Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 11 - Support Vector Regression and KKT-based analysis

Recap: KKT Conditions for Constrained Optimization

Recap: Constrained (Convex) Problem

- The general optimization problem we consider with (convex) inequality and (linear) equality constraints is:

$$\min_{\mathbf{w}} f(\mathbf{w})$$

$$\text{subject to } g_i(\mathbf{w}) \leq 0; 1 \leq i \leq m$$

$$h_j(\mathbf{w}) = 0; 1 \leq j \leq p$$

Recap: KKT conditions

- Here, $\mathbf{w} \in \mathbb{R}^n$ and the domain is the intersection of all functions. Lagrangian is: $L(\mathbf{w}, \lambda, \mu) = f(\mathbf{w}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{w}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{w})$
- KKT **necessary** conditions for all differentiable functions (i.e. f, g_i, h_j) with optimality points $\hat{\mathbf{w}}$ and $(\hat{\lambda}, \hat{\mu})$ are:
  - $\nabla f(\hat{\mathbf{w}}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{\mathbf{w}}) + \sum_{j=1}^{p} \hat{\mu}_j \nabla h_j(\hat{\mathbf{w}}) = 0$
  - $g_i(\hat{\mathbf{w}}) \leq 0; 1 \leq i \leq m$ and $h_j(\hat{\mathbf{w}}) = 0; 1 \leq j \leq p$
  - $\hat{\lambda}_i \geq 0; 1 \leq i \leq m$ and $\hat{\lambda}_i g_i(\hat{\mathbf{w}}) = 0; 1 \leq i \leq m$
- When f and $g_i, \forall i \in [1, m]$ are convex and $h_j, \forall j \in [1, p]$ are affine, KKT conditions are also **sufficient** for optimality at $\hat{\mathbf{w}}$ and $(\hat{\lambda}, \hat{\mu})$

Recap: KKT conditions for the Constrained (Convex) Problem

Recap Application 1: Equivalence of two forms of Ridge Regression

Recap: Equivalent Forms of Ridge Regression

- Values of $\mathbf{w}$ and λ that satisfy all these equations would yield an optimal solution. That is, if

$$\|\mathbf{w}^*\| = \|(\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}\| \leq \xi$$

then $\lambda = 0$ is the solution. Else, for some sufficiently large value, λ will be the solution to

$$\|\mathbf{w}^*\| = \|(\Phi^T\Phi + \lambda I)^{-1}\Phi^T\mathbf{y}\| = \xi$$

Recap: Reformulation of Constrained (Ridge) Regression

Substituting $g(\mathbf{w}) = \|\mathbf{w}\|^2 - \xi$, in the first KKT equation considered earlier:

$$\nabla_{\mathbf{w}^*}(f(\mathbf{w}) + \lambda \cdot (\|\mathbf{w}\|^2 - \xi)) = \mathbf{0}$$

This is equivalent to solving

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

for the same choice of λ. This form of **regularized** ridge regression is the **penalized ridge regression**.

Recap: Lagrangian Duality

Recap: Lagrangian Duality and KKT conditions

- With $\mathbf{w} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p$, Lagrangian is:

$$L(\mathbf{w}, \lambda, \mu) = f(\mathbf{w}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{w}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{w})$$

- Lagrange dual function is minimum of Lagrangian over $\mathbf{w}$.

Recap: Lagrangian Duality and KKT conditions

- With $\mathbf{w} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p$, Lagrangian is:

$$L(\mathbf{w}, \lambda, \mu) = f(\mathbf{w}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{w}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{w})$$

- Lagrange dual function is minimum of Lagrangian over $\mathbf{w}$.

$$L^*(\lambda, \mu) = \min_{\mathbf{w}} L(\mathbf{w}, \lambda, \mu)$$

- The Dual Optimization Problem is to maximize Lagrange dual function $L^*(\lambda, \mu)$ over (λ, μ)

Figure 4.42 of → Find hyperplance below constraint set with largest y-intercept

https://www.cse.iitb.ac.in/~cs725/notes/classNotes/BasicsOfConvexOptimization.p

Recap: Lagrangian Duality and KKT conditions

- With $\mathbf{w} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^p$, Lagrangian is:

$$L(\mathbf{w}, \lambda, \mu) = f(\mathbf{w}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{w}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{w})$$

