Exc. 1

The polynomial kernel is given by

$$k(\boldsymbol{x}, \boldsymbol{x}') = (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + c)^p \qquad \text{where } \boldsymbol{x}, \boldsymbol{x}' \ \in \mathbb{R}^d$$

1.a)

Given $\boldsymbol{x} = (x_1, x_2)$ and $\boldsymbol{x}' = (x_1', x_2')$ in the input space, what is $\phi(\boldsymbol{x})$ in the feature space of the polynomial kernel with $p = 2$ and $c = 1$ (1)? What is the dimensionality of this feature space (2)? More generally, how does the dimension of the feature space scale in p for $c = 0$ (3)?

(1) We have

$$\begin{aligned}
k(\boldsymbol{x}, \boldsymbol{x}') &= (x_1x_1' + x_2x_2' + 1)^2 \\
&= (x_1x_1' + x_2x_2' + 1)(x_1x_1' + x_2x_2' + 1) \\
&= (x_1x_1')^2 + 2 \cdot x_1x_1'x_2x_2' + x_1x_1' + (x_2x_2')^2 + x_2x_2' + x_1x_1' + x_2x_2' + 1 \\
&= (x_1x_1')^2 + (x_2x_2')^2 + 2 \cdot x_1x_1'x_2x_2' + 2 \cdot x_1x_1' + 2 \cdot x_2x_2' + 1
\end{aligned}$$

So since $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle$ we know that

$$\begin{aligned}
\langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle &= (x_1x_1')^2 + (x_2x_2')^2 + 2 \cdot x_1x_1'x_2x_2' + 2 \cdot x_1x_1' + 2 \cdot x_2x_2' + 1 \\
&= [x_1^2 \ \ x_2^2 \ \ \sqrt{2}x_1x_2 \ \ \sqrt{2}x_1 \ \ \sqrt{2}x_2 \ \ 1] \begin{bmatrix} x_1'^2 \\ x_2'^2 \\ \sqrt{2}x_1'x_2' \\ \sqrt{2}x_1' \\ \sqrt{2}x_2' \\ 1 \end{bmatrix}
\end{aligned}$$

Therefore $\phi(\boldsymbol{x}) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ 1 \end{bmatrix}$

(2) As we can see $\phi(\boldsymbol{x}) \in \mathbb{R}^6$ so the dimensionality of the feature space is 6.

(3) If $c = 0$ then dimensionality of the feature space as a function of p is $\binom{d+p-1}{p}$ (Liu, 2007, p. 110). So then the feature space scales in p as $\mathcal{O}(d^p)$, i.e. it grows exponentially in p.

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.

No, we don't need to represent the feature space explicitly for non-linear kernels. Instead, we can rely on the kernel trick to implicitly work in the feature space via (new) inner products.

From the lecture we know that the optimization dual problem for the hard-margin SVM looks as follows:

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^n}{\text{maximize}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \quad \forall i \in \{1, ..., n\} : \alpha_i \geq 0$$

The key insight is that the features appear *only* in form of an inner product $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$. If we non-linearly map our features to a higher dimension via a transformation $\phi(\cdot)$ then this inner product gets replaced by a new inner product $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle_{new} = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$.

Now the beauty of using a (non-linear) kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is that it gives us $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle_{new}$ directly (with less computational effort), i.e. $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle_{new}$ (so we plug the kernel into the dual opt. problem, i.e. do the kernel trick) . So we don't need to use the transformation function $\phi(\cdot)$ and hence the feature space is never explicitly calculated/represented.

Exc. 2

A kernel function $k(\boldsymbol{x}, \boldsymbol{x}') \in \mathbb{R}$ is a valid kernel if the following conditions hold:

(1) $k(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}', \boldsymbol{x})$ (symmetry)

(2) $\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0 \quad \forall c_1, ..., c_n \in \mathbb{R},\ \forall \boldsymbol{x}_1, ..., \boldsymbol{x}_n \in \mathbb{R}^d,\ \forall n \in \mathbb{N}$ (Gram matrix is p.s.d.)

