Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 8 - Support Vector Regression and Optimization Basics

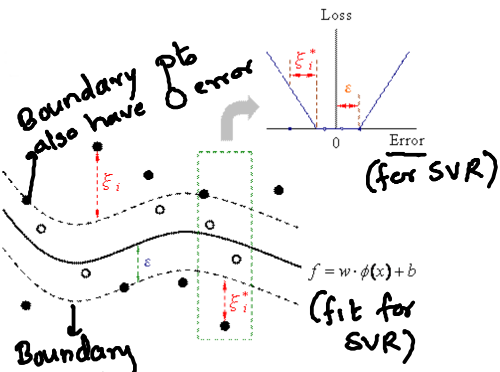
Building on questions on Least Squares Linear Regression

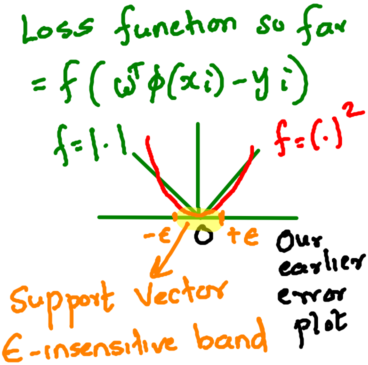
1. Is there a probabilistic interpretation?
   - Gaussian Error, Maximum Likelihood Estimate
2. Addressing overfitting
   - Bayesian and Maximum Aposteriori Estimates, Regularization, **Support Vector Regression**
3. How to minimize the resultant and more complex error functions?
   - Level Curves and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

Support Vector Regression

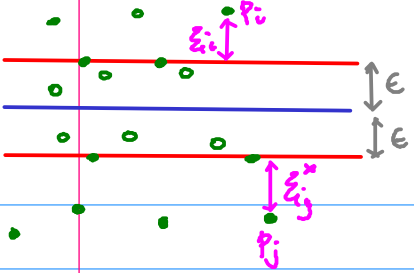
One more formulation before we look at Tools of Optimization/duality

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



For all pts with ϵ band, we expect $\xi_i = \xi_i^* = 0$

The regression curve may not pass through any point at all!

$\boxed{\text{You need both } \xi_i \text{ \& } \xi_i^* \text{ for each } p_i \text{ since } p_i \text{ could lie above or below}}$

$p_i \rightarrow \xi_i, \xi_i^*$

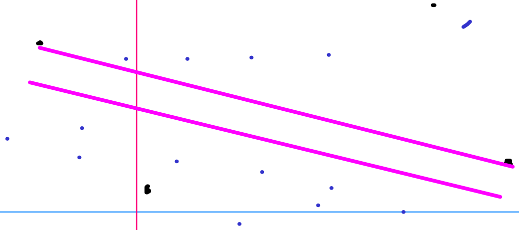
$p_j \rightarrow \xi_j, \xi_j^*$

A pt can be either above OR below

$\therefore \xi_i > 0 \Rightarrow \xi_i^* = 0$ and vice versa

ξ_i = distance of p_i above the ϵ band

ξ_i^* = distance of p_i below the ϵ-band



$\epsilon = 0.5$

H/W: Think of a setting in which no point will land up in the ϵ-band

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

$$\xi_i, \xi_i^*, \epsilon, \mathcal{D} = \{(x_i, y_i)\}$$

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

Constraints?

Objective desired to be minimized?

Constraints First!

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

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

$w^T\phi(x_i) + b$

ⓓ • y_i, $w^T\phi(x_i) + b - y_i - \epsilon \le 0$

$\updownarrow \epsilon$

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

- The tolerance ϵ is fixed
- It is desirable that $\forall i$:
  - ① $y_i - w^\top \phi(x_i) - b \le \epsilon + \xi_i$ → Even if $y_i \ge w^T\phi(x_i) + b$ [Summarizes cases ⓐ & ⓑ]
  - $b + w^\top \phi(x_i) - y_i \le \epsilon + \xi_i^*$ → Even if $y_i \le w^T\phi(x_i) + b$ [summarizes cases ⓒ & ⓓ]

Alternatively:

$$\xi_i = \max\left(y_i - w^T\phi(x_i) - b - \epsilon, 0\right)$$

$$\xi_i^* = \max\left(b + w^T\phi(x_i) - y_i - \epsilon, 0\right)$$

SVR objective

Think of C as $\frac{1}{2\lambda}$

- 1-norm Error, and L_2 regularized:

$$\underbrace{\frac{1}{2}\|w\|_2^2}_{L_2 \text{ regularizer}} + \underbrace{C\left(\sum_{i=1}^{m} \xi_i + \xi_i^*\right)}_{\text{1-norm Error}}$$

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

$$w^T\phi(x_i) + b - y_i \le \epsilon + \xi_i^* \quad (2)$$

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

$\sum_{i=1}^{m}(\xi_i + \xi_i^*)$ = sum of errors over all m data points

Q1 : Why is $\xi_i, \xi_i^* \ge 0$ required ?

Q2 : How does this formulation ensure that one of ξ_i & ξ_i^* will be 0 ?

Ans to Q2: Proof by contradiction.

↳ Suppose $\xi_i > 0$ & $\xi_i^* > 0$

ie $y_i - w^T\phi(x_i) - b \leq \epsilon + \xi_i$ ①

$w^T\phi(x_i) + b - y_i \leq \epsilon + \xi_i^*$ ②

Claim: $\xi_i^* = 0$ would also satisfy ② [Complete proof

⇒ By setting $\xi_i^* = 0$, I can lower my error in objective while still satisfying ① and ②

⇒ Contradicts assumption that $\xi_i > 0$ $\xi_i^* > 0$ was an "optimal" solution!

SVR objective

- 1-norm Error, and L_2 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 Error, and L_2 regularized:

$+ C \sum_i (\xi_i^2 + \xi_i^{*2})$

$\xi_i, \xi_i^* \geq 0 \Leftarrow$ Not required! (H/W)

SVR objective

- 1-norm Error, and L_2 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 Error, and L_2 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

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):** → SVR

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

s.t. $\forall i,\ y_i - w^\top\phi(x_i) - b \leq \epsilon + \xi_i;\ b + w^\top\phi(x_i) - y_i \leq \epsilon + \xi_i^*$

- **Equivalence**: λ (Penalized) $\equiv \theta$ (Constrained)
- **Duality**: Dual of Support Vector Regression: Kernels, non-linear Non-parametric

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?