

Lecture 13: More on Kernels, PSD kernels, Mercer Kernels, etc

Instructor: Prof. Ganesh Ramakrishnan

Recall the Kernelized version of SVR

- The kernelized dual problem:

$$\begin{aligned} \max_{\alpha_i, \alpha_i^*} & -\frac{1}{2} \sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(\mathbf{x}_i, \mathbf{x}_j) \\ & -\epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i (\alpha_i - \alpha_i^*) \end{aligned}$$

s.t.

- ▶ $\sum_i (\alpha_i - \alpha_i^*) = 0$
- ▶ $\alpha_i, \alpha_i^* \in [0, C]$
- The kernelized decision function:
 $f(\mathbf{x}) = \sum_i (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b$
- Using any point \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

An example Kernel

- Let $K(x_1, x_2) = (1 + x_1^\top x_2)^2$
- What $\phi(x)$ will give $\phi^\top(x_1)\phi(x_2) = K(x_1, x_2) = (1 + x_1^\top x_2)^2$
- Is such a ϕ guaranteed to exist?
- Is there a unique ϕ for given K ?

- We can prove that such a ϕ exists
- For example, for a 2-dimensional x_i :

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

- $\phi(x_i)$ exists in a 5-dimensional space
- Thus, to compute $K(x_1, x_2)$, all we need is $x_1^\top x_2$, and there is no need to compute $\phi(x_i)$

Introduction to the Kernel Trick

- **Kernels** operate in a *high-dimensional, implicit* feature space without ever computing the coordinates of the data in that space, but rather by simply computing the Kernel function
- This approach is called the "*kernel trick*" and will talk about *valid kernels*
- This operation is often computationally cheaper than the explicit computation of the coordinates
- 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

- ▶ Proof: $\mathbf{b}^T \mathcal{K} \mathbf{b} = \sum_{i,j} b_i \mathcal{K}_{ij} b_j = \sum_{i,j} b_i b_j \langle \phi(x_i), \phi(x_j) \rangle$
 $= \langle \sum_i b_i \phi(x_i), \sum_j b_j \phi(x_j) \rangle = \left\| \sum_i b_i \phi(x_i) \right\|_2^2 \geq 0$

Basis function expansion and the Kernel trick

- We started off with the functional form¹

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

Each ϕ_j is called a *basis function* and this representation is called *basis function expansion*²

- And we landed up with an equivalent

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

for Ridge regression and Support Vector Regression

- For $p \in [0, \infty)$, with what K , kind of regularizers, loss functions, etc., will these dual representations hold?³

¹The additional b term can be either absorbed in ϕ or kept separate as discussed on several occasions.

²Section 2.8.3 of Tibshi

³Section 5.8.1 of Tibshi.

Existence of basis expansion ϕ for symmetric K ?

- *Positive-definite kernel*: For any dataset $\{x_1, x_2, \dots, x_n\}$ and for any n , the Gram matrix \mathcal{K} must be positive definite

$$\mathcal{K} = \begin{bmatrix} K(x_1, x_1) & \dots & K(x_1, x_n) \\ \dots & K(x_i, x_j) & \dots \\ K(x_n, x_1) & \dots & K(x_n, x_n) \end{bmatrix}$$

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 decomposition⁴:

$$K(x_1, x_2) = \sum_{j=1}^{\infty} \alpha_j \phi_j(x_1) \phi_j(x_2) \text{ where } \alpha_j \geq 0 \text{ and } \sum_{j=1}^{\infty} \alpha_j^2 < \infty$$

- *Mercer kernel* and *Positive-definite kernel* turn out to be equivalent if the input space $\{x\}$ is *compact*⁵

⁴Eigen-decomposition wrt linear operators. See

https://en.wikipedia.org/wiki/Mercer%27s_theorem

⁵That is, if every Cauchy sequence is convergent.

Mercer's theorem


- **Mercer kernel:** $K(x_1, x_2)$ is a Mercer kernel if $\int \int K(x_1, x_2)g(x_1)g(x_2) dx_1 dx_2 \geq 0$ for all square integrable functions $g(x)$
($g(x)$ is square integrable iff $\int (g(x))^2 dx$ is finite)

- **Mercer's theorem:**

An implication of the theorem:

for any Mercer kernel $K(x_1, x_2)$, $\exists \phi(x) : \mathbb{R}^n \mapsto H$,
s.t. $K(x_1, x_2) = \phi^\top(x_1)\phi(x_2)$

- ▶ where H is a *Hilbert space*⁶, the infinite dimensional version of the Euclidian space.
- ▶ Euclidian space: $(\mathbb{R}^n, \langle \cdot, \cdot \rangle)$ where $\langle \cdot, \cdot \rangle$ is the standard dot product in \mathbb{R}^n
- ▶ Formally, *Hilbert Space* is an inner product space with associated norms, where every Cauchy sequence is convergent

⁶Do you know Hilbert? No? Then what are you doing in his space? :) 

Prove that $(x_1^\top x_2)^d$ is a Mercer kernel ($d \in \mathbb{Z}^+$, $d \geq 1$)

