Lecture 09-b: Support Vector Regression in some details

Instructor: Prof. Ganesh Ramakrishnan

Recall motivations for regularizer: $\frac{1}{2}\|w\|_2^2$

① Slides 17 to 19 of https://www.cse.iitb.ac.in/~cs725/notes/lecture-slides/lecture-06-unannotated.pdf

Analytical motivation

Empirical motivation

$\phi w = y$ for $\phi = \begin{bmatrix} 1 & x_1^1 & x_1^2 & \cdots & x_1^{p-1} \\ \vdots \\ 1 & x_m^1 & x_m^2 & \cdots & x_m^{p-1} \end{bmatrix}$

- will have 1 or more solutions when $p \geq m$ and no solutions (unless rank of $\phi < m$) otherwise. Thus, with $p \geq m$, you have more chances of fitting regression curve on all data points than when $p < m$
- will therefore tend to "overfit" training data at the risk of making errors on test data as p increases
- $p > p' \equiv$ obtaining $w^{p'}$ by setting higher coefficient indices of w^p to zero

$w^{p'} = w^p(1, 2 \dots p')$

- ∴ Determining "right" p' is like finding a "sparse" solution for larger value of p

(see pages 3 to 5 of https://www.cse.iitb.ac.in/~cs725/notes/lecture-slides/lecture-07-annotated.pdf)

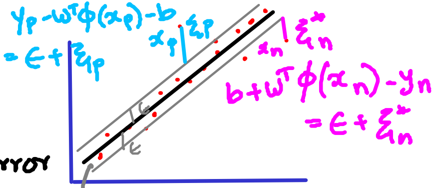
Observe how test error increases for $t \geq 7$ whereas train error keeps decreasing for a while.

This indicates that some regularization (to penalize non-zero w_i's) could have helped avoid overfitting

Again see some of the arguments below that are common to the empirical & analytic reasoning

Also see: Slide 12 of https://www.cse.iitb.ac.in/~cs725/notes/lecture-slides/lecture-09-unannotated.pdf for how the ISTA algo for LASSO attains this sparsity (for L_1 regularization) by thresholding $\hat{w}_i^{(k+1)}$ and only letting "promising" values of $\hat{w}_i^{(k+1)}$ to be non-zero and setting the rest to zero

Support Vector Regression (SVR)



Zero error in this ϵ-band

Note: We want

$$\left|\left(y_i - \left(w^T\phi(x_i) + b\right)\right)\right| \leq \epsilon + \xi \text{ error}$$

- $\min_{w,b,\xi_i,\xi_i^*} \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i + \xi_i^*)$
  s.t. $\forall i$,
  $y_i - w^\top\phi(x_i) - b \leq \epsilon + \xi_i$,
  $b + w^\top\phi(x_i) - y_i \leq \epsilon + \xi_i^*$,
  $\xi_i, \xi_i^* \geq 0$

Claim: Exactly one of ξ_i & $\xi_i^* > 0$

Claim: At optimal soln, one of the two will be an equality for all points that lie outside the ϵ-band

- Let's consider the lagrange multipliers α_i, α_i^*, μ_i and μ_i^* corresponding to the above-mentioned constraints respectively.

Claim: For points within ϵ-band, $\xi_i = \xi_i^* = 0$

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

s.t

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

$$w^T\phi(x_i) + b - y_i \leq \epsilon + \xi_i^* \rightarrow (3)$$

$$\xi_i^*, \xi_i \geq 0 \longrightarrow (4)$$

for each i

Claims 1 & 2: Suppose $\xi_i > 0$ then as per (2)

$y_i - w^T\phi(x_i) - b \leq \epsilon + \xi_i$ (Multiplying both sides by -1)

$$-y_i + w^T\phi(x_i) + b \geq -\epsilon - \xi_i \rightarrow (2')$$

Claim: By setting $\xi_i^* = 0$, (3) will be satisfied & objective reduced in contrast with $\xi_i^* > 0$

↳ Claim: $y_i - w^T\phi(x_i) - b > \epsilon$ (since if LHS $\leq \epsilon$ then $\xi_i = 0$ should have been the soln)

↳ $-y_i + w^T\phi(x_i) + b < -\epsilon < 0$

↳ $-y_i + w^T\phi(x_i) + b < 0 + \xi_i^*$

See next slide

Similarly if $\exists\ \xi_i$ for which $y_i - w^T\phi(x_i) - b = \epsilon + \xi_i$, that ξ_i will be chosen

