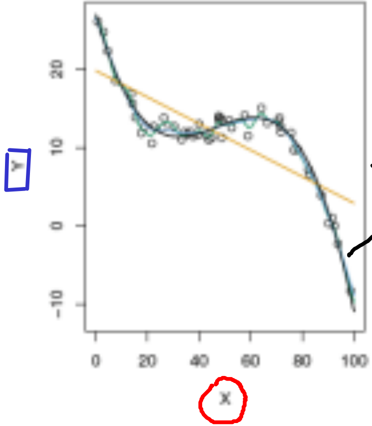
Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 8 - Support Vector Regression and Optimization Basics

Recap: Overfitting and Regularization through Illustration



$y \approx w_0 + w_1 x + w_2 x^2 + w_3 x^3$

- 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 [Prob 8 of Tut 3 & 4]

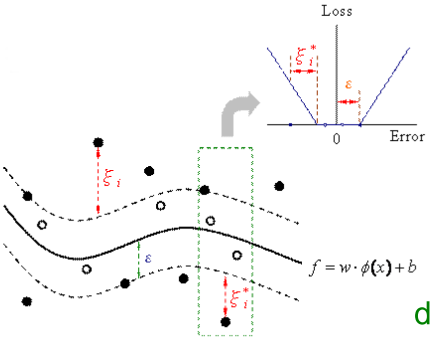
Support Vector Regression

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

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, $\Omega(\|w\|)$
   - $L(w,x,y)$ ← **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 (SVR)



SVR attempts to avoid overfitting also through the loss component by introducing an epsilon insensitive loss (band)

$$\min_{w} \; \underbrace{L_\epsilon(w, D)} + \Omega(w)$$

discover the regression curve s.t points around it within epsilon band have not been penalized

- ϵ-*insensitive loss*: Any point in the band (of ϵ) is not penalized.
- Any point outside the band is penalized, and has slackness ξ_i or ξ_i^* SVR is graceful on epsilon measurement errors and penalizes rest
- The SVR model curve may not pass through any training point

$$\text{Loss}(x_i) \geq 0$$

$$\xi_i = y_i - (w^T\phi(x_i) + b + \epsilon) \quad \text{if } y_i \geq w^T\phi(x_i) + b + \epsilon$$

$$\xi_i^* = w^T\phi(x_i) + b - \epsilon - y_i \quad \text{if } y_i \leq w^T\phi(x_i) + b - \epsilon$$

Need a compact(er) expression for the Loss!!

$$\text{Loss} = \sum_i \text{Loss}(x_i) = \sum_i (\xi_i + \xi_i^*)$$ [since at most one of them should be non-zero]

INTENT

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

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

Execution: min a s.t. $a = \max(b, c) \iff$ min a s.t. $a \geq b$, $a \geq c$

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

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

Regression curve (line)

(b) $\xi_i = 0$

$+\epsilon$ band

(c) $\xi_i^* = 0$

$-\epsilon$ band

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

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

[Necessary conditions that become sufficient when also minimize $\sum_i (\xi_i + \xi_i^*)$]

$0 \leq \xi_i, \xi_i^*$

obvious from construction

$\xi_i \cdot \xi_i^* = 0$ → claim: This is redundant constraint

ϵ is a hyperparameter like λ or σ^2

Claim: If $\xi_i > 0$ & $\xi_i^* > 0$ we get a contradiction

Proof: ① $\xi_i \geq \underbrace{y_i - w^T\phi(x_i) - b}_{k_i} - \epsilon$

$\xi_i = \max(0, k_i - \epsilon)$

$\left\{ \min_{w, \xi_i, \xi_i^*} C \sum (\xi_i + \xi_i^*) + \cdots \right.$

② $\xi_i^* \geq \underbrace{w^T\phi(x_i) + b - y_i}_{-k_i} - \epsilon$

$\xi_i^* = \max(0, -k_i - \epsilon)$

If $\xi_i > 0$ then $\xi_i = k_i - \epsilon > 0 \Rightarrow k_i > \epsilon$

Additionally

If $\xi_i^* > 0$ then $\xi_i^* = -k_i - \epsilon > 0 \Rightarrow -k_i > \epsilon$

Contradiction

SVR objective

- 1-norm Error, and L_2 regularized:

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

Instead of C in the error, you could we a λ in regularizer!