Exc. 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 $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} \in \mathbb{R}^d$ that Gram matrix is positive semi-definite.

We have the linear kernel as

$$k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle = \boldsymbol{x}^t \boldsymbol{x}' = \sum_{i}^{d} x_i x_i'$$

(1) symmetry is fullfilled because dot product is symmetric, i.e. $\boldsymbol{x}^t \boldsymbol{x}' = \boldsymbol{x}' \boldsymbol{x}^t$

(2) Proof p.s.d. of Gram matrix:

$$\begin{aligned}
\sum_{i,j=1}^{n} c_i c_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \sum_{i,j=1}^{n} c_i c_j \ \boldsymbol{x}_i^T \boldsymbol{x}_j \\
&= \sum_{i,j=1}^{n} c_i c_j \ \sum_{k=1}^{d} x_{ik} x_{jk} \\
&= \sum_{k=1}^{d} \left(\sum_{i=1}^{n} c_i x_{ik} \right) \left(\sum_{j=1}^{n} c_j x_{jk} \right) \\
&= \sum_{k=1}^{d} \left(\sum_{i=1}^{n} c_i x_{ik} \right) \left(\sum_{i=1}^{n} c_i x_{ik} \right) \\
&= \sum_{k=1}^{d} \left(\sum_{i=1}^{n} c_i x_{ik} \right)^2 \geq 0 \quad \forall c_1, ..., c_n, \ \forall \boldsymbol{x}_1, ..., \boldsymbol{x}_n, \ \forall n
\end{aligned}$$

Note that in the second last step we could exchange j for i because they are just dummy indices belonging to the same set $\{1, ..., n\}$. Since (1) and (2) are fullfilled, the linear kernel constitutes a valid kernel.

Exc. 2.b

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

We have

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

(1) Symmetry is fullfilled because dot product is symmetric.

(2) Proof p.s.d. of Gram matrix:

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \ k(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \ \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle \\
&= \left\langle \ \sum_{i}^{n} c_i \phi(\boldsymbol{x}_i), \sum_{j}^{n} c_j \phi(\boldsymbol{x}_j) \ \right\rangle \\
&= \left\| \sum_{i}^{n} c_i \ \phi(\boldsymbol{x}_i) \right\|^2 \geq 0 \quad \forall c_1, ..., c_n, \ \forall \boldsymbol{x}_1, ..., \boldsymbol{x}_n, \ \forall n
\end{aligned}$$

Since (1) and (2) are fullfilled, the dot product in any feature space is a kernel.

Exc. 2.c

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

$\underline{k_3(\boldsymbol{x}, \boldsymbol{x}') = k_1(\boldsymbol{x}, \boldsymbol{x}') + k_2(\boldsymbol{x}, \boldsymbol{x}') :}$

(1) $k_3(\boldsymbol{x}, \boldsymbol{x}')$ fullfills symmetry since it is a sum of two valid, i.e. symmetric kernels.

(2) proof that Gram matrix of $k_3(\boldsymbol{x}, \boldsymbol{x}')$ is p.s.d.:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \ k_3(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \ (k_1(\boldsymbol{x}_i, \boldsymbol{x}_j) + k_2(\boldsymbol{x}_i, \boldsymbol{x}_j)) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \ k_1(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \ k_2(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0 \quad \forall c_1, ..., c_n, \ \forall \boldsymbol{x}_1, ..., \boldsymbol{x}_n, \ \forall n\end{aligned}$$

Since the left summand is ≥ 0 because k_1 is a valid kernel, and the same holds true for the right summand because k_2 is a valid kernel, then their sum must be ≥ 0. Therefore, the Gram matrix of k_3 is p.s.d., and since k_3 is also symmetric it constitutes valid kernel.

$\underline{k_4(\boldsymbol{x}, \boldsymbol{x}') = \lambda k_1(\boldsymbol{x}, \boldsymbol{x}'), \ \lambda \in \mathbb{R}^+ :}$

(1) k_4 obviously fullfills symmetry since k_1 does and scaling k_1 by λ doesn't change anything.