- We want to prove that

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

for all square integrable functions $g(x)$

- Here, x_1 and x_2 are vectors s.t $x_1, x_2 \in \mathbb{R}^t$

- Thus, $\int_{x_1} \int_{x_2} (x_1^\top x_2)^d g(x_1) g(x_2) dx_1 dx_2$

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

s.t. $\sum_{i=1}^t n_i = d$
(taking a leap)

Prove that $(x_1^\top x_2)^d$ is a Mercer kernel ($d \in \mathbb{Z}^+$, $d \geq 1$)

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

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

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

(integral of decomposable product as product of integrals)

$$\text{s.t. } \sum_i^t n_i = d$$

Prove that $(x_1^\top x_2)^d$ is a 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:

$$\sum_{n_1 \dots n_t} \frac{d!}{n_1! \dots n_t!} \left(\int_{x_1} (x_{11}^{n_1} \dots x_{1t}^{n_t}) g(x_1) dx_1 \right)^2 \geq 0$$

(the square is non-negative for reals)

- Thus, we have shown that $(x_1^\top x_2)^d$ is a Mercer kernel.

What about $\sum_{d=1}^r \alpha_d (x_1^\top x_2)^d$ s.t. $\alpha_d \geq 0$?

- $K(x_1, x_2) = \sum_{d=1}^r \alpha_d (x_1^\top x_2)^d$
- Is $\int_{x_1} \int_{x_2} \left(\sum_{d=1}^r \alpha_d (x_1^\top x_2)^d \right) g(x_1) g(x_2) dx_1 dx_2 \geq 0$?
- We have

$$\begin{aligned} & \int_{x_1} \int_{x_2} \left(\sum_{d=1}^r \alpha_d (x_1^\top x_2)^d \right) g(x_1) g(x_2) dx_1 dx_2 \\ &= \sum_{d=1}^r \alpha_d \int_{x_1} \int_{x_2} (x_1^\top x_2)^d g(x_1) g(x_2) dx_1 dx_2 \end{aligned}$$

What about $\sum_{d=1}^r \alpha_d (x_1^\top x_2)^d$ s.t. $\alpha_d \geq 0$?

- We have already proved that $\int_{x_1} \int_{x_2} (x_1^\top x_2)^d g(x_1) g(x_2) dx_1 dx_2 \geq 0$
- Also, $\alpha_d \geq 0, \forall d$
- Thus,

$$\sum_{d=1}^r \alpha_d \int_{x_1} \int_{x_2} (x_1^\top x_2)^d g(x_1) g(x_2) dx_1 dx_2 \geq 0$$

- By which, $K(x_1, x_2) = \sum_{d=1}^r \alpha_d (x_1^\top x_2)^d$ is a Mercer kernel.
- Examples of Mercer Kernels: Linear Kernel, Polynomial Kernel, Radial Basis Function Kernel

Kernels in SVR

- Recall:

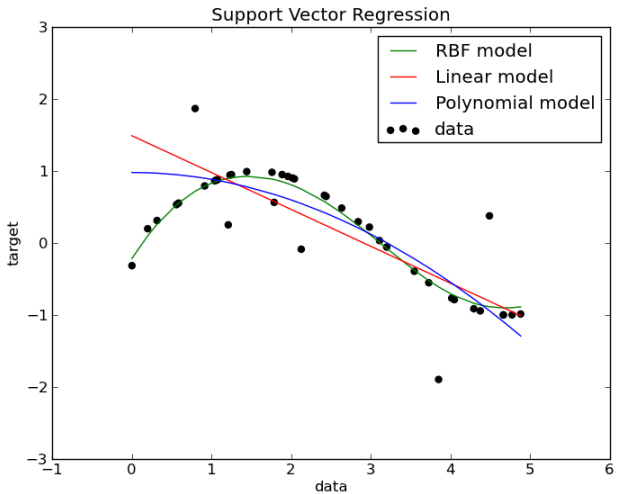
$$\max_{\alpha_i, \alpha_i^*} - \frac{1}{2} \sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(x_i, x_j) - \epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i (\alpha_i - \alpha_i^*)$$

and the decision function:

$$f(x) = \sum_i (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

are all in terms of the kernel $K(x_i, x_j)$ only

- One can now employ any Mercer kernel in SVR or Ridge Regression to implicitly perform linear regression in higher dimensional spaces*



Basis function expansion & Kernel: Part 1

We saw that for $p \in [0, \infty)$, under certain conditions on K , the following can be equivalent representations



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

- And

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

- For what kind of regularizers, loss functions and $p \in [0, \infty)$ will these dual representations hold?⁷

⁷Section 5.8.1 of Tibshirani.

Basis function expansion & Kernel: Part 2

- We could also begin with

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

and impose no constraints on K .

- *E.g.:* $K_k(x_q, x) = I(\|x_q - x\| \leq \|x_{(k)} - x_0\|)$ where $x_{(k)}$ is the training observation ranked k^{th} in distance from x and $I(S)$ is the indicator of the set S
- This is precisely the Nearest Neighbor Regression model
- Kernel regression and density models are other examples of such *local regression* methods⁸

⁸Section 2.8.2 of Tibshi