Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 14 - Positive Definite Kernels, Mercer's Theorem

Recap: The Gram (Kernel) Matrix

- For any dataset $\underbrace{\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}}$ (Fixed set of m pts) and for any m, the Gram matrix $\mathcal{K}$ is defined as

$$\mathcal{K} = \begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & \ldots & K(\mathbf{x}_1, \mathbf{x}_n) \\ \ldots & K(\mathbf{x}_i, \mathbf{x}_j) & \ldots \\ K(\mathbf{x}_m, \mathbf{x}_1) & \ldots & K(\mathbf{x}_m, \mathbf{x}_m) \end{bmatrix}$$

- Claim: If $\mathcal{K}_{ij} = K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ are entries of an $n \times n$ **Gram Matrix** $\mathcal{K}$ then
  - $\mathcal{K}$ must be positive semi-definite → In this area, referred to as positive def.
  - Proof: $\mathbf{b}^T \mathcal{K} \mathbf{b} = \sum_{i,j} b_i \mathcal{K}_{ij} b_j = \sum_{i,j} b_i b_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$

$$= \langle \sum_i b_i \phi(\mathbf{x}_i), \sum_j b_j \phi(\mathbf{x}_j) \rangle = ||\sum_i b_i \phi(\mathbf{x}_i)||_2^2 \geq 0$$

Recap: Basis expansion ϕ for symmetric K?

Let us see an eg.

- *Positive-definite kernel:* For any dataset $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ and for any m, the Gram matrix $\mathcal{K}$ must be positive definite so that $\mathcal{K} = U\Sigma U^T = (U\Sigma^{\frac{1}{2}})(U\Sigma^{\frac{1}{2}})^T = RR^T$ where rows of U are linearly independent and Σ is a positive diagonal matrix
- *Mercer kernel:* Extending to eigenfunction decomposition[1]:

$$K(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{\infty} \alpha_j \phi_j(\mathbf{x}_1)\phi_j(\mathbf{x}_2) \rightarrow [\sqrt{\alpha_1}\phi_1(.), \sqrt{\alpha_2}\phi_2(.) \dots]$$

where $\alpha_j \geq 0$ and $\sum_{j=1}^{\infty} \alpha_j^2 < \infty$

(is $\exists$ $j \in \{1 .. \infty\}$ & scalars functions α_j & $\phi_j(.)$)

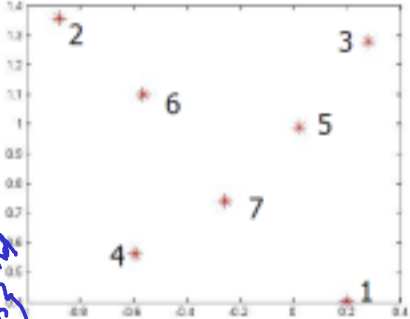
Eigen values

Eigen fns

- *Mercer kernel* and *Positive-definite kernel* turn out to be equivalent if the input space $\{x\}$ is *compact*[2]

1 Eigen-decomposition wrt linear operators. See https://en.wikipedia.org/wiki/Mercer%27s_theorem

2 That is, if every Cauchy sequence is convergent.



Finite example

Choose 7 2D points

Choose a kernel k

w.r.t matrix

Empirically we will check if this "gram" matrix can give me back a ϕ

w.r.t function

Later we recover ϕ in a principled manner

Function does not depend on $\{x_1 .. x_m\}$

$K_{ij} = \exp(-|x_i - x_j|^2/10)$ can be calculated.

