Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 13 - Kernel Trick, Positive Definite Kernels, Mercer's Theorem

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

- $\mathbf{w} = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi(x_i) \Rightarrow$ 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 $\alpha_j \in (0, C)$.

  b evaluated using (x_j, y_j)
- 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**

Recap: 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 $f(x) = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$
- Computation of α_i is specific to the objective function being minimized: Closed form exists for Ridge regression but NOT for SVR

Recap: The Kernelized version of SVR

- The kernelized dual problem:

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

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

  - such that $\sum_i(\alpha_i-\alpha_i^*)=0$ and $\alpha_i,\alpha_i^*\in[0,C]$
- Kernelized decision function: $f(\mathbf{x})=\sum_i(\alpha_i-\alpha_i^*)\underline{K(\mathbf{x}_i,\mathbf{x})}+b$
- Using any $\mathbf{x}_j$ with $\alpha_j\in(0,C)$: $b=y_j-\sum_i(\alpha_i-\alpha_i^*)K\underline{(\mathbf{x}_i,\mathbf{x}_j)}$
- 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: Kernelizing Ridge Regression

- 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**

Recap: Basis function expansion and Kernel

- 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.

Recap: 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

   Obviously holds for ridge regression

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

   Note: Result is wrt point of optimal soln

   can be always written as $\underline{\phi^T(\mathbf{x})\mathbf{w}^* + b^* = \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}^{(i)})}$, 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**

The Representer Theorem and SVR

1. The SVR solution $(\mathbf{w}^*, b^*, \xi_i^*) =$

$$\arg\min_{\mathbf{w}, b, \xi_i} C \sum_{i=1}^{m} (\xi_i + \xi_i^*) + \frac{1}{2} \|\mathbf{w}\|_2^2$$

s.t. $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \le \epsilon + \xi_i$, and $\longrightarrow$ $\xi_i = \max(0, y_i - w^T\phi(x_i) - b - \epsilon)$

$w^\top \phi(\mathbf{x}_i) + b - y_i \le \epsilon + \xi_i^*$, and $\longrightarrow$ $\xi_i^* = \max(0, w^T\phi(x_i) + b - y_i - \epsilon)$

$\xi_i, \xi^* \ge 0, \forall i = 1, \ldots, n$

2. Can be rewritten as $(\mathbf{w}^*, b^*, \xi_i^*) =$

$$\arg\min_{w, b, \xi_i, \xi_i^*} C \sum_{i=1}^{m} \max(0, y_i - w^T\phi(x_i) \ldots) + \max(0, w^T\phi(x_i) - y_i \ldots) + \frac{1}{2}\|w\|^2$$

The Representer Theorem and SVR

1. The SVR solution $(\mathbf{w}^*, b^*, \xi_i^*) =$

$$\arg\min_{\mathbf{w},b,\xi_i} C\sum_{i=1}^{m}(\xi_i + \xi_i^*) + \frac{1}{2}\|\mathbf{w}\|_2^2$$

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
$\xi_i, \xi^* \geq 0, \forall i = 1, \ldots, n$

2. Can be rewritten as $(\mathbf{w}^*, b^*, \xi_i^*) =$

$$\arg\min_{\mathbf{w},b,\xi_i} C\sum_{i=1}^{m}\underbrace{\max\left\{-\epsilon \pm \left(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b\right), 0\right\}}_{E(f(x_i),\,y_i)} + \underbrace{\frac{1}{2}\|\mathbf{w}\|_2^2}_{\Omega(\|w\|_2)}$$

The Representer Theorem and SVR (contd.)

1. If $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b$ and $K(\mathbf{x}', \mathbf{x}) = \phi^T(\mathbf{x})\phi(\mathbf{x}')$ and given the SVR solution $(\mathbf{w}^*, b^*, \xi_i^*) =$

$$\arg\min_{\mathbf{w},b,\xi_i} C\sum_{i=1}^{m} \max\left\{-\epsilon \pm \left(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b\right), 0\right\} + \frac{1}{2}\|\mathbf{w}\|_2^2$$

2. Setting $\mathbf{E}\left(f\left(\mathbf{x}^{(i)}\right), y^{(i)}\right) = \underline{C\max\left\{-\epsilon \pm \left(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b\right)\right.}, 0\}$ and $\Omega(\|\mathbf{w}\|_2) = \frac{1}{2}\|\mathbf{w}\|_2^2$, we can apply the Representer theorem to SVR, so that $\phi^T(\mathbf{x})\mathbf{w}^* + b = \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}^{(i)})$