- Lagrange dual function is minimum of Lagrangian over $\mathbf{w}$.

$$L^*(\lambda, \mu) = \min_{\mathbf{w}} L(\mathbf{w}, \lambda, \mu)$$

- The Dual Optimization Problem is to maximize Lagrange dual function $L^*(\lambda, \mu)$ over (λ, μ)

$$\operatorname*{argmax}_{\lambda, \mu} L^*(\lambda, \mu) = \operatorname*{argmax}_{\lambda, \mu} \min_{\mathbf{w}} L(\mathbf{w}, \lambda, \mu)$$

Extra: Lagrangian Duality and KKT conditions

- The dual function yields lower bound for minimizer of the primal formulation.
- Max of dual function $L^*(\lambda, \mu)$ over (λ, μ) is also therefore a lower bound

For intuition see Figure 4.42 of https://www.cse.iitb.ac.in/~cs725/notes/classNotes/BasicsOfConvexOptimizat

$$\max_{\lambda,\mu} \min_{w} L(w,\lambda,\mu) \leq \min_{w} f(w)$$

s.t $g_i(w) \leq 0$

$h_j(w) = 0$

The extent of inequality is the duality gap

Extra: Lagrangian Duality and KKT conditions

- The dual function yields lower bound for minimizer of the primal formulation.
- Max of dual function $L^*(\lambda, \mu)$ over (λ, μ) is also therefore a lower bound

$$\max_{\lambda,\mu} L^*(\lambda, \mu) = \max_{\lambda,\mu} \min_{\mathbf{w}} L(\mathbf{w}, \lambda, \mu) \leq L(\mathbf{w}, \lambda, \mu)$$

- **Duality Gap:** The gap between primal and dual solutions. In the KKT conditions, $\hat{\mathbf{w}}$ correspond to primal optimal and $(\hat{\lambda}, \hat{\mu})$ to dual optimal points $\Rightarrow$ Duality gap is $f(\hat{\mathbf{w}}) - L^*(\hat{\lambda}, \hat{\mu})$
- Duality gap characterizes suboptimality of the solution and can be approximated by $f(\mathbf{w}) - L^*(\lambda, \mu)$ for any feasible $\mathbf{w}$ and corresponding λ and μ

Extra: Lagrangian Duality and KKT conditions

- The dual function yields lower bound for minimizer of the primal formulation.
- Max of dual function $L^*(\lambda, \mu)$ over (λ, μ) is also therefore a lower bound

Equality (zero gap) for Convexity

$$\max_{\lambda,\mu} L^*(\lambda, \mu) = \max_{\lambda,\mu} \min_{\mathbf{w}} L(\mathbf{w}, \lambda, \mu) \leq L(\mathbf{w}, \lambda, \mu)$$

- **Duality Gap:** The gap between primal and dual solutions. In the KKT conditions, $\hat{\mathbf{w}}$ correspond to primal optimal and $(\hat{\lambda}, \hat{\mu})$ to dual optimal points $\Rightarrow$ Duality gap is $f(\hat{\mathbf{w}}) - L^*(\hat{\lambda}, \hat{\mu})$
- Duality gap characterizes suboptimality of the solution and can be approximated by $f(\mathbf{w}) - L^*(\lambda, \mu)$ for any feasible $\mathbf{w}$ and corresponding λ and μ

KKT conditions for the Constrained (Convex) Problem

Application 2: SVR and its Dual: : Assume the $\hat{}$ on values of $\left\{\hat{\mathbf{w}}, \hat{b}, \hat{\xi}, \hat{\xi}^*, \hat{\alpha}, \hat{\alpha}^*, \hat{\mu}, \hat{\mu}^*\right\}$ at KKT when not explicitly specified