K=	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	1.0000	0.8131	0.9254	0.9369	0.9630	0.8987	0.9683
x_2	0.8131	1.0000	0.8745	0.9312	0.9102	0.9837	0.9264
x_3	0.9254	0.8745	1.0000	0.8806	0.9851	0.9286	0.9440
⋮	0.9369	0.9312	0.8806	1.0000	0.9457	0.9714	0.9857
	0.9630	0.9102	0.9851	0.9457	1.0000	0.9653	0.9862
⋮	0.8987	0.9837	0.9286	0.9714	0.9653	1.0000	0.9779
x_7	0.9683	0.9264	0.9440	0.9857	0.9862	0.9779	1.0000

Matrix depends on $\{x_1 .. x_m\}$

$[U,D]=\mathrm{svd}(\mathbf{K}),\ UDU^T=\mathbf{K},\ UU^T=I$

U =

-0.3709	0.5499	0.3392	0.6302	0.0992	-0.1844	-0.0633
-0.3670	-0.6596	-0.1679	0.5164	0.1935	0.2972	0.0985
-0.3727	0.3007	-0.6704	-0.2199	0.4635	-0.1529	0.1862
-0.3792	-0.1411	0.5603	-0.4709	0.4938	0.1029	-0.2148
-0.3851	0.2036	-0.2248	-0.1177	-0.4363	0.5162	-0.5377
-0.3834	-0.3259	-0.0477	-0.0971	-0.3677	-0.7421	-0.2217
-0.3870	0.0673	0.2016	-0.2071	-0.4104	0.1628	0.7531

D =

6.6315	0	0	0	0	0	0
0	0.2331	0	0	0	0	0
0	0	0.1272	0	0	0	0
0	0	0	0.0066	0	0	0
0	0	0	0	0.0016	0	0
0	0	0	0	0	0.000	0
0	0	0	0	0	0	0.000

Eigenvalues of K (diag elements of D) are nonnegative

Mapped points=sqrt(D)*U^T

Mapped points =

First row of $U * \sqrt{\lambda_1}$

-0.9551	-0.9451	-0.9597	-0.9765	-0.9917	-0.9872	-0.9966
0.2655	-0.3184	0.1452	-0.0681	0.0983	-0.1573	0.0325
0.1210	-0.0599	-0.2391	0.1998	-0.0802	-0.0170	0.0719
0.0511	0.0419	-0.0178	-0.0382	-0.0095	-0.0079	-0.0168
0.0040	0.0077	0.0185	0.0197	-0.0174	-0.0146	-0.0163
-0.0011	0.0018	-0.0009	0.0006	0.0032	-0.0045	0.0010
-0.0002	0.0004	0.0007	-0.0008	-0.0020	-0.0008	0.0028
$\phi(x_1)$	$\phi(x_2)$	$\phi(x_3)$	$\phi(x_4)$	$\phi(x_5)$	$\phi(x_6)$	$\phi(x_7)$

This ϕ is not a function. It is defined ONLY on these 7 points and not for any other pt

You can check now that

$$\langle \phi(x_i), \phi(x_j) \rangle \doteq \phi(x_i)^T \phi(x_j) = \exp(-|x_i - x_j|^2/10) \ \forall i, j$$

We want a $\phi(\cdot)$ that is a function (eigenfunction) OR atleast assured existence of such a ϕ.

Mercer and Positive Definite Kernels

Can we show that $\exists$ a ϕ for the function $\exp\left(-\frac{|x_i - x_j|^2}{10}\right) = \phi^T(x_i)\,\phi(x_j)$

without depending on some subset $\{x_1 \ldots x_m\}$

Mercer's theorem

- **Mercer kernel:** $K(\mathbf{x}_1, \mathbf{x}_2)$ is a Mercer kernel if $\int_{x_1} \int_{x_2} K(\mathbf{x}_1, \mathbf{x}_2) g(\mathbf{x}_1) g(\mathbf{x}_2)\, d\mathbf{x}_1 d\mathbf{x}_2 \geq 0$ for all square integrable functions $g(\mathbf{x})$ Extending psd definition using quadratic expansion to functions
  ($g(\mathbf{x})$ is square integrable *iff* $\int (g(\mathbf{x}))^2\, dx$ is finite)
- **Mercer's theorem:** (Assures existence of such ϕ)
  An implication of the theorem:
  for any *Mercer kernel* $K(\mathbf{x}_1, \mathbf{x}_2)$, $\exists\, \phi(\mathbf{x}) : \mathbb{R}^n \mapsto H$,
  s.t. $K(\mathbf{x}_1, \mathbf{x}_2) = \phi^\top(\mathbf{x}_1)\phi(\mathbf{x}_2)$
  - where H is a *Hilbert space*, the infinite dimensional version of the Eucledian space, which is......
  - $(\Re^n, < ., . >)$ where $< ., . >$ is the standard dot product in $\Re^n$
  - Advanced: Formally, *Hibert Space* is an inner product space with associated norms, where every Cauchy sequence is convergent