C is inversely related to λ

$\frac{1}{2}$ is only to simplify future derivations & can be ignored

s.t. Constraints discussed earlier hold

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:

$\sum_i (\xi_i + \xi_i^*)^2$

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. $\forall i$,
    $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i$, → if $\xi_i < 0$ satisfies constraints so will $\xi_i = 0$ -- since objective will be happier with smaller ξ_i^2, $\xi_i = 0$
    $b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*$
  - Here, the constraints $\xi_i, \xi_i^* \geq 0$ are not necessary

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

Regression through the eyes of Optimization

Need for Optimization so far

- **Unconstrained (Penalized) Optimization:** (Eg: Ridge)

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

- **Constrained Optimization 1:**

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

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

Need for Optimization so far (contd.)

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

SVR

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

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

- **Equivalence**: λ (Penalized) $\equiv$ θ (Constrained)
- **Iteratively Solving:** Lasso, Regression with L_0 norm, Support Vector Regression
- **Duality**: Dual of Support Vector Regression $\Big\}$ Kernelization

Foundations: Level curves and surfaces

- A level curve of a function $\mathbf{f}(\mathbf{x})$ is defined as a curve along which the value of the function remains unchanged while we change the value of its argument x.
- Formally we can define a level curve as :

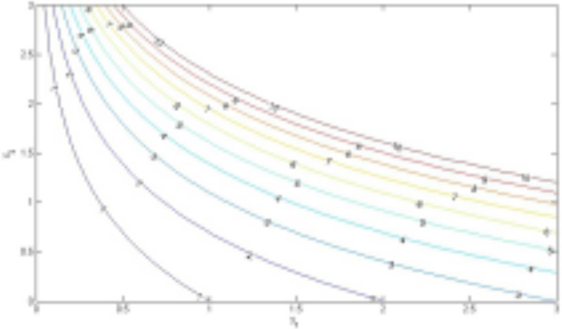
$$L_c(\mathbf{f}) = \left\{ \mathbf{x} | \mathbf{f}(\mathbf{x}) = \mathbf{c} \right\} \tag{1}$$

where c is a constant.

Foundations: Level curves and surfaces

Region of quick change

x_2 ↑

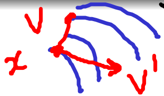


Region of slow Change

→ x_1

Figure 1: 10 level curves for the function $\mathbf{f}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1 \mathbf{e}^{\mathbf{x}_2}$ (Figure 4.12 from https://www.cse.iitb.ac.in/~CS725/notes/classNotes/BasicsOfConvexOptimization.pdf)

Foundations: Directional Derivatives



- Directional derivative: Rate at which the function changes at a given point $\mathbf{x}$ in a given direction $\mathbf{v}$
- The *directional derivative* of a function f in the direction of a unit vector $\mathbf{v}$ at a point $\mathbf{x}$ can be defined as :

$$D_{\mathbf{v}}(f,\mathbf{x}) = \lim_{h\to 0} \frac{f(\mathbf{x}+h\mathbf{v}) - \mathbf{f}(\mathbf{x})}{h} \tag{2}$$

$$s.t. \; ||\mathbf{v}||_2 = \mathbf{1} \tag{3}$$

Normalization

Foundations: Gradient Vector

- The **g**radient vector of a function f at a point $\mathbf{x}$ is defined as:

$D_v(f, x) = v^T \nabla f(x)$

$\|\nabla f(x)\| = \max_v D_v(f, x)$

$$\nabla f_{\mathbf{x}^*} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \cdot \\ \cdot \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \epsilon \mathbb{R}^n \qquad (4)$$