KKT and Dual for SVR

- $$\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*)$$
  s.t. $\forall i$,
  $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \le \epsilon + \xi_i$, (α_i)
  $b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \le \epsilon + \xi_i^*$, (α_i^*)
  μ_i $\xi_i, \xi_i^* \ge 0$ μ_i^*
- Consider corresponding lagrange multipliers α_i, α_i^*, μ_i and μ_i^*
- The Lagrange Function is

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

$\alpha_i \;\; \alpha_i^* \;\; \mu_i \;\; \mu_i^* \ge 0$

KKT and Dual for SVR

- $$\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*)$$
  s.t. $\forall i$,
  $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i$,
  $b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*$,
  $\xi_i, \xi_i^* \geq 0$
- Consider corresponding lagrange multipliers α_i, α_i^*, μ_i and μ_i^*
- The Lagrange Function is $L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) =$
  $$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i\left(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) +$$
  $$\sum_{i=1}^m \alpha_i^*\left(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i\xi_i - \sum_{i=1}^m \mu_i^*\xi_i^*$$

KKT conditions for SVR

$$L(\mathbf{w},\alpha,\alpha^*,\mu,\mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + \cancel{C\sum_i(\xi_i+\xi_i^*)} + \sum_{i=1}^{m}\alpha_i\left(\cancel{y_i} - \mathbf{w}^\top\phi(\mathbf{x}_i) - \cancel{b-\epsilon-\xi_i}\right) +$$

$$\sum_{i=1}^{m}\alpha_i^*\left(\cancel{b} + \mathbf{w}^\top\phi(\mathbf{x}_i) \cancel{- y_i - \epsilon - \xi_i^*}\right) - \sum_{i=1}^{m}\cancel{\mu_i\xi_i} - \sum_{i=1}^{m}\cancel{\mu_i^*\xi_i^*}$$

$(\hat{w}, \hat{\alpha}, \hat{\alpha}_i^*, \ldots\ldots)$ soln to KKT

Differentiating the Lagrangian w.r.t

- $\mathbf{w}$: $\hat{w} - \sum_i \hat{\alpha}_i \phi(x_i) + \sum_i \hat{\alpha}_i^* \phi(x_i) = 0$ (At optimality)

$$\hat{w} = \sum_i (\hat{\alpha}_i - \hat{\alpha}_i^*)\phi(x_i)$$

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \cancel{\frac{1}{2}\|\mathbf{w}\|^2} + C\sum_i(\xi_i + \cancel{\xi_i^*}) + \sum_{i=1}^m \alpha_i \left(\cancel{y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon} - \xi_i\right) +$$

$$\sum_{i=1}^m \alpha_i^* \left(\cancel{b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*}\right) - \sum_{i=1}^m \mu_i\xi_i - \cancel{\sum_{i=1}^m \mu_i^*\xi_i^*}$$

Differentiating the Lagrangian w.r.t

- $\nabla_{\mathbf{w}} L$

  $\mathbf{w}$: $\mathbf{w} - \alpha_i\phi(\mathbf{x}_i) + \alpha_i^*\phi(\mathbf{x}_i) = 0$ *i.e.*, $\mathbf{w} = \sum_{i=1}^m (\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$
- ξ_i : $\dfrac{\partial L}{\partial \xi_i} = \dfrac{\partial}{\partial \xi_i}\Big(\cancel{C\sum_{j \neq i}\xi_j} + C\xi_i - \cancel{\sum_{j \neq i}\alpha_j\xi_j} - \alpha_i\xi_i - \cancel{\sum_{j\neq i}\mu_j\xi_j} - \mu_i\xi_i\Big)$

  $= C - \hat{\alpha}_i - \hat{\mu}_i = 0 \Rightarrow \hat{\alpha}_i + \hat{\mu}_i = C$

& given $\alpha_i \mu_i \geq 0$ $\Rightarrow \hat{\alpha}_i, \hat{\mu}_i \in [0, C]$

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i \left(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) +$$
$$\sum_{i=1}^m \alpha_i^* \left(b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i \xi_i - \sum_{i=1}^m \mu_i^* \xi_i^*$$

