Homework 5

Timon Keller

6.12.2021

Exercise 1

(a)

$$K(x,y) = (\langle x,y\rangle + 1)^2 = (\sum_{i=1}^{2} x_i y_i + 1)^2 = (x_1y_1 + x_2y_2 + 1)^2$$
$$= (x_1y_1 + x_2y_2 + 1)(x_1y_1 + x_2y_2 + 1)$$

$$= ((x_1y_1)^2 + x_1y_1x_2y_2 + x_1y_1) + (x_2y_2x_1y_1 + (x_2y_2)^2 + x_2y_2) + (x_1y_1 + x_2y_2 + 1)$$

$$= (x_1y_1)^2 + (x_2y_2)^2 + 2(x_1y_1x_2y_2) + 2(x_1y_1) + 2(x_2y_2) + 1$$

$$\varphi(x) = \underline{\langle x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{x_1}, \sqrt{x_2}, 1}\rangle$$

Dimensionality = 6

For c = 0, the dimensionality scales p + 1

$$\text{for p = 2: } \langle x, x'\rangle^2 = (\sum_{i=1}^{2} x_i x_i')'2 = (x_1x_1' + x_2x_2')^2$$
$$= (x_1x_1')^2 + 2x_1x_1'x_2x_2' + (x_2x_2')^2$$

-> Dimesionality = 3

(b)

No, since for SVM, we can use the kernel trick, meaning we do not have to represent the feature space specifically for non-linear kernels, since the kernel chosen can be any arbitrary kernel.

Exercise 2

(a)

$$K_{ij} = k(x_ix_j) = \sum_{i=1}^{d} x_{i,k}x_{j,k}$$

$$\sum_{i,j} c_ic_jK_{i,j} = \sum_{i,j} c_ic_jk(x_ix_j) = \sum_{i,j} c_ic_j\langle x_i, x_j >\rangle = \sum_{i,j} \langle c_ix_i, c_jx_j >\rangle$$

$$= \sum_{i,j} \sum_{k}^{d} c_i x_{i,k} c_j x_{j,k} = \sum_{i} \sum_{j} \sum_{k}^{d} c_i x_{i,k} c_j x_{j,k}$$

$$= \sum_{k}^{d} (\sum_{i} c_i x_{i,k})(\sum_{j} c_j x_{j,k}) = \sum_{k}^{2} (\sum_{i} c_i x_{i,k})^2 \geq 0$$

(b)

$$K_{ij} = k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

$$\sum_{i,j} c_i c_j K_{ij} = \sum_{i,j} c_i c_j \langle \phi(x_i), \phi(x_j) \rangle$$

$$= \sum_{i,j} \langle c_i \phi(x_i), c_j \phi(x_j) \rangle = \sum_{i} \sum_{j} \langle c_i \phi(x_i), c_j \phi(x_j) \rangle = \sum_{i} \langle c_i \phi(x_i), c_j \phi(x_j) \rangle^2 \geq 0$$

(c) (1)

$$k_3(x, x') = k_1(x, x') + k_2(x, x')$$
$$\sum_{i,j} c_i c_j k_3(x_i, x_j) = \sum_{i,j} c_i c_j k_1(x_i, x_j) + \sum_{i,j} c_i c_j k_2(x_i, x_j) \geq 0$$

Both k1 and k2 are kernels and are therefore both positive semi-definite. Their sum is therefore also a kernel.

(2)

$$k_4(x, x') = \lambda k_1(x, x'), \lambda \in R^+$$

$$\sum_{i,j} c_i c_j k_4(x_i, x_j) = \sum_{i,j} \lambda c_i c_j k_1(x_i, x_j) = \lambda \sum_{i,j} c_i c_j k_1(x_i, x_j) \geq 0$$

Since both λ and $K_1 \geq 0$

Exercise 3

(a)

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

$$= \lambda k_a(x, x') + k_b(x, x') + k_c(x, x')$$

$->$ also a kernel due to the closure properties proven in 2.c)

$$k_a = \text{polynomial kernel}$$

$$k_b = \text{linear kernel}$$

$$k_c = \text{RBF kernel}$$

(b)

$$k_G XY(X, X') = \begin{cases} 0, & \text{if one or both 3-mers don't start with G} \\ \text{number of perfect matches}, & \text{else} \end{cases}$$

$$= \sum_{s \in S, s' \in S'} k_{delta}(s, s')$$

$$\text{with } k_{delta} = \begin{cases} 1, & x = x' \\ 0, & \text{else} \end{cases}$$

S and S' being all possible 3-mers of X and X'

(c)

See Python program

(d)

If two sequences vary greatly in their length (for example seq1 = 3bp, seq = 99bp), our values of kGXY could potentially be extremely high, even though the sequences are obviously not similar in their length.

I don't see a problem with the lengths of the sequences not being multiples of 3, as this just increases the amount of 3-mers we get. If our sequences are of equal length, they still have the same amount of 3-mers.