Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 12 - Support Vector Regression and its Dual using Optimization Principles, Kernel Trick

Recap: Lagrange Function 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. $\forall i$,
  $$y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \le \epsilon + \xi_i,$$
  $$b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \le \epsilon + \xi_i^*,$$
  $$\xi_i, \xi_i^* \ge 0$$
- Consider corresponding lagrange multipliers α_i, α_i^*, μ_i and μ_i^*
- The Lagrange Function is $L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) =$

  ξ_i, ξ_i^*

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

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

Recap: KKT conditions for the Constrained (Convex) Problem

Assume the ˆ on values of $\left\{\underline{\hat{\mathbf{w}}, \hat{b}, \hat{\xi}, \hat{\xi}^*, \hat{\alpha}, \hat{\alpha}^*, \hat{\mu}, \hat{\mu}^*}\right\}$ at KKT when not explicitly specified

Recap: Necessary and Sufficient SVR 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:
  $\alpha_i (y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$

Set of pts Qualified by KKT conditions (& therefore wearing "^")

Support Vectors: Non-zero contribution $\alpha_i - \alpha_i^*$ outside ϵ-band

[Analysis at the KKT constraint set/soln pts ?]

- **For any point $(\mathbf{x}_i, y_i)$, the product $\alpha_i\alpha_i^* = 0$.**

Suppose $\alpha_i > 0$ $\alpha_i^* > 0$

$$\cancel{y_i} - \cancel{w^T\phi(x_i)} - \cancel{b} - \epsilon - \xi_i = 0$$

$$+ \quad -\cancel{y_i} + \cancel{w^T\phi(x_i)} + \cancel{b} - \epsilon - \xi_i^* = 0$$

} By complementary slackness

$$-2\epsilon = \xi_i + \xi_i^*$$

$\because$ KKT includes original constraint set as well

$\xi_i, \xi_i^* \geq 0$

A contradiction!!!

Support Vectors: Non-zero contribution $\alpha_i - \alpha_i^*$ outside ϵ-band

- **For any point $(\mathbf{x}_i, y_i)$, the product $\alpha_i \alpha_i^* = 0$.**
  - Let $\alpha_i > 0$ and $\alpha_i^* > 0$. This leads to a contradiction.
  - By Complimentary slackness, $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i = 0$ AND $b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^* = 0$. Adding up the two equalities gives us: $\xi_i + \xi_i^* = -2\epsilon$.
  - Since only one of ξ_i and ξ_i^* can be non-zero, $\Longrightarrow$ the non-zero component is negative, which is a contradiction since $\xi_i, \xi_i^* \geq 0$
  - Thus, $\alpha_i - \alpha_i^* \propto \max\{\alpha_i, \alpha_i^*\}$
- **For points within the ϵ-insensitive tube $\alpha_i = 0$ and $\alpha_i^* = 0$:**

We saw this last time

Support Vectors: Non-zero contribution $\alpha_i - \alpha_i^*$ outside ϵ-band

- **For any point $(\mathbf{x}_i, y_i)$, the product $\alpha_i \alpha_i^* = 0$.**
  - Let $\alpha_i > 0$ and $\alpha_i^* > 0$. This leads to a contradiction.
  - By Complimentary slackness, $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i = 0$ AND $b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^* = 0$. Adding up the two equalities gives us: $\xi_i + \xi_i^* = -2\epsilon$.
  - Since only one of ξ_i and ξ_i^* can be non-zero, $\Longrightarrow$ the non-zero component is negative, which is a contradiction since $\xi_i, \xi_i^* \geq 0$
  - Thus, $\alpha_i - \alpha_i^* \propto \max\{\alpha_i, \alpha_i^*\}$
- **For points within the ϵ-insensitive tube $\alpha_i = 0$ and $\alpha_i^* = 0$:**
  - If $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i < 0$, then $\alpha_i = 0$, $\mu_i = C$ and $\xi_i = 0$. Similarly, $b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon < 0$ leading to $\alpha_i^* = 0$.