Differentiating the Lagrangian w.r.t

- $\mathbf{w}$: $\mathbf{w} - \alpha_i \phi(\mathbf{x}_i) + \alpha_i^* \phi(\mathbf{x}_i) = 0$ *i.e.,* $\mathbf{w} = \sum_{i=1}^m (\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$
- ξ_i : $C - \alpha_i - \mu_i = 0$ *i.e.,* $\alpha_i + \mu_i = C$
- ξ_i^* : $C - \alpha_i^* - \mu_i^* = 0$ i.e. $\alpha_i^* + \mu_i^* = C$

KKT conditions for SVR

$$L(\mathbf{w},\alpha,\alpha^*,\mu,\mu^*) = \cancel{\frac{1}{2}\|\mathbf{w}\|^2} + C\cancel{\sum_i(\xi_i+\xi_i^*)} + \sum_{i=1}^m \alpha_i\left(\cancel{y_i - \mathbf{w}^\top\phi(\mathbf{x}_i)} - b \cancel{- \epsilon - \xi_i}\right) +$$

$$\sum_{i=1}^m \alpha_i^*\left(b \cancel{+ \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*}\right) \cancel{- \sum_{i=1}^m \mu_i\xi_i - \sum_{i=1}^m \mu_i^*\xi_i^*}$$

Differentiating the Lagrangian w.r.t

- $\mathbf{w}$: $\mathbf{w} - \alpha_i\phi(\mathbf{x}_i) + \alpha_i^*\phi(\mathbf{x}_i) = 0$ *i.e.*, $\mathbf{w} = \sum_{i=1}^m (\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$
- ξ_i : $C - \alpha_i - \mu_i = 0$ *i.e.*, $\alpha_i + \mu_i = C$
- ξ_i^* : $\alpha_i^* + \mu_i^* = C$
- b : $-\sum_i \alpha_i + \sum_i \alpha_i^* = \sum_i (\alpha_i^* - \alpha_i) = 0$

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i\left(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) +$$
$$\sum_{i=1}^m \alpha_i^*\left(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i\xi_i - \sum_{i=1}^m \mu_i^*\xi_i^*$$

Differentiating the Lagrangian w.r.t

$\nabla L = 0$

- $\mathbf{w}$: $\mathbf{w} - \alpha_i\phi(\mathbf{x}_i) + \alpha_i^*\phi(\mathbf{x}_i) = 0$ *i.e.*, $\mathbf{w} = \sum_{i=1}^m(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$
- ξ_i : $C - \alpha_i - \mu_i = 0$ *i.e.*, $\alpha_i + \mu_i = C$
- ξ_i^* : $\alpha_i^* + \mu_i^* = C$
- b : $\sum_i(\alpha_i^* - \alpha_i) = 0$
- Complimentary slackness:

$\hat{\alpha}_i\left(y_i - \hat{w}^T\phi(x_i) - \hat{b} - \epsilon - \hat{\xi}_i\right) = 0$, $\hat{\alpha}_i^*\left(\hat{b} + \hat{w}^T\phi(x_i) - y_i - \epsilon - \hat{\xi}_i^*\right) = 0$

$\hat{\mu}_i\hat{\xi}_i = 0$ $\hat{\mu}_i^*\hat{\xi}_i^* = 0$

KKT conditions for SVR (One way to solve SVR ... we want to see MORE)

$$L(\mathbf{w},\alpha,\alpha^*,\mu,\mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i+\xi_i^*) + \sum_{i=1}^m \alpha_i\left(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) +$$
$$\sum_{i=1}^m \alpha_i^*\left(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i\xi_i - \sum_{i=1}^m \mu_i^*\xi_i^*$$

Differentiating the Lagrangian w.r.t

- $\mathbf{w}$: $\mathbf{w} - \alpha_i\phi(\mathbf{x}_i) + \alpha_i^*\phi(\mathbf{x}_i) = 0$ *i.e.*, $\mathbf{w} = \sum_{i=1}^m(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$
- ξ_i : $C - \alpha_i - \mu_i = 0$ *i.e.*, $\alpha_i + \mu_i = C$
- ξ_i^* : $\alpha_i^* + \mu_i^* = C$
- b : $\sum_i(\alpha_i^* - \alpha_i) = 0$
- Complimentary slackness:
  $\alpha_i(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$ AND $\mu_i\xi_i = 0$ AND
  $\alpha_i^*(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*) = 0$ AND $\mu_i^*\xi_i^* = 0$

