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

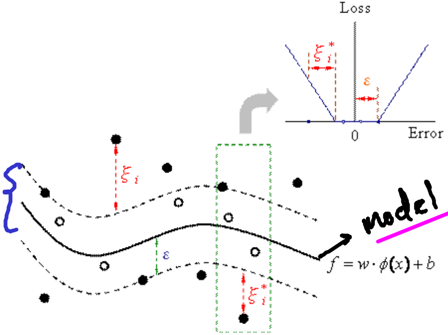
→ SVR motivates Duality (Kernelized representation)

→ Equivalences → Penalized regression & constrained regression

Support Vector Regression

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

Support Vector Regression (SVR)



Region of 0-error

model curve

So far: Generalize (avoid overfitting) through $\Omega(w)$ (L_1, L_2, L_0)

Now: Avoid overfitting (generalize) also through Loss component

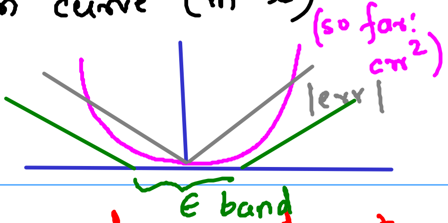
$$\min_{w} \text{Loss}(w, D) + \Omega(w)$$

- Any point in the band (of ϵ) is not penalized. Thus the loss function is known as ϵ-*insensitive loss*
  ↳ Acknowledging that measurement errors be ignored
- 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

In words: Give me a regression curve (in x)

$$f(x) = w^T \phi(x) + b$$

such that:



① Training Pts in an ϵ-band around curve do not contribute to any loss (New part)

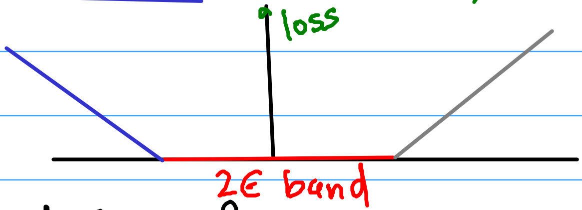
② Loss / Error contribution from pts outside the ϵ-band is minimized (as $\epsilon \to 0$ problem tends to earlier problem)

③ $\Omega(w)$ is penalty (continued from before)

We will formulate & solve SVR problem, which as $\epsilon \to 0$ also helps solve problem of $\min |err|$ discussed in TUTORIAL 1

Form of SVR ϵ error:

$$\text{loss} = \max\left(-\text{err} - \epsilon,\ 0,\ \text{err} - \epsilon\right)$$



Constrained form for above

$$\left.\begin{aligned} \text{loss}(x_i) &\geq 0 \\ &\geq -\left(f(x_i) - y_i\right) - \epsilon \\ &\geq \left(f(x_i) - y_i\right) - \epsilon \end{aligned}\right\} \ \& \ \left\{ \min \ \text{loss}(x_i) \right.$$

$y_i > f(x_i)$ outside ϵ

Intent

$$\xi_i = \max(y_i - f(x_i) - \epsilon, 0) = \max(err_i - \epsilon, 0)$$

$$\xi_i^* = \max(f(x_i) - y_i - \epsilon, 0) = \max(-err_i - \epsilon, 0)$$

$f(x_i) > y_i$ outside ϵ

& $\xi_i \xi_i^* = 0$ → pt i can be either above OR below ϵ band!

- The tolerance ϵ is fixed
- It is desirable that $\forall i$: Constraints

Execution

$$y_i - (w^T \phi(x_i) + b) - \epsilon \leq \xi_i$$

$$0 \leq \xi_i$$

$$w^T \phi(x_i) + b - y_i - \epsilon \leq \xi_i^*$$

$$0 \leq \xi_i^*$$

and

$$\min \sum_i \xi_i + \sum_i \xi_i^*$$

ϵ is a hyperparameter in as much as λ was!

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

& $\min_{\xi_i \xi_i^*} \sum_i (\xi_i + \xi_i^*)$

$(\xi_i \geq 0, \xi_i^* = 0)$

$(\xi_i = \xi_i^* = 0)$

$(\xi_i^* = \xi_i = 0)$

$(\xi_i^* \geq 0, \xi_i \leq 0)$

Geometric view of intent

(a) $y_i, \xi_i = y_i - w^T \phi(x_i) - b - \epsilon$

(b) $y_i, y_i - w^T \phi(x_i) - b - \epsilon \leq 0$

(c) $y_i, w^T \phi(x_i) + b - y_i - \epsilon \leq 0$

(d) $y_i, \xi_i^* = w^T \phi(x_i) + b - y_i - \epsilon$

Prove: That $\xi_i \xi_i^* \leq 0$

Claim: If $\xi_i > 0$ & $\xi_i^* > 0$ .. prove contradiction result

$$\text{(1)} \quad \underbrace{y_i - w^T\phi(x_i) - b}_{k} - \epsilon \le \xi_i$$

$$\text{(2)} \quad \underbrace{w^T\phi(x_i) + b - y_i}_{-k} - \epsilon \le \xi_i^*$$

$$0 \le \xi_i$$

$$0 \le \xi_i^*$$

$$\min_i \left(\xi_i + \xi_i^*\right)$$

$$k - \epsilon = \xi_i \Rightarrow k > \epsilon$$

Proof :- $\xi_i > 0 \Rightarrow$ (1) is an equality $\Rightarrow k - \epsilon > 0 \Rightarrow -k - \epsilon < 0$

$\Rightarrow \xi_i^* = 0$ yields lower value of objective while also satisfying (2)

SVR objective

- 1-norm Error, and L_2 regularized:

$$\underbrace{\min_i C\left(\sum_i (\xi_i + \xi_i^*)\right)}_{\text{Absolute value of error outside } \epsilon \text{ band}} + \|w\|_2^2$$

C is tradeoff between loss & regularizer

s.t. $\forall i$: $y_i - w^T\phi(x_i) - b - \epsilon \le \xi_i$

$w^T\phi(x_i) + b - y_i - \epsilon \le \xi_i^*$

$\xi_i, \xi_i^* \ge 0$

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. $\forall i$,
    $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \le \epsilon + \xi_i$,
    $b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \le \epsilon + \xi_i^*$,
    $\xi_i, \xi_i^* \ge 0$
- 2-norm Error, and L_2 regularized: (minimizing quadratic loss : error^2)

$$\min \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i^2 + \xi_i^{*2})$$

$$\text{s.t. } y_i - w^T\phi(x_i) - b \le \epsilon + \xi_i$$
$$b + w^T\phi(x_i) - y_i \le \epsilon + \xi_i^*$$

Since $\xi_i^2 \to \infty$ as $\xi_i \to -\infty$, $\xi_i \ge 0$ is natural. Similarly $\xi_i^* \ge 0$

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. $\forall i$,
    $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \le \epsilon + \xi_i$,
    $b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \le \epsilon + \xi_i^*$,
    $\xi_i, \xi_i^* \ge 0$
- 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. $\forall i$,
    $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \le \epsilon + \xi_i$, ①
    $b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \le \epsilon + \xi_i^*$ ②
  - Here, the constraints $\xi_i, \xi_i^* \ge 0$ are not necessary

If ① is satisfied for $\xi_i < 0$ then ① is also satisfied for $\xi_i = 0$ while further reducing the objective $C\sum(\xi_i^2 + \xi_i^{*2})$

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) $\equiv \theta$ (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:**

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

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

- **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