Discussion 6: Kernels
Machine Learning, Spring 2019

1 Introduction and Background

Motivation: map a point in the original feature space $\mathcal{X} \subset \mathbb{R}^d$ (let's stay real) to a high-dimensional feature space, so that the model learned in the high-dimensional space is more expressive.

First thought: use a feature map $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^D$ (let's allow D to be infinity) to add more features.

Something nice: by Mercer's theorem: for a continuous symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, if

$$\int_{\mathcal{X}\times\mathcal{X}} k(\mathbf{x}, \mathbf{z}) f(\mathbf{x}) f(\mathbf{z}) d\mathbf{x} d\mathbf{z} \geq 0$$

for all $f \in L_2[\mathcal{X}] = \left\{ f : \left(\int_{\mathcal{X}} |f(z)|^2 \, dz \right)^{\frac{1}{2}} < \infty \right\}$, then $k(\cdot, \cdot)$ can be written as an absolutely uniformly convergent series

$$k(\mathbf{x}, \mathbf{z}) = \sum_{j=1}^{\infty} \lambda_j \psi_j(\mathbf{x}) \psi_j(\mathbf{z}).$$

where $\|\psi_j\|_{L_2} = \left(\int_{\mathcal{X}} |\psi_j(z)|^2 \, dz \right)^{\frac{1}{2}} = 1$ and $\lambda_j \geq 0$.

So, we this kernel $k(\cdot, \cdot)$ is associated with the feature map: $\phi(\mathbf{x}) = [\psi_1(\mathbf{x}), \psi_2(\mathbf{x}), \psi_3(\mathbf{x}), \dots]$. This means adding features can be so simple, since we don't have to go to the actual feature space.

RKHS: for a kernel satisfying the Mercer's condition, use x $k(\cdot, x)$ and the feature map. The space

$$\mathcal{H}_k = \overline{\left\{ f(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i) : m \in \mathbb{N}, x_i \in \mathcal{X}, \alpha_i \in \mathbb{R} \right\}}$$

is a Hilbert space with inner product being: for $f = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i)$ and $g = \sum_{j=1}^{m'} \beta_j k(\cdot, x'_j)$,

$$\langle f, g \rangle_{\mathcal{H}_k} = \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j).$$

The kernel k is reproducing in the sense that $\langle k(\cdot, x), f(\cdot) \rangle_{\mathcal{H}_k} = f(x)$.

Representer Theorem: Fix a set $\mathcal{X}$, a kernel k, and let $\mathcal{H}_k$ be the corresponding RKHS. For any function $l : \mathbb{R}^2 \to \mathbb{R}$, the solutions of the optimization problem:

$$f^* \in \operatorname{argmin}_{f \in \mathcal{H}_k} \sum_{i=1}^{n} l(f(x_i), y_i) + \|f\|^2_{\mathcal{H}_k}$$

can all be expressed in the form

$$f^* = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot).$$

2 Properties

(a) If k_1 and k_2 are valid kernels, show that

(a) $k(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z})k_2(\mathbf{x}, \mathbf{z})$ is a valid kernel.

(b) $k(\mathbf{x}, \mathbf{z}) = g(\mathbf{x})g(\mathbf{z})$ for some $g : \mathcal{X} \to \mathbb{R}$ is a valid kernel.

(b) Let $\mathbf{B}$ be a matrix with negative eigenvalues. Show that $\mathbf{B}$ cannot be used to define an inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{B} \mathbf{y}$.

(c) Let $\mathbf{A}$ be a positive semi-definite (PSD) matrix. Show that an inner product defined as $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{A} \mathbf{y}$ may not be valid.

3 Visualizing kernels

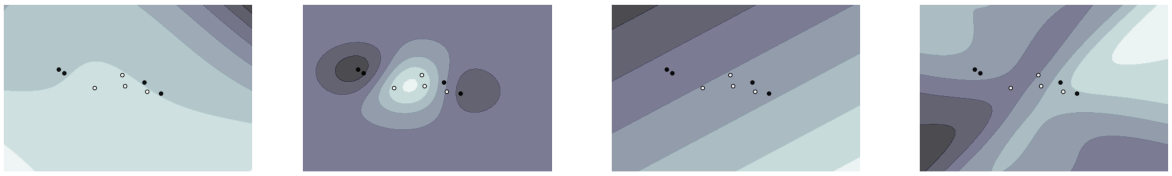
3.1 a

Consider the following dataset:

Item	1	2	3	4	5	6	7	8
$\boldsymbol{x}$	(-2,1.5)	(-1.8,1.3)	(-.7,.5)	(.3,1.2)	(.4,.6)	(1.2,.3)	(1.1,.8)	(1.7,.2)
y	-1	-1	1	1	1	1	-1	-1

In Figure 1, match the subfigures to the following kernels:

- radial basis function (RBF) kernel: $k(\mathbf{u}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{u}-\mathbf{v}\|_2^2}{\gamma}\right)$.
- linear kernel: $k(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v}$.
- polynomial kernel: $k(\mathbf{u}, \mathbf{v}) = \left(\mathbf{u}^\top \mathbf{v} + c\right)^d$ with $d > 0$.
- sigmoid kernel: $k(\mathbf{u}, \mathbf{v}) = \tanh\left(\gamma \mathbf{u}^\top \mathbf{v} + \theta\right)$



(a) *(b)* *(c)* *(d)*

Figure 1: Contour plots for four kernels.

3.2 b

For each of the data sets below, circle **ALL** kernels that could possibly be used to perfectly classify the data with an SVM. If no kernel can be used to perfectly classify the data, instead circle **NONE OF THESE**. The three types of kernels to consider are:

- **QUADRATIC**: A polynomial kernel of degree 2
- **CUBIC**: A polynomial kernel of degree 3
- **RBF**: A radial basis function kernel

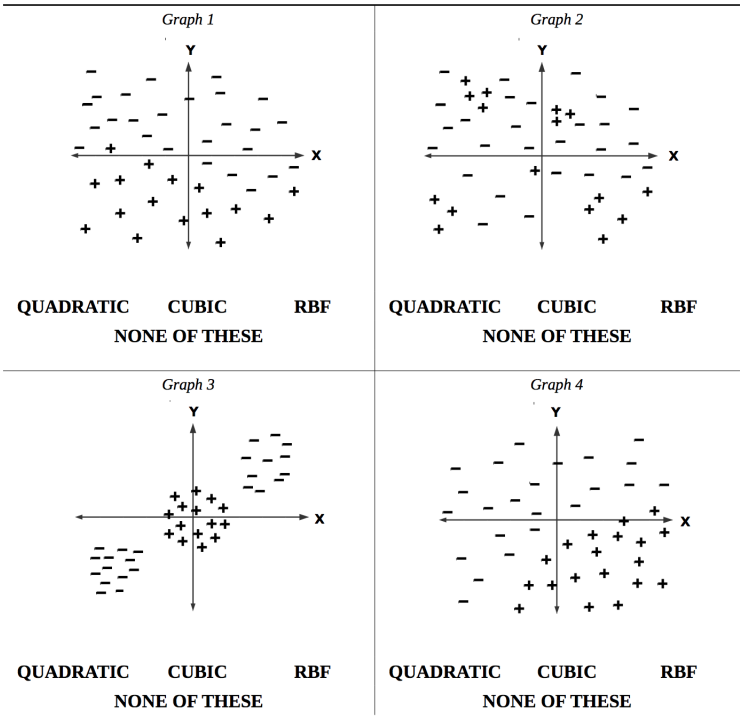


Figure 2

4 SVM with a non-linear kernel

Given data in Figure 3 and the kernel function $k(\mathbf{u}, \mathbf{v}) = 2\,\|\mathbf{u}\|\,\|\mathbf{v}\|$, compute the decision boundary for the SVM with kernel k.

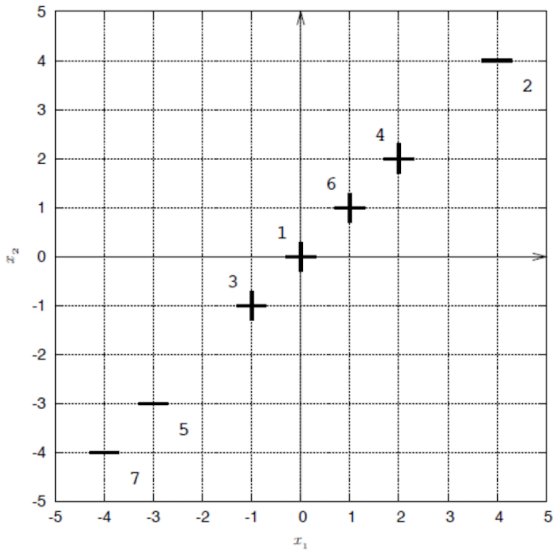


Figure 3

Step 1: Compute the feature map $\boldsymbol{\phi}$ associated with the kernel function k.

Step 2: Compute the positions for the data points in the new feature space.

Step 3: Compute the decision boundary in the new feature space and thus the decision boundary in the original space.