Solution to Homework 5: Kernels

Dr. Carlos Oliver
carlos.oliver@bsse.ethz.ch

Dr. Sarah Brüningk
sarah.brueningk@bsse.ethz.ch

Thomas Gumbsch
thomas.gumbsch@bsse.ethz.ch

Bowen Fan
bowen.fan@bsse.ethz.ch

Prof. Dr. Karsten Borgwardt
karsten.borgwardt@bsse.ethz.ch

Submission deadline: 14.12.2022, 08:00

Objectives

The goal of this homework is to understand the principle of kernel functions, short *kernels*.

Problem Overview

In this homework you will investigate kernels. In Part 1 you will investigate the dot product in feature space. In Part 2 you will investigate the property of positive definiteness of kernels. In Part 3 you will investigate constructing kernels with the closure properties.
Always document all steps of your calculations unless stated otherwise.

Homework Part 1: Dot Product in Feature Space

Introduction

Kernels are very important tools in data mining. They transform a linear classifier into a non-linear one, they avoid the representation of data in the high-dimensional feature space and they extend the support vector machine (SVM) framework to structured data such as strings or graphs.

Given two vectors $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, the linear kernel is defined as the dot product between vectors:

$$k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle = \sum_i^d x_i x_i'$$

However, this kernel only applies to vectorial data. To handle structured data (e.g. strings), we therefore need to map the data to a feature space $\mathcal{H}$ (e.g. $\mathbb{R}^d$):

$$\mathbf{x} \rightarrow \phi(\mathbf{x}) \in \mathcal{H}$$

then we can calculate the dot product in the feature space:

$$k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$$

Even for objects that are represented by vectors, we can use a non-linear mapping ϕ to handle non-linear cases.

Exercise 1

The polynomial kernel is given by

$$k(\mathbf{x}, \mathbf{x}') = (\langle \mathbf{x}, \mathbf{x}' \rangle + c)^p$$

Answer the following theoretical questions:

Exercise 1.a Given $\mathbf{x} = (x_1, x_2)$ and $\mathbf{x}' = (x_1', x_2')$ in the input space, what is $\phi(\mathbf{x})$ in the feature space of the polynomial kernel with $p = 2$ and $c = 1$? What is the dimensionality of this feature space? More generally, how does the dimension of the feature space scale in p for $c = 0$?

By expanding the inner product and applying the multinomial theorem, we can factor the kernel value into the inner product of two feature vectors $\phi(\mathbf{x})$ as:
$\phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, 1)$
The dimensionality of the feature space is 6.
The Gaussian RBF kernel admits an infinite-dimensional feature space.

Exercise 1.b The dimension of the feature space can be very high. Do we need to represent the feature space explicitly for non-linear kernels when using an SVM classifier? Give a reason for your answer.

There is no need to represent feature space explicitly. As shown in the lecture slides, the SVM and its decision function are only related to the inner product between the training points and the predicting points. Therefore, we only need to specify the kernel function without representing the feature space.

Homework Part 2: Positive Semi-Definite Kernels

Introduction

A kernel is a similarity measure, but not all the similarity measures are kernels. How to define a valid kernel? Suppose a similarity function k that gives a real number between two objects $\mathbf{x}$ and $\mathbf{x}'$ (e.g. $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$), that is:

$$k(\mathbf{x}, \mathbf{x}') \in \mathbb{R}.$$

Given a set of objects $\{\mathbf{x_1}, ..., \mathbf{x_n}\}$ and the function k, the Gram Matrix (or kernel matrix) of k is defined as an $n \times n$ matrix K with respect to $\{\mathbf{x_1}, ..., \mathbf{x_n}\}$:

$$K_{ij} = k(\mathbf{x_i}, \mathbf{x_j})$$

A matrix K is positive semi-definite, if for all $c_i \in \mathbb{R}$ K satisfies the following inequality:

$$\sum_{i,j} c_i c_j K_{ij} \geq 0$$

Let $\mathbb{X}$ be a nonempty set. If for all $n \in \mathbb{N}$ and all $\{\mathbf{x_1}, ..., \mathbf{x_n}\} \in \mathbb{X}$, the Gram matrices Ks of the function k are all positive semi-definite, then k is a semi-definite kernel function or *kernel*, for short. The definitions shown above are from [2].