Quadratic constraints

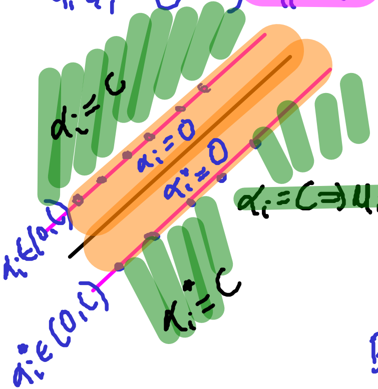
Algos such as Primal-Dual Path following (Predictor-corrector) solve these directly to find soln

Conclusions from the KKT conditions:

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

$\because \mu_i \xi_i = 0 \Rightarrow \xi_i = 0 \Rightarrow y_i - w^T\phi(x_i) - b - \epsilon - \xi_i = 0$

$\Rightarrow y_i - w^T\phi(x_i) - b - \epsilon = 0$



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

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

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

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

$\alpha_i^* = 0 \Rightarrow \mu_i^* = C, \xi_i^* = 0, w^T\phi(x_i) + b - y_i - \epsilon \leq 0$

By symmetry,

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

Recap: $w = \sum_i (\alpha_i - \alpha_i^*) \phi(x_i)$

We know from prev slide

Points within ϵ-band have $\alpha_i = \alpha_i^* = 0$

$\Rightarrow$ w wont change by perturbation of pts within ϵ-band

Just an observation.. Not soln.

KKT conditions

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,
  $w - \alpha_i \phi(\mathbf{x}_i) + \alpha_i^* \phi(\mathbf{x}_i) = 0$
  i.e. $\mathbf{w} = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) \phi(\mathbf{x}_i)$
- 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,
  $\sum_i^m (\alpha_i^* - \alpha_i) = 0$
- Complimentary slackness:
  $\alpha_i (y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$

SVR KKT Conditions: Necessary and Sufficient for Optimality

For Support Vector Regression, since the original objective and the constraints are convex, any $(\mathbf{w}, b, \alpha, \alpha^*, \mu, \mu^*, \xi, \xi^*)$ that satisfy the necessary KKT conditions gives optimality (conditions are also sufficient)

Therefore primal-dual path following algos directly solve SVR KKT to find soln

Some observations based on KKT conditions

$$\alpha_i(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$$

and

$$\alpha_i^*(b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*) = 0$$

$\Rightarrow$?

Some observations based on KKT conditions

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

$$(C - \alpha_i)\xi_i = 0 \Rightarrow ?$$

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

$$(C - \alpha_i^*)\xi_i^* = 0 \Rightarrow ?$$

More observations

- $\alpha_i, \alpha_i^* \geq 0$, $\mu_i, \mu_i^* \geq 0$, $\alpha_i + \mu_i = C$ and $\alpha_i^* + \mu_i^* = C$
  Thus, $\alpha_i, \mu_i, \alpha_i^*, \mu_i^* \in [0, C], \forall i$
- If $0 < \alpha_i < C$, then $0 < \mu_i < C$
  (as $\alpha_i + \mu_i = C$)
- $\mu_i \xi_i = 0$ and $\alpha_i(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$ are complementary slackness conditions
  So $0 < \alpha_i < C \Rightarrow \xi_i = 0$ and $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b = \epsilon + \xi_i = \epsilon$
  - All such points lie on the boundary of the ϵ band
  - Using any point $\mathbf{x}_j$ (that is with $\alpha_j \in (0, C)$) on margin, we can recover b as:
    $b = y_j - \mathbf{w}^\top \phi(\mathbf{x}_j) - \epsilon$ → for every $\alpha_i \in (0, C)$

Numerically sensitive.. In practice $b = avg\,(y_i - w^T\phi(x_i) - \epsilon)$ for $\alpha_i \in (0, C)$

Support Vector Regression

Next: Dual Objective