (LCQP):

  $$\mathbf{w}^* = \underset{\mathbf{w}}{\textbf{argmin}} \|\phi \mathbf{w} - \mathbf{y}\|^{\mathbf{2}} \text{ s.t. } \sum_{\mathbf{i=1}}^{\mathbf{n}} \xi_{\mathbf{i}} \leq \eta, \;\; \mathbf{w_i} \leq \xi_{\mathbf{i}}, \;\; -\mathbf{w_i} \leq \xi_{\mathbf{i}} \tag{3}$$

Lasso: Continued

- KKT conditions:

$$2(\phi^T\phi)\mathbf{w} - \mathbf{2}\boldsymbol{\phi}^{\mathbf{T}}\mathbf{y} + \sum_{\mathbf{i=1}}^{\mathbf{n}}(\boldsymbol{\theta}_{\mathbf{i}} - \boldsymbol{\lambda}_{\mathbf{i}}) = \mathbf{0}$$

$$\beta(\sum_{i=1}^{n}\xi_i - \eta) = 0$$

$$\forall\ i,\ \theta_i(\mathbf{w_i} - \boldsymbol{\xi}_{\mathbf{i}}) = \mathbf{0}\ \mathbf{and}\ \boldsymbol{\lambda}_{\mathbf{i}}(-\mathbf{w_i} - \boldsymbol{\xi}_{\mathbf{i}}) = \mathbf{0}$$

- Like Ridge Regression, an equivalent Lasso formulation can be shown to be:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\textbf{argmin}}\|\phi\mathbf{w} - \mathbf{y}\|^{\mathbf{2}} + \lambda\|\mathbf{w}\|_{\mathbf{1}} \tag{4}$$

- The justification for the equivalence between (2) and (4) as well as the solution to (4) requires *subgradient*.

Subgradients

- An equivalent condition for convexity of $f(\mathbf{x})$:

$$\forall\, \mathbf{x}, \mathbf{y} \in \mathbf{dmn(f)},\ \ \mathbf{f}(\mathbf{y}) \geq \mathbf{f}(\mathbf{x}) + \nabla^{\top}\mathbf{f}(\mathbf{x})(\mathbf{y} - \mathbf{x})$$

- $\mathbf{g_f}(\mathbf{x})$ is a *subgradient* for a function f at $\mathbf{x}$ if

$$\forall\, \mathbf{y} \in \mathbf{dmn(f)},\ \ \mathbf{f}(\mathbf{y}) \geq \mathbf{f}(\mathbf{x}) + \mathbf{g_f}(\mathbf{x})^{\top}(\mathbf{y} - \mathbf{x})$$

- Any convex (even non-differentiable) function will have a subgradient at any point in the domain!
- If a convex function f is differentiable at $\mathbf{x}$ then $\nabla f(\mathbf{x}) = \mathbf{g_f}(\mathbf{x})$
- $\mathbf{x}$ is a point of minimum of (convex) f if and only if $\mathbf{0}$ is a subgradient of f at $\mathbf{x}$

Subgradients and Lasso

- Claim (out of syllabus): If $\mathbf{w}^*(\eta)$ is solution to (2) and $\mathbf{w}^*(\lambda)$ is solution to (4) then
  - Solution to (2) with $\eta = ||\mathbf{w}^*(\lambda)||$ is also $\mathbf{w}^*(\lambda)$ and
  - Solution to (4) with λ as solution to $\phi^T(\phi\mathbf{w} - \mathbf{y}) = \lambda\mathbf{g_x}$ is also $\mathbf{w}^*(\eta)$
- The unconstrained form for Lasso in (4) has no closed form solution
- But it can be solved using a generalization of gradient descent called *proximal subgradient descent*[1]

[1] https://www.cse.iitb.ac.in/~cs709/notes/enotes/lecture27b.pdf

Proximal Subgradient Descent for Lasso[2]

