CS6350 - Homework/Assignment-5

Arnab Das(u1014840)

November 14, 2016

1: Margins

(1) For a xor function in two dimension of x_1, x_2 and label y, the examples sets in the form of tuple (x_1, x_2, y) are $(-1,-1,-1)$, $(-1,1,1)$, $(1,-1,1)$ and $(1,1,-1)$, where variables are boolean and takes $\{-1,1\}$. It is not linearly separable in the euclidean space. However, the transformation, ϕ, of mapping $[x_1, x_2]$ to $[x_1, x_1x_2]$ makes it linearly separable in which the datapoints now (x_1, x_1x_2, y) becomes $(-1,1,-1)$, $(-1,-1,1)$, $(1,-1,1)$ and $(1,1,-1)$. The line $x_1x_2 = 0$ is a separating classifier. SInce $x_1x_2 = 0$ is equaidistant from all the 4 points in the transformed space, it gives the maximum margin, which is the distance of any of the points(since equidistant) from this line. and equal to **1 unit**. The linear classifier in the transformed space when mapped back to the original euclidean space , will be combination of the lines $x_1 = 0$ and $x_2 = 0$, as shown in Figure-1(c). This is because in the transformed space, since $x_1x_2 = 0$ means that line satisfies all points which has $x_1 = 0$ or/and $x_2 = 0$, hence in the euclidean space it is a combination of both.

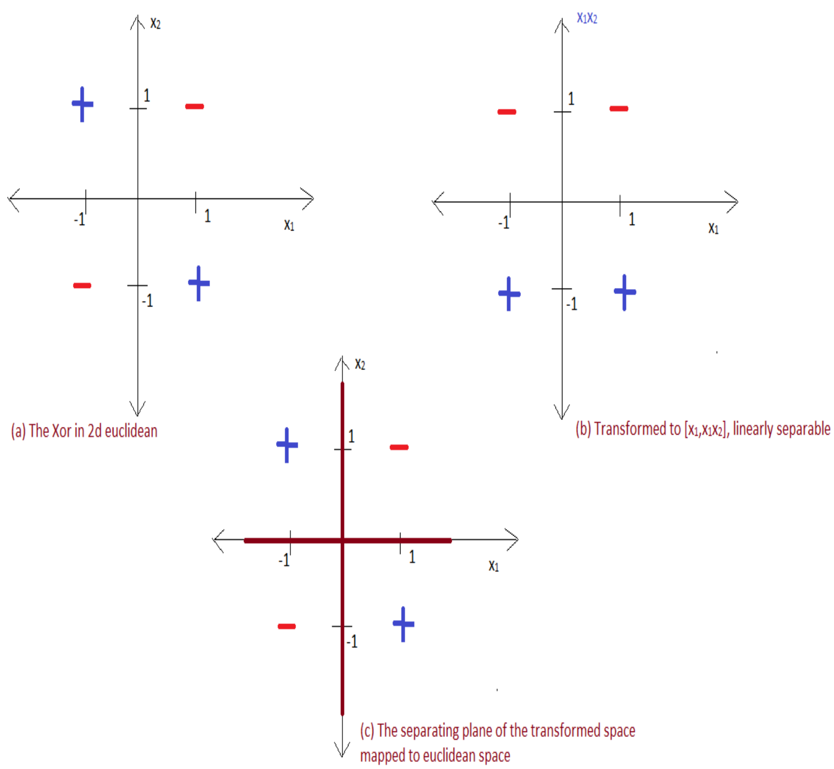

Figure 1: Space transformation for Xor to make linearly separable

(2a) For $D_1 = \{x_1, x_2, x_3, x_5, x_7\}$, the linear classifier with the maximum margin will be parallel to the line joining x_1 and x_3, and the distance of this classifer will be equal from $x_1 and x_3$ on one side and x_5 on the other side. Hence, the maximum possible margin for D_1 will be half the distance of x_5 from the line joining x_1 and x_3. The line joining x_1, x_3 is $x_1 - x_2 = 0$. Then, margin for D_1 will be:

$$D_{1_{marginMax}} = \frac{1}{2 \times \sqrt[2]{2}}$$

For $D_2 = \{x_1, x_5, x_6, x_8\}$, the linear classifier with the maximum margin will be parallel to the line joining x_5, x_6, and the distance of this classifier will be equal from x_5 and x_6 on one side and x_1 on the other side. Hence, the maximum possible margin for D_2 will be half of the distance of x_1 from the line joining x_5 and x_6. The line joining x_5, x_6 is $\sqrt[2]{3}x_1 + x_2 - \sqrt[2]{3} = 0$. Then the margin for D_2 will be:

$$D_{2_{marginMax}} = \frac{\sqrt[2]{3}}{4}$$

For $D_3 = \{x_3, x_4, x_5, x_7\}$, the linear classifier with the maximum margin will be parallel to the line joining x_4 and x_3, and the distance of this classifier will be equal from x_4 and x_3 from one side and from x_5 on the other side. Hence, the maximum possible margin for D_3 will be half of the distance of x_5 from the line joining x_3 and x_4. The line joining x_3 and x_4 is $2x_1 - x_2 - 1 = 0$. Then the margin for D_3 will be:

$$D_{3_{marginMax}} = \frac{1}{2 \times \sqrt[2]{5}}$$

(2b) For D_1, $R = \frac{3}{2}$ and $\gamma = \frac{1}{2 \times \sqrt[2]{2}}$, perceptron mistake bound for $D_1 = 18$.
For D_2, $R = 1$ and $\gamma = \frac{\sqrt[2]{3}}{4}$, perceptron mistake bound for $D_2 = \frac{16}{3}$.
For D_3, $R = \frac{3}{2}$ and $\gamma = \frac{1}{2 \times \sqrt[2]{5}}$, perceptron mistake bound for $D_3 = 45$.
D_3 has the greatest mistake bound.

(2c) A higher mistake bound indicates how well the classifier can fit the training data by making only this bounded number of mistakes. Hence, a low mistake bound will means the classifier fits the training data quickly. However, that provides no guarantees on the test data. Since the perceptron learns by making mistakes, hence a lower number of mistakes indicate that the learning performed by the perceptron has been less, and hence its predictive power intuitively reduces on the test data. To put it simply, a classifier learns less if it makes less number of mistakes because that is its only entry point towards learning and updates. Hence, the classifier with a higher mistakes bound is easier to learn and the one with a small mistake bound is difficult to learn. Thus the ranking in order of ease of ranking will be D_3, D_1, D_2 .

2: Kernels

(1a) Given valid kernels , $K_1(x, z)$ and $K_2(x, z)$, we need to show the product of these two kernels is also a kernel. For the defined space $x_1, x_2, \ldots, x_n \in S$, we define the respective Gram matrices as:

$$C = \{c_{ij}\} = K_1(x_i, x_j)$$
$$D = \{d_{ij}\} = K_2(x_i, x_j)$$

We define the newKernel $K = K_1 \times K_2$ as the product of these kernels , such that its Gram matrix looks like:

$$E = \{e_{ij}\} = \{c_{ij}\}\{d_{ij}\} = K(x_i, x_j)$$

Now, since the kernels K_1 and K_2 are spd, hence the elements are symmetric. So , we can write :

$$C = \{c_{ij}\} = K_1(x_i, x_j) = \{c_{ji}\} = K_1(x_j, x_i)$$
$$D = \{d_{ij}\} = K_2(x_i, x_j) = \{d_{ji}\} = K_2(x_j, x_i)$$

Then , the corresponding elements in the matric for the new kernel, will be:

$$\{e_{ji}\} = \{c_{ji}\}\{d_{ji}\} = \{c_{ij}\}\{d_{ij}\} = \{e_{ij}\}$$

Hence, the new kernel is also symmetric. Now, we need to prove it is symmetric positive definite. Let $u \in R^n$, we need to show $u^T E u \geq 0$. We can write:

$$u^T E u = \sum_{ij} u_i u_j e_{ij} = \sum_{ij} u_i u_j c_{ij} d_{ij} \tag{1}$$

Now, any matrix A that is non-singular, can be turned into a symmetric positive definite matrix by multiplication with its transpose, that is $A^T A$ is always symmetric positive definite. Since, the matrix A can be any matrix without any restriction other than being non-singular, this means that a given symmetric positive definite matrix can be considered to be formed as a product of matrix and its transpose. So, we break down C as a product of a general non-singular matrix A and its transpose, and D as the product of a general non-singullar matrix B and its transpose.

$$C = A^T A = \{c_{ij}\} = a_i^T a_j = \sum_k a_{ik} a_{jk}$$

$$D = B^T B = \{d_{ij}\} = b_i^T b_j = \sum_l b_{il} b_{jl}$$

Plugging these back to equation(1), we get:

$$u^T E u = \sum_{ij} u_i u_j e_{ij} = \sum_{ij} u_i u_j \sum_k a_{ik} a_{jk} \sum_l b_{il} b_{jl} = \sum_{kl} \sum_{ij} u_i u_j a_{ik} a_{jk} b_{il} b_{jl}$$

$$u^T E u = \sum_{kl} \sum_{ij} u_i u_j a_{ik} a_{jk} b_{il} b_{jl}$$

Since the i,j do not depend on each other, we can separate these as below

$$u^T E u = \sum_{kl} (\sum_i u_i a_{ik} b_{il}) (\sum_j u_j a_{jk} b_{jl})$$

The terms for j are completely independent of i, and exactly identical to i, so we can remove the j terms and place the i terms as square which is greater than equal to 0

$$u^T E u = \sum_{kl} (\sum_i u_i a_{ik} b_{il})^2 \geq 0$$

Hence, the gram matrix E is symmetric positive definite, which means K is also a kernel.(Proved).

