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

Instructor: Prof. Ganesh Ramakrishnan

Recall the Kernelized version of SVR

- The kernelized dual problem:

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

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

  s.t.
  - $\sum_i(\alpha_i-\alpha_i^*)=0$
  - $\alpha_i,\alpha_i^*\in[0,C]$
- The kernelized decision function:
  $f(x)=\sum_i(\alpha_i-\alpha_i^*)K(x_i,x)+b$
- Using any point x_j with $\alpha_j\in(0,C)$:
  $b=y_j-\sum_i(\alpha_i-\alpha_i^*)K(x_i,x_j)$
- Computing $K(x_1,x_2)$ often does not even require computing $\phi(x_1)$ or $\phi(x_2)$ explicitly

An example Kernel

polynomial kernel of degree 2

- Let $K(x_1, x_2) = (1 + x_1^\top x_2)^2$ → polynomial kernel of degree 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? → Yes
- 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 $\phi(x_i)$

$K(x_1, x_2) = (1 + x_1^T x_2)^d$ .. what is a ϕ ?

Eg: $x_1, x_2 \in \mathbb{R}^2$

$$(1 + x_1^T x_2)^d = (\underbrace{1}_{c_1} + \underbrace{x_{11} x_{21}}_{c_2} + \underbrace{x_{12} x_{22}}_{c_3})^d$$

$$= \sum_{\substack{n_1, n_2, n_3 \\ \text{s.t } n_1 + n_2 + n_3 = d}} \left(c_1^{n_1} c_2^{n_2} c_3^{n_3}\right) \frac{\lfloor d}{\lfloor n_1 \lfloor n_2 \lfloor n_3}$$

$$\phi(x_1) = \left[\sqrt{\frac{\lfloor d}{\lfloor n_1 \lfloor n_2 \lfloor n_3}} \left(c_1^{n_1}\right)_{\underbrace{x_1}} \left(c_2^{n_2}\right)_{\underbrace{x_1}} \left(c_3^{n_3}\right)_{\underbrace{x_1}} \right]$$

Restricted to

Verify this abstraction wrt $d = 2$ which we have already derived... Similarly you can derive ϕ for $x_1, x_2 \in \mathbb{R}^k$

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* (Extending necessary condition of psd from linear regression)
- This operation is often computationally cheaper than the explicit computation of the coordinates
- Claim: If $\mathcal{K}_{ij} = K(x_i, x_j) = \underbrace{\langle \phi(x_i), \phi(x_j) \rangle}$ $\to \phi^T(x_i)\phi(x_j)$ are entries of an $n \times n$ **Gram Matrix** $\mathcal{K}$ then
  - $\mathcal{K}$ must be positive semi-definite : K is a symmetric matrix since $\phi^T(x_i)\phi(x_j) = \phi^T(x_j)\phi(x_i)$
  - Proof: $\mathbf{b}^T\mathcal{K}\mathbf{b} = \sum_{i,j} b_i\mathcal{K}_{ij}b_j = \sum_{i,j} b_ib_j\langle\phi(x_i), \phi(x_j)\rangle$ $(\forall\, b \in \mathbb{R}^n)$

$$= \langle \sum_i b_i\phi(x_i), \sum_j b_j\phi(x_j)\rangle = ||\sum_i b_i\phi(x_i)||_2^2 \geq 0$$

Since $\sum_i\sum_j b_ib_j\phi^T(x_i)\phi(x_j) = \left(\sum_i b_i\phi(x_i)\right)^T\left(\sum_j b_j\phi(x_j)\right)$

Basis function expansion and the Kernel trick

- We started off with the functional form[1]

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

ϕ could be complex & infinite dimension

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

- And we landed up with an equivalent

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

some formulations show existence of ϕ & others dont care

K might be simple to compute

could be thought of as a basis fn $h_i(x)$

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]

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

[2]Section 2.8.3 of Tibshi

[3]Section 5.8.1 of Tibshi.

$$\sum_j w_j \phi_j(x) \xleftarrow{} f(x) \xrightarrow{} \sum_i \alpha_i K(x, x_i) \quad \left(= \sum_i \alpha_i h_i(x)\right)$$

① LHS leads to RHS for loss functions & regularizers such as Ridge regression & SVR

↳ Q: Does LHS lead to RHS for Lasso? (Tut 5)

② "Valid" K could mean existence of ϕ s.t $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$: Psd kernel fns

③ Some models directly begin with RHS form without bothering abt existence of ϕ

Nearest neighbr regression

Local regression

Non-parametric regressi-on

Existence of basis expansion ϕ for symmetric K?