Prove that $(\mathbf{x}_1^\top \mathbf{x}_2)^d$ is Mercer kernel ($d \in \mathbb{Z}^+, d \geq 1$)

- We want to prove that $\int_{\mathbf{x}_1}\int_{\mathbf{x}_2}(\mathbf{x}_1^\top \mathbf{x}_2)^d g(\mathbf{x}_1)g(\mathbf{x}_2)\,d\mathbf{x}_1 d\mathbf{x}_2 \geq 0$, for all square integrable functions $g(\mathbf{x})$
- Here, $\mathbf{x}_1$ and $\mathbf{x}_2$ are vectors s.t $\mathbf{x}_1, \mathbf{x}_2 \in \Re^t$
- Thus, $\int_{\mathbf{x}_1}\int_{\mathbf{x}_2}(\mathbf{x}_1^\top \mathbf{x}_2)^d g(\mathbf{x}_1)g(\mathbf{x}_2)\,d\mathbf{x}_1 d\mathbf{x}_2$

$$= \int_{x_{11}} .. \int_{x_{1t}} \int_{x_{21}} .. \int_{x_{2t}} \left[\sum_{n_1..n_t} \frac{d!}{n_1!..n_t!} \prod_{j=1}^{t} (x_{1j}x_{2j})^{n_j}\right] g(x_1)g(x_2)\,dx_{11}..dx_{1t}dx_{21}..dx_{2t}$$

\# of ways of sampling from d $(n_1 .. n_t)$ with n_j objects of type $x_{1j}x_{2j}$

s.t. $\sum_{i=1}^{t} n_i = d$

(taking a leap)

You can also leave it as some fn $f(n_1 \ldots n_t)$

Prove that $(\mathbf{x}_1^\top \mathbf{x}_2)^d$ is Mercer kernel ($d \in \mathbb{Z}^+$, $d \geq 1$)

$\sum_n \int_x = \int_x \sum_n$

$$= \sum_{n_1 \dots n_t} \frac{d!}{n_1! \dots n_t!} \int_{\mathbf{x}_1} \int_{\mathbf{x}_2} \prod_{j=1}^{t} (x_{1j} x_{2j})^{n_j} g(x_1) g(x_2)\, dx_1 dx_2$$

$$= \sum_{n_1 \dots n_t} \frac{d!}{n_1! \dots n_t!} \int_{\mathbf{x}_1} \int_{\mathbf{x}_2} (x_{11}^{n_1} x_{12}^{n_2} \dots x_{1t}^{n_t}) g(x_1) (x_{21}^{n_1} x_{22}^{n_2} \dots x_{2t}^{n_t}) g(x_2)\, dx_1 dx_2$$

Treat x_2 as x_1 by change of var. names

$$= \sum_{n_1 \dots n_t} \frac{d!}{n_1! \dots n_t!} \left(\int_{x_1} x_{11}^{n_1} x_{12}^{n_2} \dots x_{1t}^{n_t} g(x_1)\, dx_1 \right) \left(\int_{x_2} x_{21}^{n_1} x_{22}^{n_2} \dots x_{2t}^{n_t} g(x_2)\, dx_2 \right)$$

Note: $\int_x \int_y g(x) g(y)\, dx\, dy = \left(\int_x g(x)\, dx \right) \left(\int_y g(y)\, dy \right) = \left(\int_x g(x)\, dx \right) \left(\int_x g(x)\, dx \right)$

