Introduction to Machine Learning - CS725

Instructor: Prof. Ganesh Ramakrishnan

Lecture 07 - Support Vector Regression and Optimization Basics

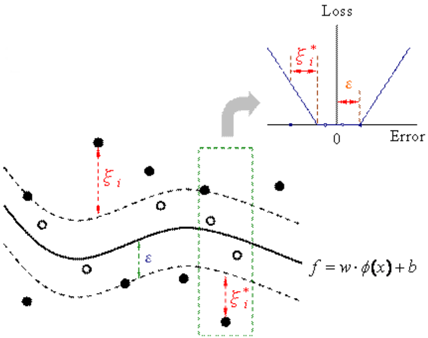
Building on questions on Least Squares Linear Regression

1. Is there a probabilistic interpretation?
   - Gaussian Error, Maximum Likelihood Estimate
2. Addressing overfitting
   - Bayesian and Maximum Aposteriori Estimates for Gaussian and Laplacian (and Beta) priors, L_0, L_1 and L_2 Regularization, **Support Vector Regression**
3. How to minimize the resultant and more complex error functions?
   - Level Curves and Surfaces, Gradient Vector, Directional Derivative, Gradient Descent Algorithm, Convexity, Necessary and Sufficient Conditions for Optimality

Support Vector Regression

One more formulation before we look at Tools of Optimization/duality

Support Vector Regression (SVR)



- 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 $\forall i$:

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

SVR objective

- 1-norm Error, and L_2 regularized:

SVR objective

- 1-norm Error, and L_2 regularized:
  - $\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\left\|\mathbf{w}\right\|^2 + C\sum_i(\xi_i + \xi_i^*)$
    s.t. $\forall i$,
    $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i$,
    $b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*$,
    $\xi_i, \xi_i^* \geq 0$
- 2-norm Error, and L_2 regularized:

SVR objective

- 1-norm Error, and L_2 regularized:
  - $\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*)$
    s.t. $\forall i$,
    $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i$,
    $b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*$,
    $\xi_i, \xi_i^* \geq 0$
- 2-norm Error, and L_2 regularized:
  - $\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i^2 + \xi_i^{*2})$
    s.t. $\forall i$,
    $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i$,
    $b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*$
  - Here, the constraints $\xi_i, \xi_i^* \geq 0$ are not necessary

Need for Optimization so far

- **Unconstrained (Penalized) Optimization:**

$$\mathbf{w}_{Reg} = \arg\min_{\mathbf{w}} \; ||\Phi\mathbf{w} - \mathbf{y}||_2^2 + \Omega(\mathbf{w})$$

- **Constrained Optimization 1:**

$$\mathbf{w}_{Reg} = \arg\min_{\mathbf{w}} \; ||\Phi\mathbf{w} - \mathbf{y}||_2^2$$

$$such\ that\ \Omega(\mathbf{w}) \leq \theta$$

- **Constrained Optimization 2 ($t = 1$ or 2):**

$$\arg\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i (\xi_i^t + \xi_i^{*t})$$

s.t. $\forall i,\ y_i - w^\top\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i;\ b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*$

- **Equivalence**: λ (Penalized) $\equiv$ θ (Constrained)
- **Duality**: Dual of Support Vector Regression

Solving Unconstrained Minimization Problem

- Intuitively: Minimize by setting derivative (gradient) to 0 and hoping to find **closed form** solution.
- When is such a solution a global minimum?
- For most optimization problems, finding closed form solutions is difficult. Even for linear regression (for which closed form solution exists), are there alternative methods?
  - Eg: Consider, $\mathbf{y} = \Phi\mathbf{w}$,where Φ is a matrix with full column rank, the least squares solution, $\mathbf{w}^* = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$. Now, imagine that Φ is a very large matrix. with say, 100,000 columns and 1,000,000 rows. Computation of closed form solution might be challenging.
- How about iterative methods?

SVR objective

- 1-norm Error, and L_2 regularized:
  - $\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*)$
    s.t. $\forall i$,
    $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i$,
    $b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*$,
    $\xi_i, \xi_i^* \geq 0$
- 2-norm Error, and L_2 regularized:
  - $\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i^2 + \xi_i^{*2})$
    s.t. $\forall i$,
    $y_i - \mathbf{w}^\top\phi(x_i) - b \leq \epsilon + \xi_i$,
    $b + \mathbf{w}^\top\phi(x_i) - y_i \leq \epsilon + \xi_i^*$
  - Here, the constraints $\xi_i, \xi_i^* \geq 0$ are not necessary

Need for Optimization so far

- **Unconstrained (Penalized) Optimization:**

$$\mathbf{w}_{Reg} = \arg\min_{\mathbf{w}} \ ||\Phi\mathbf{w} - \mathbf{y}||_2^2 + \Omega(\mathbf{w})$$

- **Constrained Optimization 1:**

$$\mathbf{w}_{Reg} = \arg\min_{\mathbf{w}} \ ||\Phi\mathbf{w} - \mathbf{y}||_2^2$$

$$\textit{such that } \Omega(\mathbf{w}) \leq \theta$$

- **Constrained Optimization 2 ($t = 1$ or 2):**

$$\arg\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i^t + \xi_i^{*t})$$

s.t. $\forall i,\ y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i;\ b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*$

- **Equivalence**: λ (Penalized) $\equiv$ θ (Constrained)
- **Duality**: Dual of Support Vector Regression

Solving Unconstrained Minimization Problem

- Intuitively: Minimize by setting derivative (gradient) to 0 and hoping to find **closed form** solution.
- When is such a solution a global minimum?
- For most optimization problems, finding closed form solutions is difficult. Even for linear regression (for which closed form solution exists), are there alternative methods?
  - Eg: Consider, $\mathbf{y} = \phi\mathbf{w}$,where ϕ is a matrix with full column rank, the least squares solution, $\mathbf{w}^* = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$. Now, imagine that ϕ is a very large matrix. with say, 100,000 columns and 1,000,000 rows. Computation of closed form solution might be challenging.
- How about iterative methods?