Support Vectors: Non-zero contribution $\alpha_i - \alpha_i^*$ outside ϵ-band

- $\alpha_i = C$ **and** $\alpha_i^* = C$ **correspond to points lying either outside or on the** $\epsilon-$**tube:**

We have seen this

Support Vectors: Non-zero contribution $\alpha_i - \alpha_i^*$ outside ϵ-band

- **$\alpha_i = C$ and $\alpha_i^* = C$ correspond to points lying either outside or on the $\epsilon-$tube:**
  - If $\alpha_i = C$, then $\mu_i = 0$ and $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon = \xi_i \geq 0$.
  - Similarly, $\alpha_i^* = C$ corresponds to points lying below (or beyond) the lower $\epsilon-$band.
- **For points on boundary of the ϵ-insensitive tube $\alpha_i \in [0, C]$:**

We argued that $\alpha_i \in (0, C)$ should correspond to pts "on the boundary" of ϵ-tube!

Support Vectors: Non-zero contribution $\alpha_i - \alpha_i^*$ outside ϵ-band

- **$\alpha_i = C$ and $\alpha_i^* = C$ correspond to points lying either outside or on the $\epsilon-$tube:**
  - If $\alpha_i = C$, then $\mu_i = 0$ and $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon = \xi_i \geq 0$.
  - Similarly, $\alpha_i^* = C$ corresponds to points lying below (or beyond) the lower $\epsilon-$band.
- **For points on boundary of the ϵ-insensitive tube $\alpha_i \in [0, C]$:**
  - For any point on the upper margin, $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon = 0$ and $\xi_i = 0 \implies \mu_i \geq 0 \implies \alpha_i \in [0, C]$. Similarly, $\alpha_i^* \in [0, C]$ for points lying on the margin of the lower $\epsilon-$band.

Support Vector Regression (SVR)

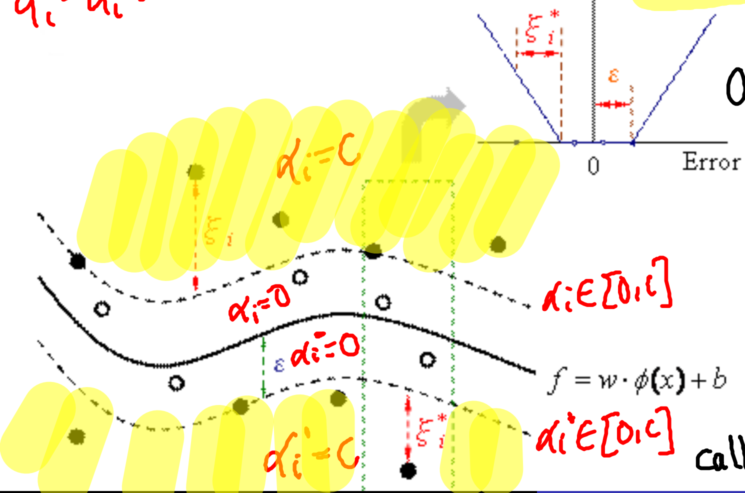
$\alpha_i - \alpha_i^* \propto \max(\alpha_i, \alpha_i^*)$

$w = \sum(\alpha_i - \alpha_i^*)\,\phi(x_i)$

$\Downarrow$

Only contribution to w is from pts on the bndry or outside the bndry

These pts are called SUPPORT VECTORS



Recap: Retrieving solution for b

- $\mu_i \xi_i = 0$ 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:
    $b = y_j - \mathbf{w}^\top \phi(\mathbf{x}_j) - \epsilon$

Hereafter we will not bother abt b much

Support Vector Regression

Dual Objective

Weak Duality and SVR

Defined for $\alpha, \alpha^*, \mu, \mu^* \geq 0$

- $L^*(\alpha, \alpha^*, \mu, \mu^*) = \min_{\mathbf{w}, b, \xi, \xi^*} L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$
- By weak duality theorem, for any $\alpha_i, \alpha_i^* \geq 0$ and $\mu_i, \mu_i^* \geq 0$:

(For $\alpha, \alpha^*, \mu, \mu^* \geq 0$) $L^*(\alpha, \alpha^*, \mu, \mu^*) \leq \min_{w, b, \xi, \xi^*} L(w \ldots)$

Weak Duality and SVR

- $L^*(\alpha, \alpha^*, \mu, \mu^*) = \min_{\mathbf{w}, b, \xi, \xi^*} L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$
- By weak duality theorem, for any $\alpha_i, \alpha_i^* \geq 0$ and $\mu_i, \mu_i^* \geq 0$:
  $\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*) \geq L^*(\alpha, \alpha^*, \mu, \mu^*)$ s.t $\alpha, \alpha^*, \mu, \mu^* \geq 0$
  s.t. $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$, and $\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$
  and $\xi_i, \xi^* \geq 0, \forall i = 1, \ldots, n$
- Thus,