Prove that $(\mathbf{x}_1^\top \mathbf{x}_2)^d$ is Mercer kernel ($d \in \mathbb{Z}^+$, $d \geq 1$)

$$= \sum_{n_1 \dots n_t} \frac{d!}{n_1! \dots n_t!} \int_{\mathbf{x}_1} \int_{\mathbf{x}_2} \prod_{j=1}^{t} (x_{1j} x_{2j})^{n_j} \, g(x_1) g(x_2) \, dx_1 dx_2$$

$$= \sum_{n_1 \dots n_t} \frac{d!}{n_1! \dots n_t!} \int_{\mathbf{x}_1} \int_{\mathbf{x}_2} (x_{11}^{n_1} x_{12}^{n_2} \dots x_{1t}^{n_t}) g(x_1) \, (x_{21}^{n_1} x_{22}^{n_2} \dots x_{2t}^{n_t}) g(x_2) \, dx_1 dx_2$$

$$= \sum_{n_1 \dots n_t} \frac{d!}{n_1! \dots n_t!} \left(\int_{\mathbf{x}_1} (x_{11}^{n_1} \dots x_{1t}^{n_t}) g(x_1) \, dx_1 \right) \left(\int_{\mathbf{x}_2} (x_{21}^{n_1} \dots x_{2t}^{n_t}) g(x_2) \, dx_2 \right)$$

(integral of decomposable product as product of integrals)

s.t. $\sum_{i}^{t} n_i = d$

Prove that $(\mathbf{x}_1^\top \mathbf{x}_2)^d$ is Mercer kernel $(d \in \mathbb{Z}^+, d \geq 1)$

- Realize that both the integrals are basically the same, with different variable names
- Thus, the equation becomes:

$$\boxed{\sum_{n_1 \ldots n_t} \frac{d!}{n_1! \ldots n_t!} \left(\int_{\mathbf{x}_1} (x_{11}^{n_1} \ldots x_{1t}^{n_t}) g(x_1)\, dx_1 \right)^2 \geq 0}$$

(the square is non-negative for reals)

- Thus, we have shown that $(\mathbf{x}_1^\top \mathbf{x}_2)^d$ is a Mercer kernel.

Recall: We showed that for $d=2$ & $t=2$, a $\phi(\cdot)$ exists.
We have now more general result

What about $\sum_{d=1}^{r} \alpha_d (\mathbf{x}_1^\top \mathbf{x}_2)^d$ s.t. $\alpha_d \geq 0$?

(Note: $\int_x \sum_y p(x,y) = \sum_y \int_x p(x,y)$)

- $K(\mathbf{x}_1, \mathbf{x}_2) = \sum_{d=1}^{r} \alpha_d (\mathbf{x}_1^\top \mathbf{x}_2)^d$
- Is $\int_{\mathbf{x}_1} \int_{\mathbf{x}_2} \left(\sum_{d=1}^{r} \alpha_d (x_1^\top x_2)^d \right) g(\mathbf{x}_1) g(\mathbf{x}_2)\, d\mathbf{x}_1 d\mathbf{x}_2 \geq 0$?

We have shown this is ≥ 0

$= \sum_d \alpha_d \int_{x_1} \int_{x_2} (x_1^\top x_2)^d g(x_1) g(x_2)\, dx_1 dx_2 \Rightarrow \geq 0$

What about $\sum_{d=1}^{r} \alpha_d (\mathbf{x}_1^\top \mathbf{x}_2)^d$ s.t. $\alpha_d \geq 0$?

- Is $\int_{\mathbf{x}_1} \int_{\mathbf{x}_2} \left(\sum_{d=1}^{r} \alpha_d (x_1^\top x_2)^d \right) g(\mathbf{x}_1) g(\mathbf{x}_2) \, d\mathbf{x}_1 d\mathbf{x}_2 \geq 0$?
- We have

$$\int_{\mathbf{x}_1} \int_{\mathbf{x}_2} \left(\sum_{d=1}^{r} \alpha_d (\mathbf{x}_1^\top \mathbf{x}_2)^d \right) g(x_1) g(x_2) \, dx_1 dx_2 =$$

What about $\sum_{d=1}^{r} \alpha_d (\mathbf{x}_1^\top \mathbf{x}_2)^d$ s.t. $\alpha_d \geq 0$?