Continued discussion on last point from previous page

Claim: If $\hat{\xi}_i$ s.t. $y_i - w^T\phi(x_i) - b < \epsilon + \hat{\xi}_i$ is optimal, then I claim that $\xi_i < \hat{\xi}_i$

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

will give objective $= \frac{1}{2}\|w\|_2^2 + C\xi_i + \text{o/rs}$

that is less than $\frac{1}{2}\|w\|_2^2 + C\hat{\xi}_i + \text{o/rs}$

Lagrange fn

$$L(\omega, b, \xi_i, \xi_i^*, \alpha_i, \alpha_i^*, \mu_i, \mu_i^*) = \frac{1}{2}\|\omega\|^2 + C\sum\left(\xi_i + \xi_i^*\right) + \sum_i \alpha_i^*\left(\omega^T\phi(x_i) + b - y_i - \epsilon - \xi_i^*\right)$$
$$+ \sum_i \alpha_i\left(y_i - \omega^T\phi(x_i) - b - \epsilon - \xi_i\right) - \sum_i \mu_i \xi_i - \sum_i \mu_i^* \xi_i^*$$

$$\min_{b,\omega,\xi,\xi_i^*} \frac{1}{2}\|\omega\|^2 + C\sum_i(\xi_i + \xi_i^*)$$

s.t. $y_i - \omega^T\phi(x_i) - b \le \epsilon + \xi_i \rightarrow \alpha_i$ (2)

$\omega^T\phi(x_i) + b - y_i \le \epsilon + \xi_i^* \rightarrow \alpha_i^*$ (3)

$\left.\begin{array}{l} \xi_i \ge 0 \longrightarrow \mu_i \\ \xi_i^* \ge 0 \longrightarrow \mu_i^* \end{array}\right\}$ (4)

$$\nabla_\omega L(\omega, b, \xi_i, \xi_i^*, \alpha_i, \alpha_i^*, \mu_i, \mu_i^*) = 0 \Rightarrow \omega + \sum_i -\alpha_i\phi(x_i) + \alpha_i^*\phi(x_i) = 0$$

ie $\omega = \sum_i (\alpha_i^* - \alpha_i)\phi(x_i)$

$$\nabla_b L(\omega, b, \xi_i, \xi_i^*, \alpha_i, \alpha_i^*, \mu_i, \mu_i^*) = 0 \Rightarrow \sum(\alpha_i - \alpha_i^*) = 0$$

$\nabla_{\xi_i} L = 0 \Rightarrow C - \alpha_i - \mu_i = 0$ similarly: $C - \alpha_i^* - \mu_i^* = 0$

KKT conditions

- Differentiating the Lagrangian w.r.t. w,
  $w - \alpha_i\phi(x_i) + \alpha_i^*\phi(x_i) = 0$
  i.e. $w = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\phi(x_i)$
- Differentiating the Lagrangian w.r.t. ξ_i,
  $C - \alpha_i - \mu_i = 0$
  i.e. $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,
  $\alpha_i^* + \mu_i^* = C$
- Differentiating the Lagrangian w.r.t b,
  $\sum_i(\alpha_i^* - \alpha_i) = 0$
- Complimentary slackness:
  $\alpha_i(y_i - w^\top\phi(x_i) - b - \epsilon - \xi_i) = 0$
  $\mu_i\xi_i = 0$
  $\alpha_i^*(b + w^\top\phi(x_i) - y_i - \epsilon - \xi_i^*) = 0$
  $\mu_i^*\xi_i^* = 0$

(1) If $\xi_i > 0$ then $\mu_i = 0$ & $\alpha_i = C$
If $\xi_i^* > 0$ then $\mu_i^* = 0$ & $\alpha_i^* = C$

(2) $\alpha_i \in (0, C) \Rightarrow$ $\mu_i \in (0, C) \Rightarrow$ $\xi_i = 0$ & $y_i - w^T\phi(x_i) - b - \epsilon - \xi_i = 0$
$\Rightarrow y_i - w^T\phi(x_i) - b = \epsilon$

Conclusions from the KKT conditions:

$$\alpha_i \in (0, C) \Rightarrow ?$$

$$\alpha_i^* \in (0, C) \Rightarrow ?$$