(2) Proof that Gram matrix of k_4 is p.s.d.:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \ k_4(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \ \lambda k_1(\boldsymbol{x}, \boldsymbol{x}') \\ &= \lambda \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \ k_1(\boldsymbol{x}, \boldsymbol{x}') \geq 0 \quad \forall c_1, ..., c_n, \ \forall \boldsymbol{x}_1, ..., \boldsymbol{x}_n, \ \forall n, \ \forall \lambda\end{aligned}$$

Since $\lambda > 0$ and $\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \ k_1(\boldsymbol{x}, \boldsymbol{x}') \geq 0$ (because k_1 is a valid kernel), the above inequality holds and thus, the Gram matrix of k_4 is p.s.d.. Further, k_4 fullfills symmetry as well and therefore, it constitutes a valid kernel.

Exc. 3

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

Exc. 3.a

Show that

$$k(\boldsymbol{x}, \boldsymbol{x}') = 3\langle \boldsymbol{x}, \boldsymbol{x}' \rangle^4 + 1 + \boldsymbol{x}^t \boldsymbol{x}' + \exp(-\frac{1}{2\sigma^2}||\boldsymbol{x} - \boldsymbol{x}'||^2)$$

is a kernel.

Define

1) $k_1(\boldsymbol{x}, \boldsymbol{x}') = 3\langle \boldsymbol{x}, \boldsymbol{x}' \rangle^4$

2) $k_2(\boldsymbol{x}, \boldsymbol{x}') = 1$

3) $k_3(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^t \boldsymbol{x}'$

4) $k_4(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\frac{1}{2\sigma^2}||\boldsymbol{x} - \boldsymbol{x}'||^2)$

1) From lecture slides and Exc. 2 we know that linear kernel $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle$ is a kernel. From lecture we also know that the product of kernels is a kernel, i.e. $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle \cdot \langle \boldsymbol{x}, \boldsymbol{x}' \rangle \cdot \langle \boldsymbol{x}, \boldsymbol{x}' \rangle \cdot \langle \boldsymbol{x}, \boldsymbol{x}' \rangle = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle^4$ is a kernel. Further, from the lecture we know that multiplying a kernel by a positive scalar yields a kernel, i.e. $3\langle \boldsymbol{x}, \boldsymbol{x}' \rangle^4 = k_1(\boldsymbol{x}, \boldsymbol{x}')$ is a kernel.

2) We know from lecture slides that $k_2(\boldsymbol{x}, \boldsymbol{x}') = 1$ is a kernel.

3) We know from lecture slides and Exc. 2 that linear kernel is a kernel and hence $k_3(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^t \boldsymbol{x}' = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle$ is kernel.

4) We know from lecture that the RBF kernel, $k_4(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\frac{1}{2\sigma^2}||\boldsymbol{x} - \boldsymbol{x}'||^2)$ is a kernel.

Now from Exc.2 and lecture slides we know the that the sum of kernels yields a kernel. Hence

$$k(\boldsymbol{x}, \boldsymbol{x}') = k_1(\boldsymbol{x}, \boldsymbol{x}') + k_2(\boldsymbol{x}, \boldsymbol{x}') + k_3(\boldsymbol{x}, \boldsymbol{x}') + k_4(\boldsymbol{x}, \boldsymbol{x}') = 3\langle \boldsymbol{x}, \boldsymbol{x}' \rangle^4 + 1 + \boldsymbol{x}^t \boldsymbol{x}' + \exp(-\frac{1}{2\sigma^2}||\boldsymbol{x} - \boldsymbol{x}'||^2)$$

is a kernel.

Exc. 3.b

Define X and X' as two amino acid sequences, i.e. two sequences of letters. Then the substructure S is the set of all 3-mers of X and the substructure S' is the set of all 3-mers of X'. Further let s denote a single 3-mer of the set S and s' a single 3-mer of the set S', i.e. $s \in S$ and $s' \in S'$.

Moreover, let's define an indicator function

$$\mathbb{1}\{Y = y\} = \begin{cases} 1, & \text{if } Y = y \\ 0, & \text{otw.} \end{cases}$$

Furthermore, let s_j describe the j^{th} element of s, where $j \in \{1, 2, 3\}$. E.g. s_1 would be the first letter of the 3-mer s.