- Is $\int_{\mathbf{x}_1} \int_{\mathbf{x}_2} \left(\sum_{d=1}^{r} \alpha_d (x_1^\top x_2)^d \right) g(\mathbf{x}_1) g(\mathbf{x}_2)\, d\mathbf{x}_1 d\mathbf{x}_2 \geq 0$?
- We have

$$\int_{\mathbf{x}_1} \int_{\mathbf{x}_2} \left(\sum_{d=1}^{r} \alpha_d (\mathbf{x}_1^\top \mathbf{x}_2)^d \right) g(x_1) g(x_2)\, dx_1 dx_2 =$$

$$\sum_{d=1}^{r} \alpha_d \int_{\mathbf{x}_1} \int_{\mathbf{x}_2} (\mathbf{x}_1^\top \mathbf{x}_2)^d g(\mathbf{x}_1) g(\mathbf{x}_2)\, d\mathbf{x}_1 d\mathbf{x}_2$$

What about $\sum_{d=1}^{r} \alpha_d (\mathbf{x}_1^\top \mathbf{x}_2)^d$ s.t. $\alpha_d \geq 0$?

- Since $\alpha_d \geq 0$, $\forall d$ and since we have already proved that $\int_{\mathbf{x}_1}\int_{\mathbf{x}_2} (\mathbf{x}_1^\top \mathbf{x}_2)^d g(\mathbf{x}_1) g(\mathbf{x}_2)\, d\mathbf{x}_1 d\mathbf{x}_2 \geq 0$
- We must have,

$$\sum_{d=1}^{r} \alpha_d \int_{\mathbf{x}_1}\int_{\mathbf{x}_2} (\mathbf{x}1^\top \mathbf{x}_2)^d g(\mathbf{x}_1) g(\mathbf{x}_2)\, d\mathbf{x}_1 d\mathbf{x}_2 \geq 0$$

- By which, $K(\mathbf{x}_1, \mathbf{x}_2) = \sum_{d=1}^{r} \alpha_d (\mathbf{x}_1^\top \mathbf{x}_2)^d$ is a Mercer kernel.
- Examples of Mercer Kernels: Linear Kernel, Polynomial Kernel, Radial Basis Function Kernel

r=1

$$K(x_1, x_2) = \exp\left(-\gamma \|x_1 - x_2\|^2\right)$$

Recall Taylor series (polynomial) expansion for

$$e^t = \sum_{n=0}^{\infty} \frac{t^n}{n!}$$

Suggest you apply Taylor series only on terms involving <x1,x2>

$$K(x_1, x_2) = \exp\left(-\gamma\left(\|x_1\|^2 + \|x_2\|^2 - 2x_1^T x_2\right)\right)$$

$$= \underbrace{\exp\left(-\gamma\left(\|x_1\|^2\right)\right)\exp\left(-\gamma\|x_2\|^2\right)}_{\text{these terms seperate out into the two integrals and are the same within integrals}} \exp\left(2\gamma x_1^T x_2\right)$$

$$\int_{x_1}\int_{x_2} g(x_1)\, g(x_2) \exp\left(-\Gamma \|x_1 - x_2\|_2^2\right) dx_1\, dx_2$$

$$= \int_{x_1}\int_{x_2} \overbrace{g(x_1) \exp\left(-\Gamma \|x_1\|^2\right)}^{g'(x_1)} \overbrace{g(x_2) \exp\left(-\Gamma \|x_2\|^2\right)}^{g'(x_2)} \times \underbrace{\exp\left(2\Gamma\, x_1^T x_2\right)}_{\sum_{n=0}^{\infty} (2\Gamma)^n \frac{(x_1^T x_2)^n}{n!}} dx_1\, dx_2$$

$$= \sum_{n=0}^{\infty} \frac{(2\Gamma)^n}{n!} \overbrace{\int_{x_1}\int_{x_2} g'(x_1)\, g'(x_2) \left(x_1^T x_2\right)^n dx_1\, dx_2}^{\text{we showed this is } \geq 0}$$