L.HS. $\geq \max_{\alpha, \alpha^*, \mu, \mu^* \geq 0}$ RHS

Weak Duality and SVR

- $L^*(\alpha, \alpha^*, \mu, \mu^*) = \min_{\mathbf{w}, b, \xi, \xi^*} L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$
- By weak duality theorem, for any $\alpha_i, \alpha_i^* \geq 0$ and $\mu_i, \mu_i^* \geq 0$:
$$\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m (\xi_i + \xi_i^*) \geq L^*(\alpha, \alpha^*, \mu, \mu^*)$$
s.t. $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$, and $\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$
and $\xi_i, \xi^* \geq 0, \forall i = 1, \ldots, n$
- Thus,
$$\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m (\xi_i + \xi_i^*) \geq \max_{\alpha, \alpha^*, \mu, \mu^*} L^*(\alpha, \alpha^*, \mu, \mu^*)$$
s.t. $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$, and $\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$
and $\xi_i, \xi^* \geq 0, \forall i = 1, \ldots, n$

SVR Dual objective

Discussed earlier

- Assume: By convexity, KKT conditions are necessary and sufficient and strong duality holds (for $\alpha, \alpha^* \geq 0$): Now
  $\min_{\mathbf{w},b,\xi,\xi^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) = \max_{\alpha,\alpha^*,\mu,\mu^*} L^*(\alpha, \alpha^*, \mu, \mu^*)$
  s.t. $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$, and $w^\top\phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$
  and $\xi_i, \xi^* \geq 0, \forall i = 1, \ldots, n$
- This value is precisely obtained at the $\left\{\hat{\mathbf{w}}, \hat{b}, \hat{\xi}, \hat{\xi}^*, \hat{\alpha}, \hat{\alpha}^*, \hat{\mu}, \hat{\mu}^*\right\}$ that satisfies the necessary (and sufficient) KKT optimality conditions [**KKT Constraint Set**]

SVR Dual objective (contd)

- For $\alpha, \alpha^* \geq 0$ and $\left\{\hat{\mathbf{w}}, \hat{b}, \hat{\xi}, \hat{\xi}^*, \hat{\alpha}, \hat{\alpha}^*, \hat{\mu}, \hat{\mu}^*\right\}$ from [**KKT Constraint Set**]:
$$\min_{\mathbf{w},b,\xi,\xi^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^m(\xi_i + \xi_i^*) = \max_{\alpha,\alpha^*,\mu,\mu^*} L^*(\alpha, \alpha^*, \mu, \mu^*)$$
s.t. $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$, and $w^\top\phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$
and $\xi_i, \xi^* \geq 0, \forall i = 1, \ldots, n$
- Given strong duality, we can equivalently solve:
$$\max_{\hat{\alpha},\hat{\alpha}^*,\hat{\mu},\hat{\mu}^*} \underbrace{L^*(\hat{\alpha}, \hat{\alpha}^*, \hat{\mu}, \hat{\mu}^*)}_{\text{L evaluated along KKT constraint set}}$$

- $L(\hat{\alpha}, \hat{\alpha}^*, \hat{\mu}, \hat{\mu}^*) = \frac{1}{2}\|\hat{\mathbf{w}}\|^2 + C\sum_{i=1}^{m}(\hat{\xi}_i + \hat{\xi}_i^*) + \sum_{i=1}^{m}\Big(\hat{\alpha}_i(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \hat{\xi}_i) + \hat{\alpha}_i^*(\mathbf{w}^\top\phi(\mathbf{x}_i) + b - y_i - \epsilon - \hat{\xi}_i^*$ $\sum_{i=1}^{m}(\hat{\mu}_i\hat{\xi}_i + \hat{\mu}_i^*\hat{\xi}_i^*)$

Substitute KKT into L(. . .)

- We obtain $\hat{\mathbf{w}}$, $\hat{b}$, $\hat{\xi}_i$, $\hat{\xi}_i^*$ in terms of $\hat{\alpha}$, $\hat{\alpha}^*$, $\hat{\mu}$ and $\hat{\mu}^*$ by using the KKT conditions derived earlier as $\hat{\mathbf{w}} = \sum_{i=1}^{m}(\hat{\alpha}_i - \hat{\alpha}_i^*)\phi(\mathbf{x}_i)$ and $\sum_{i=1}^{m}(\hat{\alpha}_i - \hat{\alpha}_i^*) = 0$ and $\hat{\alpha}_i + \hat{\mu}_i = C$ and $\hat{\alpha}_i^* + \hat{\mu}_i^* = C$

Dropping the messy $\hat{}$ hat notation...