Exercise 2

When answering the following theoretical questions about the positive semi-definiteness of kernels, you may find the scalar multiplication property of the dot product very useful:

$$c_1 c_2 \langle \mathbf{x}, \mathbf{x}' \rangle = \langle c_1 \mathbf{x}, c_2 \mathbf{x}' \rangle, \text{ for } c_1, c_2 \in \mathbb{R}$$

Exercise 2.a Let $\mathbb{X} \subset \mathbb{R}^d$, prove that the linear kernel is a kernel (show for all $n \in \mathbb{N}$ and all $\{\mathbf{x_1}, ..., \mathbf{x_n}\} \in \mathbb{R}^d$ that the Gram matrix is positive semi-definite).

$$k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle = \sum_i^d x_i \, x_i'$$

For all $n \in \mathbb{N}$, all $\{\mathbf{x_1}, ..., \mathbf{x_n}\} \in \mathbb{R}^N$ and $c \in \mathbb{R}^N$, we have:

$$\sum_{i,j} c_i c_j K_{ij} = \sum_{i,j} c_i c_j \langle \mathbf{x_i}, \mathbf{x_j} \rangle = \sum_{i,j} \langle c_i \mathbf{x_i}, c_j \mathbf{x_j} \rangle = \langle \sum_i c_i \mathbf{x_i}, \sum_j c_j \mathbf{x_j} \rangle = || \sum_i c_i \mathbf{x_i} ||^2 \geq 0$$

Thus, the gram matrix of linear kernel is positive semi-definite, and the linear kernel is a positive semi-definite kernel.

Exercise 2.b Similarly, prove that the dot product in any feature space is a kernel.

$$k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}) \rangle$$

For all $n \in \mathbb{N}$, all $\{\mathbf{x_1}, ..., \mathbf{x_n}\} \in \mathbb{R}^N$ and $c \in \mathbb{R}^N$, we have:
Similarly we have:

$$\sum_{i,j} c_i c_j K_{ij} = \sum_{i,j} c_i c_j \langle \phi(\mathbf{x_i}), \phi(\mathbf{x_j}) \rangle = \langle \sum_i c_i \phi(\mathbf{x_i}), \sum_j c_j \phi(\mathbf{x_j}) \rangle = || \sum_i c_i \phi(\mathbf{x_i}) ||^2 \geq 0$$

Thus, the dot product is a kernel in any feature space.

Exercise 2.c Assume we are given two kernels k_1 and k_2, prove that the following functions are kernels.

$$k_3(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')$$
$$k_4(\mathbf{x}, \mathbf{x}') = \lambda k_1(\mathbf{x}, \mathbf{x}'), \lambda \in \mathbb{R}^+$$

For all $n \in \mathbb{N}$, all $\{\mathbf{x_1}, ..., \mathbf{x_n}\} \in \mathbb{R}^N$, and $c \in \mathbb{R}^N$ we have:

$$\sum_{ij} c_i c_j k_1(\mathbf{x_i}, \mathbf{x_j}) \geq 0 \text{ and } \sum_{ij} c_i c_j k_2(\mathbf{x_i}, \mathbf{x_j}) \geq 0$$

Thus,

$$\sum_{ij} c_i c_j k_3(\mathbf{x_i}, \mathbf{x_j}) = \sum_{ij} c_i c_j (k_1(\mathbf{x_i}, \mathbf{x_j}) + k_2(\mathbf{x_i}, \mathbf{x_j})) \geq 0$$

$$\sum_{ij} c_i c_j k_4(\mathbf{x_i}, \mathbf{x_j}) = \sum_{ij} c_i c_j \lambda k_1(\mathbf{x_i}, \mathbf{x_j}) \geq 0 \text{ for } \lambda \in \mathbb{R}^+$$

Therefore, functions k_3 and k_4 are kernels.

