Course: Machine Learning by Dr. Seyyed Salehi

Homework: HW2

Name: Mohammad Mohammadi

Student ID: 402208592

Question 1 - Valid Kernel

According to the following information, determine whether each of the parts can be a kernel or not. If it is not a kernel, give a violation example and if it is a kernel, prove it.

- $a \in \mathbb{R}^+$
- $f : \mathbb{R}^d \to \mathbb{R}^p$
- K_1, K_2 are kernels over $\mathbb{R}^d \times \mathbb{R}^d$
- K_3 is a kernel over $\mathbb{R}^p \times \mathbb{R}^p$

(a) $K(x, z) = K_1(x, z) + K_2(x, z)$

(b) $K(x, z) = aK_1(x, z)$

(c) $K(x, z) = K_3(f(x), f(z))$

Answer 1

Part a

If K_1 and K_2 are kernels, they are symmetric and positive semi-definite. Therefore, for any set of points $x_i i = 1^n$ and any set of real coefficients $c_i i = 1^n$, we have:

$\sum_{i,j} c_i c_j K_1(x_i, x_j) \geq 0 \quad \text{and} \quad \sum_{i,j} c_i c_j K_2(x_i, x_j) \geq 0.$

Adding these two inequalities, we get:

$\sum_{i,j} c_i c_j [K_1(x_i, x_j) + K_2(x_i, x_j)] \geq 0,$

which shows that $K(x, z) = K_1(x, z) + K_2(x, z)$ is positive semi-definite, and thus a kernel.

Part b

Under question's assumption of $a \in \mathbb{R}^+$:

Since K_1 is a kernel, it is positive semi-definite. For a positive scalar $a \in \mathbb{R}^+$ and any set of points $x_i i = 1^n$ along with coefficients $c_i i = 1^n$, we have:

$\sum_{i,j} c_i c_j a K_1(x_i, x_j) = a \sum_{i,j} c_i c_j K_1(x_i, x_j) \geq 0,$

since $a > 0$ and $\sum_{i,j} c_i c_j K_1(x_i, x_j) \geq 0$ by the positive semi-definiteness of K_1.

Based on the equation above $K(x, z) = aK_1(x, z)$ is positive semi-definite and symmetric, and thus a kernel.

Part c

Based on the kernel trick we can say:

Since K_3 is a kernel over $\mathbb{R}^p \times \mathbb{R}^p$ and $f : \mathbb{R}^d \to \mathbb{R}^p$ is a function mapping from the domain of our input space to the domain compatible with K_3, the composition $K_3(f(x), f(z))$ is also a kernel. For any set of points $x_i i = 1^n$ and coefficients $c_i i = 1^n$, we have:

$\sum_{i,j} c_i c_j K_3(f(x_i), f(x_j)) \geq 0,$

because K_3 is positive semi-definite by assumption.

The mapping f does not affect the positive semi-definiteness of K_3, hence,
$K(x, z) = K_3(f(x), f(z))$ is a kernel.

Question 2 - Regularization

Consider that we want to solve the binary classification problem shown in Figure 1 using a simple logistic regression model.

$$P(y = 1 \mid x, w) = g(w_0 + w_1 x_1 + w_2 x_2) = \frac{1}{1 + \exp(-w_0 - w_1 x_1 - w_2 x_2)}$$

As it is clear from the figure, the training data can be separated from each other with zero training error. Now we want to maximize the value of the following expression for large values of C to solve the classification problem.

$$\sum_{i=1}^{n} \log P(y_i | x_i; \omega_0, \omega_1, \omega_2) - C\omega_j^2$$

In this expression, $C\omega_j^2$ is the regularization term that j∈{0,1,2}. In other words, only one of the parameters is regularized. According to the training data given in Figure 1, state that in each of the following situations, how will the training error compare with the simple logistic regression model? Give reasons for your answer.

1. Regularizing ω_2.
2. Regularizing ω_1.
3. Regularizing ω_0.

Now consider that we want to regularize both parameters ω_1 and ω_2. In other words, we want to maximize the following expression in our model:

$$\sum_{i=1}^{n} \log P(y_i|x_i; \omega_0, \omega_1, \omega_2) - C(\omega_1^2, \omega_2^2)$$

training data are the same as the data in Figure 1.

4. For large values of C, what values do we expect ω_0 to take? Give reasons for your answer.
5. This time, consider that we add some "+" data that are in class y = 1 to our training data. Assuming that the two classes are still fully separable, argue what values we expect ω_0 to take.

Figure 1. "+" data belong to class y = 1, and "O" data belong to class y = 0.

Answer

Part 1

The feature x_2 separates the data very well, as shown in the figure. Regularizing ω_2 means that we are penalizing the logistic regression model for giving too much weight to x_2. If C is very large, the regularization term $C\omega_2^2$ will dominate, and the model will try to reduce ω_2 to minimize the penalty. However, since x_2 is a good separator, minimizing ω_2 would likely increase the training error compared to a model without regularization because the model would not leverage the full separating power of x_2.