- *Positive-definite kernel:* For any dataset, the Gram matrix $\mathcal{K}$ must be positive definite

If K is psd by eigenvalue decomp

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

n is finite

so that $K = U\Sigma U^T = (U\Sigma^{\frac{1}{2}})(U\Sigma^{\frac{1}{2}})^T = RR^T$ where rows of U are linearly independent and Σ is a positive diagonal matrix

- *Mercer kernel:* Extending to eigenfunction decomposition[4]:

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

function function

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

optional reading

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

[5]That is, if every Cauchy sequence is convergent.

$$K = U\Sigma U^T = \underbrace{(U\Sigma^{1/2})}_{R}\underbrace{(U\Sigma^{1/2})}_{R}{}^T$$

$$= RR^T = \begin{bmatrix} r_1^T r_1 & r_1^T r_2 & \cdots & r_1^T r_n \\ r_2^T r_1 & \cdots & & \\ & & r_i^T r_j & \\ & & \cdots & r_n^T r_n \end{bmatrix}$$

Think of $r_i = \phi(x_i)$

$\Rightarrow K_{ij} = \phi^T(x_i)\phi(x_j)$ or $\langle \phi(x_i), \phi(x_j) \rangle$

$\therefore$ If matrix K is psd, there exists ϕ for each point in the matrix s.t $K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle$

But this procedure requires that for a given kernel fn $K(x, x')$ for all $(x_1, \ldots x_n)$ & for all n, the gram matrix K must be psd for ϕ to exist

For eg: to show that $K(x, z) = (1 + x^T z)^d$ is a valid kernel, you must:

show that $\forall (x_1 \ldots x_n)$ & $\forall n$

$$K = \begin{bmatrix} (1+x_1^T x_1)^d & (1+x_1^T x_2)^d & \cdots & (1+x_1^T x_n)^d \\ & \ddots & & \\ & & (1+x_i^T x_j)^d & \\ & & & \ddots \end{bmatrix}$$

is psd ... Not always practical!

Mercer's theorem

Like saying K is p.s.d matrix iff $\forall \|a\| < \infty, a^T K a \geq 0$ ie psd kernel

- **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. $K(x_1, x_2) = \phi^\top(x_1)\phi(x_2)$

  Treat H as $\mathbb{R}$ for now

  - where H is a *Hilbert space*[6], the infinite dimensional version of the Eucledian space.
  - Eucledian space: $(\Re^n, <.,.>)$ where $< .,. >$ is the standard dot product in $\Re^n$
  - Formally, *Hibert Space* is an inner product space with associated norms, where every Cauchy sequence is convergent

[6] Do 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, $x_1, x_2 \in R^t$
- Thus, $\int_{x_1} \int_{x_2} (x_1^\top x_2)^d g(x_1) g(x_2)\, dx_1 dx_2$

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

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

(taking a leap)

show that integral of products is the product of 2 integrals that are same

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

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

No dependence

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

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

(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.

Similarly: $(1 + x_1^\top x_2)^d$ 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

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

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

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

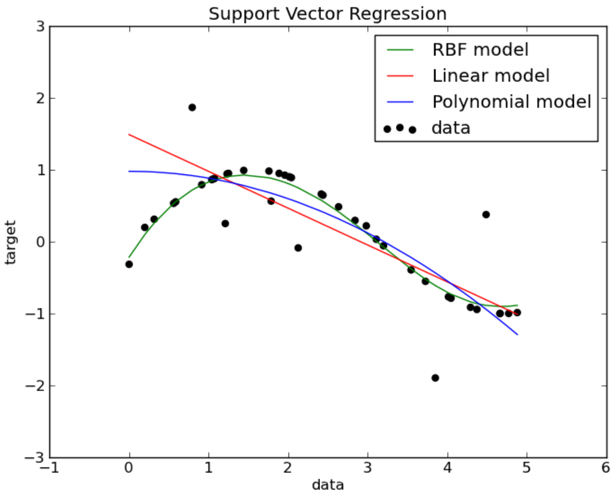
- We have already proved that $\int_{x_1} \int_{x_2} (x_1^\top x_2)^d g(x_1) g(x_2)\, dx_1 dx_2 \geq 0$
- Also, $\alpha_d \geq 0$, $\forall d$
- Thus,

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

- 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*


Support Vector Regression
RBF model
Linear model
Polynomial model
data
target
data
3
2
1
0
−1
−2
−3
−1
0
1
2
3
4
5
6

Basis function expansion & Kernel: Part 1

We saw the 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]

[7] Section 5.8.1 of Tibshi.

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 : $K_k(x_q, x) = 1$ if x_q is within k nearest nbrs of x & 0 o/w
- This is precisely the Nearest Neighbor Regression model
- Kernel regression and density models are other examples of such *local regression* methods[8]

[8] Section 2.8.2 of Tibshi