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

- Invoking $\mathbf{w} = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$ and $\sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0$ and $\alpha_i + \mu_i = C$ and $\alpha_i^* + \mu_i^* = C$, we ~~get~~ simplify using

$$\frac{1}{2}\|w\|^2 = \frac{1}{2}\sum_i\sum_j(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi(x_i)^\top\phi(x_j)$$

$$\sum_i \xi_i(C - \alpha_i - \mu_i) = 0$$

$$b\left(\sum_i -\alpha_i + \alpha_i^*\right) = 0$$

$$\sum_i(\alpha_i y_i - \alpha_i^* y_i) = \sum_i y_i(\alpha_i - \alpha_i^*) \neq 0$$

Dropping the messy ˆ hat notation...

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

- Invoking $\mathbf{w} = \sum_{i=1}^m(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$ and $\sum_{i=1}^m(\alpha_i - \alpha_i^*) = 0$ and $\alpha_i + \mu_i = C$ and $\alpha_i^* + \mu_i^* = C$, we get $L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*) = \cancel{\frac{1}{2}\sum_i\sum_j(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j)} + \sum_i \cancel{(\xi_i(C - \alpha_i - \mu_i)} + \cancel{\xi_i^*(C - \alpha_i^* - \mu_i^*)}) - \cancel{b\sum_i(\alpha_i - \alpha_i^*)} - \epsilon\sum_i(\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*) - \frac{1}{2}\sum_i\sum_j(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j)$ [Only in terms of dual vars]

Developing further..

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

Developing further..

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

SVR Dual using only dot products $\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$

- $\mathbf{w} = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi(x_i) \Rightarrow$ the final decision function $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b =$ $\sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi^T(\mathbf{x}_i)\phi(\mathbf{x}) + \overbrace{y_j - \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j) - \epsilon}^{b}$ $\mathbf{x}_j$ is any point with $\alpha_j \in (0, C)$.
- The dual optimization problem to compute the α's for SVR is:

SVR Dual using only dot products $\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$

- $\mathbf{w} = \sum_{i=1}^m (\alpha_i - \alpha_i^*)\phi(x_i) \Rightarrow$ the final decision function $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b =$
$\sum_{i=1}^m (\alpha_i - \alpha_i^*)\phi^T(\mathbf{x}_i)\phi(\mathbf{x}) + y_j - \sum_{i=1}^m (\alpha_i - \alpha_i^*)\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j) - \epsilon$
$\mathbf{x}_j$ is any point with $\alpha_j \in (0, C)$.
- The dual optimization problem to compute the α's for SVR is:
    - $\max_{\alpha_i, \alpha_i^*} \; -\frac{1}{2}\sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j) - \epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*)$
    - s.t $\sum_i (\alpha_i - \alpha_i^*) = 0$ & $\alpha_i, \alpha_i^* \in [0, C]$
- **We notice that the only way these three expressions involve ϕ is through $\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j)$, for some i, j**

Kernel function: $K(\mathbf{x}_i, \mathbf{x}_j) = \phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$

- We call $\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j)$ a **kernel function**: $K(\mathbf{x}_i, \mathbf{x}_j) = \phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j)$
- The Kernel Trick: For some important choices of ϕ, compute $K(\mathbf{x}_i, \mathbf{x}_j)$ directly and more efficiently than having to explicitly compute/enumerate $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$
- The expression for decision function becomes $f(x) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$ → Similarity of query pt x with training pt x_i
- Computation of α_i is specific to the objective function being minimized: Closed form exists for Ridge regression but NOT for SVR

Kernel Ridge Regression (Tut 5)

The Kernelized version of SVR

- The kernelized dual problem:

$$\max_{\alpha_i, \alpha_i^*} \; -\frac{1}{2}\sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(\mathbf{x}_i, \mathbf{x}_j)$$

$$-\epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*)$$

  - such that $\sum_i (\alpha_i - \alpha_i^*) = 0$ and $\alpha_i, \alpha_i^* \in [0, C]$
- Kernelized decision function: $f(\mathbf{x}) = \sum_i (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b$
- Using any $\mathbf{x}_j$ with $\alpha_j \in (0, C)$: $b = y_j - \sum_i (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}_j)$
- Computing $K(\mathbf{x}_1, \mathbf{x}_2)$ often does not even require computing $\phi(\mathbf{x}_1)$ or $\phi(\mathbf{x}_2)$ explicitly