Closure properties of Kernels

A different way of proving valid kernel helps in some other circumstances

Let $K_1(\mathbf{x}_1, \mathbf{x}_2)$ and $K_2(\mathbf{x}_1, \mathbf{x}_2)$ be positive definite (valid) kernels. Then the following are also kernels.

- $\alpha_1 K_1(\mathbf{x}_1, \mathbf{x}_2) + \alpha_2 K_2(\mathbf{x}_1, \mathbf{x}_2)$ for $\alpha_1, \alpha_2 \geq 0$.

  Proof: Hint: K_1 & K_2 have $\phi(\cdot)$ & $\tilde{\phi}(\cdot)$

$$K_1(x_1,x_2) = \phi^T(x_1)\phi(x_2) \quad K_2(x_1,x_2) = \tilde{\phi}^T(x_1)\tilde{\phi}(x_2)$$

Then $\begin{bmatrix} \sqrt{\alpha_1}\,\phi(\cdot) \\ \sqrt{\alpha_2}\,\tilde{\phi}(\cdot) \end{bmatrix}$ is the basis function (ϕ) for $\alpha_1 K_1(\cdot) + \alpha_2 K_2(\cdot)$

Closure properties of Kernels

Let $K_1(\mathbf{x}_1, \mathbf{x}_2)$ and $K_2(\mathbf{x}_1, \mathbf{x}_2)$ be positive definite (valid) kernels. Then the following are also kernels.

- $\alpha_1 K_1(\mathbf{x}_1, \mathbf{x}_2) + \alpha_2 K_2(\mathbf{x}_1, \mathbf{x}_2)$ for $\alpha_1, \alpha_2 \geq 0$.
  Proof:
- $K_1(\mathbf{x}_1, \mathbf{x}_2) K_2(\mathbf{x}_1, \mathbf{x}_2) = \left(\sum_i \phi_i(x_1)\phi_i(x_2)\right)\left(\sum_j \tilde{\phi}_j(x_1)\tilde{\phi}_j(x_2)\right)$
  Proof:

$$= \sum_i \sum_j \underbrace{\phi_i(x_1)\tilde{\phi}_j(x_1)}_{\phi_{new}(x_1)} \underbrace{\phi_i(x_2)\tilde{\phi}_j(x_2)}_{\phi_{new}(x_2)}$$

$$\begin{bmatrix} \phi_1(\cdot)\tilde{\phi}_1(\cdot) \\ \phi_1(\cdot)\tilde{\phi}_2(\cdot) \\ \vdots \end{bmatrix}$$

Kernels in SVR

How to solve to find α_i, α_i^* ?

- Recall:
  $max_{\alpha_i,\alpha_i^*} - \frac{1}{2}\sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\mathbf{x}_i, \mathbf{x}_j) - \epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*)$
  and the decision function:
  $f(x) = \sum_i (\alpha_i - \alpha_i^*)K(\mathbf{x}_i, \mathbf{x}) + b$
  are all in terms of the kernel $K(\mathbf{x}_i, \mathbf{x}_j)$ only
- *One can now employ any mercer kernel in SVR or Ridge Regression to implicitly perform linear regression in higher dimensional spaces*

Solving the SVR Dual Optimization Problem

- The SVR dual objective is:
  $max_{\alpha_i,\alpha_i^*} - \frac{1}{2}\sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j)$
  $-\epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*)$ such that $\sum_i (\alpha_i - \alpha_i^*) = 0$,
  $\alpha_i, \alpha_i^* \in [0, C]$
- This is a linearly constrained quadratic program (LCQP), just like the