Part 2

Regularizing ω_1 should not significantly affect the separation because x_1 does not separate the classes well; the data for both classes are mixed along the x_1 axis. The impact on training error would likely be less for regularizing ω_1 compared to regularizing ω_2.

Part 3

Regularizing the bias term ω_0 usually makes less sense in the context of logistic regression, as it shifts the decision boundary away from the origin but does not control the shape of the boundary. Regularizing ω_0 would affect the position of the decision boundary, possibly increasing the training error if the boundary needs to be away from the origin to separate the classes effectively (Partialy similar to our case).

However, since the data are linearly separable, it's possible to find a boundary that separates the classes with zero training error without relying on the bias term to be large.

Part 4

With both ω_1 and ω_2 being regularized, ω_0 would be the only term not penalized for being large. Hence, the optimization would likely result in ω_0 being larger relative to the other weights to achieve the best separation without incurring a penalty. Since ω_0 adjusts the threshold level for classification, it will make up for smaller values of ω_1 and ω_2 due to the high penalty on these parameters.

Part 5

If additional "+" data from class $y = 1$ are added and the classes are still fully separable, ω_0 should be adjusted to account for the new data placement.

Depending on where the new "+" data are placed, ω_0 could either increase or decrease to maintain zero training error.

If the "+" data are placed farther away from the origin along the direction where x_2 is already providing good separation, ω_0 would not need to change significantly.

If they are placed closer to the decision boundary, ω_0 might need to increase to push the decision boundary further from the origin to maintain separation.

Question 3 - Gaussian Kernel and Valid Kernel

(a) Show that Gaussian Kernel can be written as inner product of feature vectors with infinite dimenssions:

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

(b) Assume that $A \in \mathbb{R}^{p\times p}$ is a symmetric and positive semi-difinite matrix. Prove that below is a valid kernel function:

$$k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\mathrm{T}} A\mathbf{y}$$

Answer

Part a

The Gaussian Kernel by definition is defined as:

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

This kernel function can indeed be represented as an infinite-dimensional inner product, using Mercer's theorem.

The idea is based on the fact that the Gaussian kernel is a positive semi-definite function and can be expanded into a Taylor series.

The expansion of the exponential function in the Gaussian kernel corresponds to an inner product in an infinite-dimensional space.

The Taylor expansion of the exponential function is as:

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}$$

So the exponential in the Gaussian kernel can be expanded as:

$$e^{-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}} = e^{-\frac{\mathbf{x}^2}{2\sigma^2}} \cdot e^{-\frac{\mathbf{x}'^2}{2\sigma^2}} \cdot e^{\frac{\mathbf{x}\cdot\mathbf{x}'}{\sigma^2}}$$

Now we can expand $e^{\frac{\mathbf{x}\cdot\mathbf{x}'}{\sigma^2}}$ using the Taylor series expansion and consider the inner product $\mathbf{x} \cdot \mathbf{x}'$ which we can write as a sum over the individual dimensions:

$$e^{\frac{\mathbf{x}\cdot\mathbf{x}'}{\sigma^2}} = \sum_{n=0}^{\infty} \frac{(\mathbf{x} \cdot \mathbf{x}'/\sigma^2)^n}{n!}$$

Each term of this series can be thought of as an inner product of some feature mapping $\phi(x)$ in a higher (possibly infinite) dimensional space.

Part b

In order to prove that a function $k(x, y) = x^T A y$ is a valid kernel function, we need to show that it satisfies Mercer's condition, which states that the kernel must be **symmetric** and **positive semi-definite**.

For Symmetry:

Since A is symmetric and positive semi-definite, $A^T = A$ due to its symmetry, so we have:

$$k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^{\mathrm{T}} A \mathbf{y} = \mathbf{y}^{\mathrm{T}} A \mathbf{x} = k(\mathbf{y}, \mathbf{x})$$

Which states that the kernel is symmetric.

For Positive Semi-Definiteness:

For any set of points $\{x_i\}_{i=1}^n$ and any set of coefficients $\{c_i\}_{i=1}^n$, we must have:

$$\sum_{i,j} c_i c_j k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i,j} c_i c_j \mathbf{x}_i^{\mathrm{T}} A \mathbf{x}_j \geq 0$$

We can express this as a matrix operation:

$$\mathbf{c}^{\mathrm{T}} K \mathbf{c} = \mathbf{c}^{\mathrm{T}} X^{\mathrm{T}} A X \mathbf{c}$$

where X is the matrix whose columns are the vectors x_i, K is the kernel matrix with entries $K_{ij} = K(x_i, x_j)$ and $\mathbf{c}$ is the vector of coefficients c_i.

Since A is positive semi-definite, the product $X^T AX$ is also positive semi-definite, which means $\mathbf{c}^T K\mathbf{c} \geq 0$ for any $\mathbf{c}$, and thus, K is positive semi-definite.

Hence $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T A\mathbf{y}$ is a valid kernel function.