Tutorial 5: Derive kernelized expression for Ridge Regression

Tutorial 5: Kernelizing Ridge Regression

$$\Phi = \begin{bmatrix} \phi(x_1) \\ \phi(x_m) \end{bmatrix} \qquad \Phi^T\Phi = \left[\sum_k \phi_i(x_k)\phi_j(x_k) \right] \neq \left[K(x_i, x_j) \right]$$

$(i,j)^{th}$ entry

- Given $w = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T y$ and using the identity $(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = PB^T(BPB^T + R)^{-1}$
  - $\Rightarrow w = \Phi^T(\Phi\Phi^T + \lambda I)^{-1} y = \sum_{i=1}^m \alpha_i \phi(x_i)$ where $\alpha_i = \left((\Phi\Phi^T + \lambda I)^{-1} y\right)_i$
  - $\Rightarrow$ the final decision function $f(\mathbf{x}) = \phi^T(\mathbf{x})\mathbf{w} = \sum_{i=1}^m \alpha_i \phi^T(\mathbf{x})\phi(\mathbf{x}_i)$
- Again, **We notice that the only way the decision function $f(\mathbf{x})$ involves ϕ is through $\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j)$, for some i, j**

Hint: Try identity for $P, B, R \in \mathbb{R}^{1\times 1}$ (ie scalars)

$$\Phi\Phi^T = \left[K(x_i, x_j) \right]$$

Basis function expansion and the Kernel trick

- We began with functional form called *basis function expansion*[1]

$$f(\mathbf{x}) = \sum_{j=1}^{p} w_j \phi_j(\mathbf{x})$$

- And landed up with an equivalent form for Ridge and SVR

$$f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$$

- Aside: For $p \in [0, \infty)$, with what K, kind of regularizers, loss functions, *etc.*, will these dual representations hold?[2]

[1] Each ϕ_j is called a *basis function*. b can be absorbed in ϕ. See Section 2.8.3 of Tibshi

[2] Section 5.8.1 of Tibshi.

The Representer Theorem & Reproducing Kernel Hilbert Space (RKHS) [Optional Slide]

Hilbert space is ∞ dimension extension of $\mathbb{R}^n$

1. The solution $f^* \in \mathcal{H}_K$ (Hilbert space) to the following problem

$\Omega(\cdot)$ is monotonically increasing

$$f^* = \arg\min_{f \in \mathcal{H}_K} \sum_{i=1}^{m} \mathbf{E}\left(f\left(\mathbf{x}^{(i)}\right), y^{(i)}\right) + \Omega(\|f\|_K)$$

can be always written as $f^*(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}^{(i)})$, provided $\Omega(\|f\|_K)$ is a ....

Eucledian norm

Eucledian space

$\langle x, x \rangle = x^T x = \|x\|_2^2 \quad \forall x \in \mathbb{R}^n$

Can I write for L_1 norm $\|x\|_1^p = \langle x, x \rangle$

The Representer Theorem & Reproducing Kernel Hilbert Space (RKHS)

1. More specifically, if $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b$ and $K(\mathbf{x}', \mathbf{x}) = \phi^T(\mathbf{x})\phi(\mathbf{x}')$ then the solution $\mathbf{w}^* \in \Re^n$ to the following problem

Finite dim Hilbert space

Hint: Don't care for differentiability of E

$$(\mathbf{w}^*, b^*) = \arg\min_{\mathbf{w},b} \sum_{i=1}^{m} \mathbf{E}\left(f\left(\mathbf{x}^{(i)}\right), y^{(i)}\right) + \Omega(\|\mathbf{w}\|_2)$$

H/W: Express E for SVR

can be always written as $\phi^T(\mathbf{x})\mathbf{w}^* + b = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}^{(i)})$, provided $\Omega(\|\mathbf{w}\|_2)$ is a monotonically increasing function of $\|\mathbf{w}\|_2$. $\Re^n$ is the Hilbert space and $K(., \mathbf{x}) : \mathcal{X} \rightarrow \Re$ is the **Reproducing (RKHS) Kernel**