[3] https://en.wikipedia.org/wiki/Quadratic_programming#Solvers_and_scripting_.28programming.29_languages

Solving the SVR Dual Optimization Problem

- The SVR dual objective is:
$max_{\alpha_i,\alpha_i^*} - \frac{1}{2}\sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j)$
$-\epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*)$ such that $\sum_i (\alpha_i - \alpha_i^*) = 0$,
$\alpha_i, \alpha_i^* \in [0, C]$
- This is a linearly constrained quadratic program (LCQP), just like the constrained version of Lasso
- There exists no closed form solution to this formulation
- Standard QP (LCQP) solvers[3] can be used
- Question: Are there more specific and efficient algorithms for solving SVR in this form?

[3] https://en.wikipedia.org/wiki/Quadratic_programming#Solvers_and_scripting_.28programming.29_languages

Sequential Minimial Optimization Algorithm for Solving SVR

Solving the SVR Dual Optimization Problem

- It can be shown that the objective:
  $max_{\alpha_i,\alpha_i^*} - \frac{1}{2}\sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(x_i, x_j)$
  $-\epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*)$
- can be written as:
  $max_{\beta_i} - \frac{1}{2}\sum_i \sum_j \beta_i \beta_j K(\mathbf{x}_i, \mathbf{x}_j) - \epsilon \sum_i |\beta_i| + \sum_i y_i \beta_i$
  s.t.

$\beta_i = \alpha_i - \alpha_i^*$

$\alpha_i + \alpha_i^* = |\beta_i|$

$\sum_i \alpha_i - \alpha_i^* = \sum_i \beta_i = 0$

Solving the SVR Dual Optimization Problem

- It can be shown that the objective:
  $max_{\alpha_i,\alpha_i^*} - \frac{1}{2}\sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j)$
  $-\epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*)$
- can be written as:
  $max_{\beta_i} - \frac{1}{2}\sum_i \sum_j \beta_i\beta_j K(\mathbf{x}_i, \mathbf{x}_j) - \epsilon \sum_i |\beta_i| + \sum_i y_i\beta_i$
  s.t.
  - $\sum_i \beta_i = 0$ ✓
  - $\beta_i \in [-C, C], \forall i$ ✓

  [Keep all β's constant from prev iteration except β_i, β_j & solve]
- Even for this form, standard QP (LCQP) solvers[4] can be used
- Question: How about (iteratively) solving for two β_i's at a time?
  - This is the idea of the Sequential Minimal Optimization (SMO) algorithm

Sequential Minimal Optimization (SMO) for SVR

- Consider:
  $max_{\beta_i} - \frac{1}{2} \sum_i \sum_j \beta_i \beta_j K(\mathbf{x}_i, \mathbf{x}_j) - \epsilon \sum_i |\beta_i| + \sum_i y_i \beta_i$
  s.t.
  - $\sum_i \beta_i = 0$
  - $\beta_i \in [-C, C], \forall i$
- The SMO subroutine can be defined as:

Sequential Minimal Optimization (SMO) for SVR

$-\beta_i(\Sigma \quad) - \beta_j(\Sigma \quad)$

- Consider:
  $max_{\beta_i} - \frac{1}{2}\sum_i \sum_j \beta_i \beta_j K(\mathbf{x}_i, \mathbf{x}_j) - \epsilon \sum_i |\beta_i| + \sum_i y_i \beta_i$
  s.t.
  - $\sum_i \beta_i = 0$
  - $\beta_i \in [-C, C], \forall i$

  $\beta_i + \beta_j = \text{Constant}$

  $\Rightarrow \beta_j = \text{Constant} - \beta_i$
- The SMO subroutine can be defined as:
  1. Initialise $\beta_1, \ldots, \beta_n$ to some value $\in [-C, C]$
  2. Pick β_i, β_j to estimate closed form expression for next iterate (i.e. $\beta_i^{new}, \beta_j^{new}$)
  3. Check if the KKT conditions are satisfied
     - If not, choose β_i and β_j that worst violate the KKT conditions and reiterate