Homework Part 3: Constructing Kernels

In Exercise 2.c you proved some of the closure properties of kernels. You can construct new kernels by combining known kernels, employing operations with respect to which the set of kernels is closed.

Exercise 3

Use the closure properties of kernels from Exercise 2.c and the lecture slides to construct new kernels:

Exercise 3.a Show that

$$k(\mathbf{x}, \mathbf{x}') = 6\langle \mathbf{x}, \mathbf{x}' \rangle^4 + 3 + \mathbf{x}^T \mathbf{x}' + \exp\left(-\frac{1}{2\sigma^2} ||\mathbf{x} - \mathbf{x}'||^2\right)$$

is a kernel.
$k(\mathbf{x}, \mathbf{x}')$ is a combination of a polynomial kernel $(\langle \mathbf{x}, \mathbf{x}' \rangle + 0)^4$, a constant kernel $k(\mathbf{x}, \mathbf{x}') = 3$, a linear kernel $\mathbf{x}^T \mathbf{x}'$, and the RBF kernel. The combination follows the closure properties that a sum of two kernels is a kernel and that multiplying a kernel with a positive scalar $\lambda \in \mathbb{R}$ results in a kernel. Therefore, $k(\mathbf{x}, \mathbf{x}')$ is also a kernel.

Exercise 3.b Type I collagen is the most abundant protein in the human body [1]. It consists of two *alpha-1(I)* and one *alpha-2(I)* chains that form a triple helix. The precursors of these chains are produced by the genes COL1A1 and COL1A2, respectively. The triple helical domain of both genes are rich in glycine-XY (GXY) motifs where X and Y are non-glycine amino acids

Let the GXY similarity measure, $k_{\text{GXY}}(X, X')$, between two amino acid sequences X and X' be defined as the sum of the similarities of their 3-mers. More precisely, we define the similarity of two 3-mers to be

- 0 if either one or both 3-mers do not start with a glycine (G) or
- the number of perfect matches, otherwise. (e.g. $k_{\text{GXY}}(GDA, GTR) = 1$)

Based on the R-convolution kernel framework defined on slide 156 of the lecture slides, we can express this kernel as follows:

$$k_{\text{GXY}}(X, X') = \sum_{s \in S, s' \in S'} k_{\text{base}}(s, s')$$

Give a formal mathematical description of a kernel $k_{\text{base}}(s, s')$ for $k_{\text{GXY}}(X, X')$. What are the substructures S and S' for this kernel?

$$k_{\text{base}}(s, s') = \begin{cases} 0 & \text{, if } (s_0 \neq \text{G} \vee s'_0 \neq \text{G}) \vee (\exists i \in \{1, 2\} : s_i = G \vee s'_i = G) \\ \sum_{i=0}^{2} k_{\text{delta}}(s_i, s'_i) & \text{, otherwise} \end{cases}$$

NOTE: this definition accounts for the constraint that XY must NOT be G. This was not made explicit in the bullet points above so we accept solutions that do not include this requirement.

The substructures S and S' are the set of all 3-mers of X and X', respectively:

$$S = \{(X_i, X_{i+1}, X_{i+2}) | i \in \{0, \ldots, |X| - 3\}\}$$
$$S' = \{(X'_i, X'_{i+1}, X'_{i+2}) | i \in \{0, \ldots, |X'| - 3\}\}$$

Exercise 3.c In the following you see partial amino acid sequences from three different proteins. Let the sequences X_1, X_2, and X_3 be the shown amino acid sequences of the proteins encoded by COL1A1, COL1A2, and GPR 143, respectively.

```
COL1A1 : GPAGFAGPPGDA
COL1A2 : PRGDQGPVGRTG
GPR 143: GFPNFDVSVSDM
```

Calculate $k_{\text{GXY}}(X_1, X_2)$ and $k_{\text{GXY}}(X_1, X_3)$. If you use code, add the code to your submission, otherwise document all steps of your calculations.