An Example Kernel

$x_1, x_2 \in \mathbb{R}^p$

Time to compute $K(x_1, x_2)$

$\underbrace{(1+x_1^T x_2)^2}_{\text{2 mult + addition + square}} \ll \underbrace{\phi^T(x_1)\phi(x_2)}_{\text{6 mult + addition}}$

- Let $K(\mathbf{x}_1, \mathbf{x}_2) = (1 + \mathbf{x}_1^\top \mathbf{x}_2)^2$
- Which value of $\phi(\mathbf{x})$ will yield $\phi^\top(\mathbf{x}_1)\phi(\mathbf{x}_2) = K(\mathbf{x}_1, \mathbf{x}_2) = (1 + \mathbf{x}_1^\top \mathbf{x}_2)^2$
- Is such a ϕ guaranteed to exist?
- Is there a unique ϕ for given K?

For $p=2$, $K(x_1, x_2) = (1 + x_{11}x_{21} + x_{12}x_{22})^2$

$x_1 = \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix}$ $x_2 = \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix}$

$= (1 + x_{11}^2 x_{21}^2 + x_{12}^2 x_{22}^2 + 2x_{11}x_{21}x_{12}x_{22} + 2x_{11}x_{21} + 2x_{12}x_{22}) = \phi^T(x_1)\phi(x_2)$

$2 = \sqrt{2}\sqrt{2}$

An Example Kernel

- We can prove that such a ϕ exists
- For example, for a 2-dimensional $\mathbf{x}_i$:

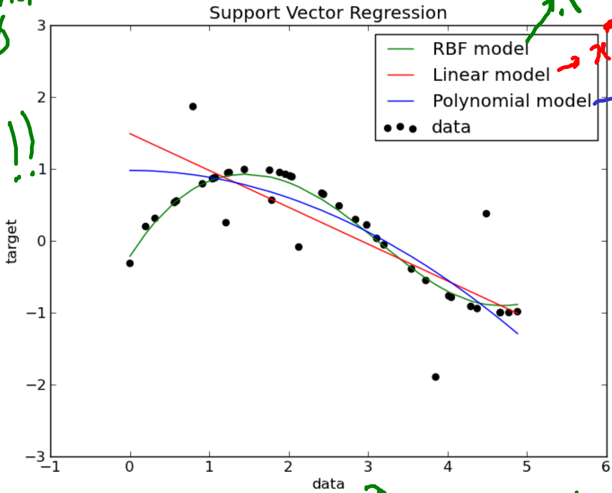
$$\phi(\mathbf{x}_i) = \begin{bmatrix} 1 \\ x_{i1}\sqrt{2} \\ x_{i2}\sqrt{2} \\ x_{i1}x_{i2}\sqrt{2} \\ x_{i1}^2 \\ x_{i2}^2 \end{bmatrix}$$

Any permutation of this $\phi(.)$ is also an answer.

- $\phi(\mathbf{x}_i)$ exists in a 6-dimensional space
- But, to compute $K(\mathbf{x}_1, \mathbf{x}_2)$, all we need is $x_1^\top x_2$ without having to enumerate $\phi(\mathbf{x}_i)$

Computing this fn is quick and easy. Using ϕ explicitly is impractical!!

$$\exp\left(\frac{-\|x_1 - x_2\|_2^2}{2\sigma^2}\right)$$



$x_1^T x_2$

$$\sum_{i=1}^{p} \alpha_i \left(x_1^T x_2\right)^i$$

To get $K(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right)$ you need an infinite dimensional $\phi(\cdot)$

More on the Kernel Trick

- **Kernels** operate in a *high-dimensional*, *implicit* feature space without necessarily computing the coordinates of the data in that space, but rather by simply computing the Kernel function
- This operation is often computationally cheaper than the explicit computation of the coordinates

The Gram (Kernel) Matrix

- For any dataset $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ and for any m, the Gram matrix $\mathcal{K}$ is defined as

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

(from ridge regression)

- 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 (for any m & any choice of $x_1 \ldots x_m$)
  - Proof:

$$v^T K v = \sum_{i,j} v_i K(x_i, x_j) v_j$$

$$= \sum_{i,j} v_i \phi^T(x_i) \phi(x_j) v_j$$

The Gram (Kernel) Matrix

- For any dataset $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ and for any m, the Gram matrix $\mathcal{K}$ is defined as