- Let $\varepsilon(\mathbf{w}) = \|\phi\mathbf{w} - \mathbf{y}\|_\mathbf{2}^\mathbf{2}$
- **Proximal Subgradient Descent Algorithm:**
  Initialization: Find starting point $\mathbf{w}^{(\mathbf{0})}$
  - Let $\widehat{\mathbf{w}}^{(\mathbf{k+1})}$ be a next gradient descent iterate for $\varepsilon(\mathbf{w}^{\mathbf{k}})$
  - Compute $\mathbf{w}^{(\mathbf{k+1})} = \underset{\mathbf{w}}{\textbf{argmin}}||\mathbf{w} - \widehat{\mathbf{w}}^{(\mathbf{k+1})}||_\mathbf{2}^\mathbf{2} + \lambda\mathbf{t}||\mathbf{w}||_\mathbf{1}$ by setting subgradient of this objective to $\mathbf{0}$. This results in:
    1. If $\widehat{w}_i^{(k+1)} > \lambda t$, then $w_i^{(k+1)} = -\lambda t + \widehat{w}_i^{(k+1)}$
    2. If $\widehat{w}_i^{(k+1)} < \lambda t$, then $w_i^{(k+1)} = \lambda t + \widehat{w}_i^{(k+1)}$
    3. 0 otherwise.
  - Set $k = k + 1$, **until** stopping criterion is satisfied (such as no significant changes in $\mathbf{w}^{\mathbf{k}}$ w.r.t $\mathbf{w}^{(\mathbf{k-1})}$)

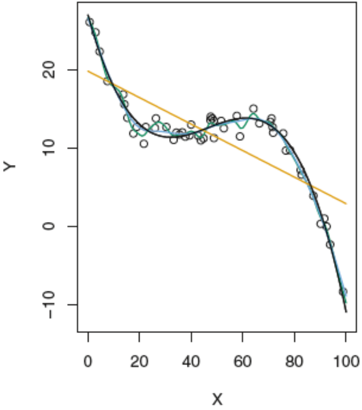
[2] https://www.cse.iitb.ac.in/~cs709/notes/enotes/lecture27b.pdf

Choice of regularizer and regularization parameter

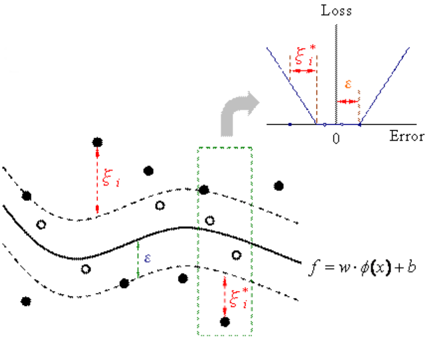
- How do we decide which value of λ to choose for the regularizer? How about choosing the regularization parameter λ through cross-validation?
- Recall the polynomial curve fitting problem we considered earlier. As we increased the degree of the polynomial how will the training error wary? What about the test error? And what is the effect of varying λ on train and test errors?
- How about a different regularizer?
  - **Lasso:** When the L_1 norm is used (instead of L_2 as in ridge regression).
- How about a different error function?
  - **Support Vector Regression**.

Support Vector Regression

Polynomial regression



- Consider a degree 3 polynomial regression model as shown in the figure
- Each bend in the curve corresponds to increase in $\|w\|$
- Eigen values of $(\phi^\top \phi + \lambda I)$ are indicative of curvature. Increasing λ reduces the curvature



- Any point in the band (of ϵ) is not penalized. Thus the loss function is known as ϵ*-insensitive loss*
- Any point outside the band is penalized, and has slackness ξ_i or ξ_i^*
- The SVR model curve may not pass through any training point

- The tolerance ϵ is fixed
- It is desirable that $\forall i$:
  - $y_i - w^\top \phi(x_i) - b \leq \epsilon + \xi_i$
  - $b + w^\top \phi(x_i) - y_i \leq \epsilon + \xi_i^*$

SVR objective

- 1-norm regularized:
  - $\min_{w,b,\xi_i,\xi_i^*} \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i + \xi_i^*)$
    s.t. $\forall i$,
    $y_i - w^\top \phi(x_i) - b \leq \epsilon + \xi_i$,
    $b + w^\top \phi(x_i) - y_i \leq \epsilon + \xi_i^*$,
    $\xi_i, \xi_i^* \geq 0$
- 2-norm regularized:
  - $\min_{w,b,\xi_i,\xi_i^*} \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i^2 + \xi_i^{*2})$
    s.t. $\forall i$,
    $y_i - w^\top \phi(x_i) - b \leq \epsilon + \xi_i$,
    $b + w^\top \phi(x_i) - y_i \leq \epsilon + \xi_i^*$
  - Here, the constraints $\xi_i, \xi_i^* \geq 0$ are not necessary