Homework 5: Kernels

Dr. Sarah Brüningk (`sarah.brueningk@bsse.ethz.ch`)
Thomas Gumbsch (`thomas.gumbsch@bsse.ethz.ch`)
Bowen Fan (`bowen.fan@bsse.ethz.ch`)
Dr. Carlos Oliver (`carlos.oliver@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. $\mathbb{R}^d$):

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

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

Among the most popular non-linear kernels are the polynomial kernel

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

,

and the Gaussian RBF kernel:

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\frac{1}{2\sigma}||\mathbf{x} - \mathbf{x}'||^2).$$

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? Without providing a formal proof, what is the dimensionality of the Gaussian RBF kernel's feature space for $\sigma = 1$?

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.

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 [Schölkopf and Smola, 2001].

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).

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

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

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

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

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

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.

Exercise 3.b   Type I collagen is the most abundant protein in the human body [Di Lullo et al., 2002]. 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. $k_{\text{GXY}}(GDR, GTR) = 2$)

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?

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.

```
1    COL1A1 :  GPAGFAGPPGDA
2    COL1A2 :  PRGDQGPVGRTG
3    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.

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.

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

References

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.

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