At the least $\mathcal{K}$ must be symmetric

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

- 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

option 1: $\sum_k \phi_k(x_i)\phi_k(x_j)$

  - $\mathcal{K}$ must be positive semi-definite
  - Proof: $\mathbf{b}^T\mathcal{K}\mathbf{b} = \sum_{i,j} b_i \mathcal{K}_{ij} b_j = \sum_{i,j} b_i b_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$

option 2: Property of inner prod

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

SAME!!! Note: i & j are indices into same "b"

Existence of basis expansion ϕ for symmetric K?

For kernels, it means positive semi-definite

- *Positive-definite kernel:* For any dataset $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ 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

$$K = \begin{bmatrix} \ddots & & \\ & k(x_i, x_j) & \\ & & \ddots \end{bmatrix} \succcurlyeq 0$$

$$\sqrt{\Sigma} = \Sigma^{1/2} = (\Sigma^{1/2})^T$$

$R \equiv$ analogous to Φ

$$K = U\Sigma U^T = \underbrace{U\sqrt{\Sigma}}_{R}\sqrt{\Sigma}U^T$$

$\Sigma_{ii} = i^{th}$ eigenvalue & Σ is diag.

U = matrix of eigenvectors.

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

We want to extend the eigenvalue decomposition applied to p.d matrix K.

to "eigen function" decomposition applied to "p.d function" $K(\cdot,\cdot)$

Recap positive definiteness for matrix : $\forall \underset{\mathbb{R}^m}{b},\ b^T \underset{\mathbb{R}^{m\times m}}{K} b \geq 0$

Extending this to function $\underset{K(\cdot,\cdot)\to\mathbb{R}}{}$ $K(\cdot,\cdot)$: $\forall g(\cdot)\ \int_{x_1,x_2} g(x_1) K(x_1,x_2) g(x_2)\, dx_1 dx_2 \geq 0$

We generalize

$\forall b \in \mathbb{R}^m$, $\quad b^T k b = \sum_i \sum_j b_i k_{ij} b_j \geq 0$

uncountably infinite sized generalizations

To

$\forall g(\cdot) \to \mathbb{R}$ $\quad \int_{x_i}\int_{x_j} g(x_i) k(x_i, x_j) g(x_j)\, dx_i\, dx_j \geq 0$

View $g(\cdot)$ as some uncountably infinite sized generalization of vector b

Existence of basis expansion ϕ for symmetric K?

- *Positive-definite kernel:* For any dataset $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ 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[3]:

$$K(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^{\infty} \alpha_j \phi_j(\mathbf{x}_1)\phi_j(\mathbf{x}_2) \text{ where } \alpha_j \geq 0 \text{ and}$$

$\sum_{j=1}^{\infty} \alpha_j^2 < \infty$

- *Mercer kernel* and *Positive-definite kernel* turn out to be equivalent if the input space $\{x\}$ is *compact*[4]

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

Mercer and Positive Definite Kernels, SMO Algorithm

Mercer's theorem

- **Mercer kernel:** $K(\mathbf{x}_1, \mathbf{x}_2)$ is a Mercer kernel if $\int\int K(\mathbf{x}_1, \mathbf{x}_2) g(\mathbf{x}_1) g(\mathbf{x}_2)\, d\mathbf{x}_1 d\mathbf{x}_2 \geq 0$ for all square integrable functions $g(\mathbf{x})$
  ($g(\mathbf{x})$ is square integrable *iff* $\int (g(\mathbf{x}))^2\, dx$ is finite)

  like saying b should have finite norm in $b^T K b$
- **Mercer's theorem:**
  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*[5], the infinite dimensional version of the Eucledian space.
  - Eucledian space: $(\Re^n, <.,.>)$ where $<.,.>$ is the standard dot product in $\Re^n$
  - Advanced: Formally, *Hibert Space* is an inner product space with

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, $\int_{\mathbf{x}_1} \int_{\mathbf{x}_2} (\mathbf{x}_1^\top \mathbf{x}_2)^d g(\mathbf{x}_1) g(\mathbf{x}_2)\, d\mathbf{x}_1 d\mathbf{x}_2$

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

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

(taking a leap)