Directional derivative along x_i

- Magnitude (euclidean norm) of gradient vector at any point indicates maximum value of directional derivative at that point
- Direction of gradient vector indicates direction of this maximal directional derivative at that point. $\nabla f(x) / \|\nabla f(x)\|$

Foundations: Gradient Vector

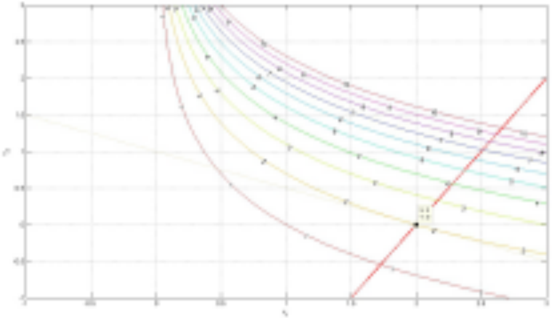


Figure 2: The level curves along with the gradient vector at (2, 0). Note that the gradient vector is perpenducular to the level curve $x_1 e^{x_2} = 2$ at (2, 0)

(Reminiscent of electrostatic flux)

Foundations: Gradient Vector

- Magnitude (euclidean norm) of gradient vector at any point indicates maximum value of directional derivative at that point
- Thus, <u>at the point of minimum</u> of a differentiable minimization objective (such as ridge regression), ...

Expect: $\nabla f(w_{MLE}) = 0$

Foundations: Gradient Vector

- Magnitude (euclidean norm) of gradient vector at any point indicates maximum value of directional derivative at that point
- Thus, at the point of minimum of a differentiable minimization objective (such as ridge regression), ...
- Ridge Regression: Find $\mathbf{w}$ such that

$O(w)$

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} ||\Phi\mathbf{w} - \mathbf{y}||^2 + \lambda||\mathbf{w}||^2 \quad (5)$$

$$= \arg\min_{\mathbf{w}} (\mathbf{w}^T\Phi^T\Phi\mathbf{w} - 2\mathbf{w}^T\phi\mathbf{y} - \mathbf{y}^T\mathbf{y} + \lambda||\mathbf{w}||^2) \quad (6)$$

$\nabla O(w^*) = 0$

Foundations: Necessary condition 1 (Solving Ridge)

- *If* $\nabla f(\mathbf{w}^*)$ *is defined &* $\mathbf{w}^*$ *is local minimum/maximum, then* $\nabla f(\mathbf{w}^*) = 0$ (A necessary condition) (Cite : Theorem 60 of CS725/notes/classNotes/BasicsOfConvexOptimization.pdf)
- Given that

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

$$\implies \ldots\ldots\ldots$$

- We would have

$$\ldots\ldots\ldots$$

$$\implies \ldots\ldots\ldots\ldots\ldots\ldots$$

$$\implies \ldots\ldots\ldots\ldots\ldots\ldots$$

Foundations: Necessary condition 1 (Solving Ridge)

- *If $\nabla f(\mathbf{w}^*)$ is defined & $\mathbf{w}^*$ is local minimum/maximum, then* $\nabla f(\mathbf{w}^*) = 0$ (A necessary condition) `(Cite : Theorem 60)` `CS725/notes/classNotes/BasicsOfConvexOptimization.pdf`
- Given that

$$f(\mathbf{w}) = (\mathbf{w}^T\Phi^T\Phi\mathbf{w} - 2\mathbf{w}^T\Phi^T\mathbf{y} - \mathbf{y}^T\mathbf{y} + \lambda||\mathbf{w}||^2) \quad (7)$$

$$\implies \nabla f(\mathbf{w}) = 2\Phi^T\Phi\mathbf{w} - 2\Phi^T\mathbf{y} + 2\lambda\mathbf{w} \quad (8)$$

- We would have

$$\nabla f(\mathbf{w}^*) = 0 \quad (9)$$

$$\implies 2(\Phi^T\Phi + \lambda I)\mathbf{w}^* - 2\Phi^T\mathbf{y} = 0 \quad (10)$$

$$\implies \mathbf{w}^* = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T\mathbf{y} \quad (11)$$