Now we can mathematically define the similarity between two 3-mers, i.e. between s and s' as:

$$k_{base}(s, s') = \begin{cases} 0, & \text{if } s_1 \neq G \quad \text{and/or} \quad s'_1 \neq G \\ \sum_{j=1}^{3} \mathbb{1}\{s_j = s'_j\}, & \text{otw.} \end{cases}$$

And we can use that to get GXY kernel:

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

Exc. 3.c

So let's define:

- $X_1 = GPAGFAGPPGAD \implies S^1 = \{GPA, GFA, GPP, GAD\}$, where a single element of the set is denoted s^1
- $X_2 = PRGDQGPVGRTG \implies S^2 = \{GDQ, GPV, GRT\}$, where a single element of the set is denoted s^2
- $X_3 = GFPNFVDSVSDM \implies S^3 = \{GFP\}$, where a single element of the set is denoted s^3

(Note: S^j is the set of 3-mers of X_j that start with a G. We can do that because if a 3-mer does not start with G then $k_{base}(. \,,.)$ will be 0 anyway and thus, doesn't impact the final result.)

Since we have to compare all pairs of 3-mers it's best to build a table and get the entries through $k_{base}(. \,,.)$ and then sum all elements of that table to get $k_{GXY}(. \,,.)$.

For $k_{GXY}(X_1, X_2)$:

		S^1			
	$k_{base}(s^1, s^2)$	GPA	GFA	GPP	GAD
S^2	GDQ	1	1	1	1
	GPV	2	1	2	1
	GRT	1	1	1	1

Summing over the above table yiels: $k_{GXY}(X_1, X_2) = 14$

For $k_{GXY}(X_1, X_3)$:

		S^1			
	$k_{base}(s^1, s^2)$	GPA	GFA	GPP	GAD
S^3	GFP	1	2	2	1

Summing over the above table yiels: $k_{GXY}(X_1, X_3) = 6$

Exc. 3.d

Two properties of the sequences in the above example might simplify our calculations.

1. **All sequences have 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 you answer.

I would argue that 1. and 2. do **not** matter in terms of results. Because what matters is the number 3-mers in two sequences that start with a G, and how similar those 3-mers (starting with G) are when compared pairwise across sequences. Remember, 3-mers that dont start with G return zero when using $k_{base}(\cdot, \cdot)$ on them - so they have no impact on result. Let's look at some examples to illustrate that point (same notation as in prev. Exc.):

Example 1 (diff. length and not multiple of 3):

Let $X_1 = GHBC$ and $X_2 = WKGHC$. Then $S^1 = \{GHB\}$ and $S^2 = \{GHC\}$
and thus, $k_{GXY}(X_1, X_2) = 2$.

Example 2 (same. length and multiple of 3):

Let $X_1 = GHB$ and $X_2 = GHC$. Then $S^1 = \{GHB\}$ and $S^2 = \{GHC\}$
and thus, $k_{GXY}(X_1, X_2) = 2$.

Example 3 (diff. length and multiple of 3):

Let $X_1 = GHB$ and $X_2 = CWKGHC$. Then $S^1 = \{GHB\}$ and $S^2 = \{GHC\}$
and thus, $k_{GXY}(X_1, X_2) = 2$.

Example 4 (same length and not multiple of 3):

Let $X_1 = GHBC$ and $X_2 = GHCK$. Then $S^1 = \{GHB\}$ and $S^2 = \{GHC\}$
and thus, $k_{GXY}(X_1, X_2) = 2$.

So all examples have the same result even though they are completely different in terms of same/not same length or multiple/not multiple of 3. The reason why they all have same result is because what matters is the number of 3-mers that start with G in the two sequences and how similar they are when compared across sequences. That is one 3-mer for X_1 (GHB) and one 3-mer for X_2 (GHC). Then applying the kernel on them yields $k_{GXY}(X_1, X_2) = 2$ in all cases.

References

Liu, B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data*. Berlin Heidelberg: Springer Science & Business Media.