We decompose each sequence into its substructures then evaluate the base kernel on all pairs of substructures.
X_1: GPA PAG AGF GFA FAG AGP GPP PPG PGD GDA
X_2: PRG RGD GDQ DQG QGP GPV PVG VGR GRT RTG
X_3: GFP FPN PNF NFD FDV DVS VSV SVS VSD SDM

$$
\begin{aligned}
k_{\text{GXY}}(X_1, X_2) &= k_{\text{base}}(GPA, GDQ) + k_{\text{base}}(GPA, GPV) + k_{\text{base}}(GPA, GRT) \\
&\quad + k_{\text{base}}(GFA, GDQ) + k_{\text{base}}(GFA, GPV) + k_{\text{base}}(GFA, GRT) \\
&\quad + k_{\text{base}}(GPP, GDQ) + k_{\text{base}}(GPP, GPV) + k_{\text{base}}(GPP, GRT) \\
&\quad + k_{\text{base}}(GDA, GDQ) + k_{\text{base}}(GDA, GPV) + k_{\text{base}}(GDA, GRT) \\
&\quad = 1+2+1+1+1+1+1+2+1+2+1+1 = 15 \\
k_{\text{GXY}}(X_1, X_3) &= k_{\text{base}}(GPA, GFP) \\
&\quad + k_{\text{base}}(GFA, GFP) \\
&\quad + k_{\text{base}}(GPP, GFP) \\
&\quad + k_{\text{base}}(GDA, GFP) \\
&\quad = 1+2+2+1 = 6
\end{aligned}
$$

Exercise 3.d Two properties of the sequences in the above example might simplify our calculations:

1. All sequences are of equal length
2. The length of all sequences is a multiple of three

How could sequences of unequal length and/or lengths that are not multiples of 3 impact our results? Do you see a problem with this scenario? Give a reason for your answer.

When comparing amino acid sequences of different lengths, we will face the problem of different scales. Assume we have an amino acid sequence X_1 with $|X_1| = 12$ consisting of 4 GXY repeats and X_2 with $|X_2| = 120$ containing the same motifs but no other GXZ sequences. While our kernel will result in the maximal possible similarity, we do not account for the fact that the majority of X_2 does not show the collagen specific pattern. One way to mitigate this issue is to normalize by the total number of amino acids when calculating the kernel. Furthermore, our kernel does not take into account whether the motif occurs in a consecutive manner. A random distribution of X_1's GXY motifs in X_2 leads to the same output as if the motifs in X_2 follow the same order as in X_1.
Depending on the definition of the substructures S and S' in exercise 3.b, the "multiple of three" property might not change the outcome. If we define the 3-mer substructures S and S' in a sliding window approach with stride t and window size w as $t = w = 3$, we will not be able to cover all amino acids if the length of the given sequence is not a multiple of three.
In order to receive full points it is enough to identify one problem or to explain why these properties do not impact the calculations.

Grading

This homework is worth a total of 100 points. Table 1 shows the points assigned to each exercise.

Table 1: Grading key for Homework 5

20 pts.	**Exercise 1**	
	10 pts.	Exercise 1.a
	10 pts.	Exercise 1.b
30 pts.	**Exercise 2**	
	10 pts.	Exercise 2.a
	10 pts.	Exercise 2.b
	10 pts.	Exercise 2.c
50 pts.	**Exercise 3**	
	10 pts.	Exercise 3.a
	15 pts.	Exercise 3.b
	10 pts.	Exercise 3.c
	15 pts.	Exercise 3.d

Acknowledgements

This exercise was created by Karsten Borgwardt and Xiao He and extended by Christian Bock.

References

[1] G. A. Di Lullo, S. M. Sweeney, J. Körkkö, L. Ala-Kokko, and J. D. San Antonio. Mapping the ligand-binding sites and disease-associated mutations on the most abundant protein in the human, type i collagen. *Journal of Biological Chemistry*, 277 (6):4223–4231, 2002.

[2] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001. ISBN 0262194759.