(2a) To Prove: Polynomial over a kernel constructed using positive coefficients is also a kernel.
If we can show that the sum of kernels with positive coefficients is a kernel, then using this result and the result of previous question(2.1.a), we can conclude that Polynomial over a kernel is also a kernel.
Given $K_1(x, z)$ and $K_2(x, z)$ are kernels, we define $K(x, z) = \alpha K_1(x, z) + \beta K_2(x, z)$ and show that K is a kernel.
Suppose K_1 has its feature map, ϕ_1, such that it is defined as $K_1(x, z) = \phi_1^T(x)\phi_1(z)$. Suppose K_2 has its feature map, ϕ_2, such that it is defined as $K_1(x, z) = \phi_2^T(x)\phi_2(z)$. Then we have ,

$$K(x, z) = \alpha K_1(x, z) + \beta K_2(x, z) =< \sqrt[2]{\alpha}\phi_1(x), \sqrt[2]{\alpha}\phi_1(z) > + < \sqrt[2]{\beta}\phi_1(x), \sqrt[2]{\beta}\phi_1(z) >$$
$$K(x, z) =< [\sqrt[2]{\alpha}\phi_1(x), \sqrt[2]{\beta}\phi_2(x)], [\sqrt[2]{\alpha}\phi_1(z), \sqrt[2]{\beta}\phi_2(z)] >$$

Which means $K(x, z)$ can be expresses as an inner product . Hence, K is an kernel. Next, when we have a polynomial over a kernel constrcuted using positive coefficients, then the terms of the polynomial are product of kernels and these products are summed up to produce the final polynomial. Since, we have already proved that the product of kernels is a kernel, and sum of kernels with positive(else $\sqrt[2]{\alpha}$ will be imaginary) coefficients are kernels, hence overall it is a kernel. (Proved).

(2) Given two examples $x \in R^2$ and $z \in R^2$, **Prove** the following is a kernel.

$$K(x,z) = 15(x^T z)^2 \exp(-||x-z||^2)$$

Let $K_1(x,z) = 15(x^T z)^2$ and $K_2 = \exp(-||x-z||^2)$. We already know from the previous results that the product of two kernels is a kernel. Hence, if we can separately prove that K_1 and K_2 are kernels, then that implies K is a kernel as well.

Proof: K_1 is a kernel: Mercer's condition says, K is a valid kernel for every kernel for every finite set $x_1, x_2, \dots, x_n$, for any choice of real valued $c_1, c_2, \dots$, if $\sum_i \sum_j K(x_i, x_j) \geq 0$. Choosing c_i, c_j to be positive real values, then for any pair of examples $x_i.x_j$, for K_1, we can write: $\sum_i \sum_j c_i c_j 15(x^T z)^2$, and since their is a square term involved with positive coefficients, hence its value is greater than equal to 0. Thus , K_1 is a valid kernel.

Proof: K_2 is a kernel: . We can break down K_2 as follows:

$$K_2(x,z) = \exp(-||x-z||^2) = \exp(-(x-z)^T(x-z)) = \exp(-<x-z, x-z>) = \exp(-(<x, x-z> - <z, x-z>))$$

$$K_2(x,z) = \exp(-(<x,x> - <x,z> - <z,x> + <z,z>)) = \exp(-(||x||^2 + ||z||^2 - 2<x,z>))$$

$$K_2(x,z) = \exp(-(||x||^2 + ||z||^2)) \exp(2<x,z>) = C \exp(2<x,z>)$$

where C = $\exp(||x||^2 + ||z||^2)$ is a constant. Then expanding the exponential we get:

$$K_2(x,z) = C \sum_{n=0}^{\infty} \frac{<x,z>^n}{n!} \tag{2}$$

Thus, we see that K_2 is formed by an infinite summ over polynomial kernels, which are further derived from the product of linear kernels$x^T z$. Since sum and product of kernels results in a kernel as proved earlier, hence K_2 is a kernel.

Since, K is formed as the product of K_1 and K_2, hence K_1 is a valid kernel.(Proved).

(3) Prove that the Gaussian kernel can be written down as the inner product of an feature space with infinite dimension.

$$K(x,z) = \exp(\frac{-||x-z||^2}{2\sigma^2})$$

Continuing from Equation(2) of the above proof and using the earlier proofs as well:

Since, we can write the sum of two kernels as a new kernel, such that

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

and their corresponding transformations be ϕ , ϕ_1 and ϕ_2 respectively. This implies that ϕ is defined such that it forms vectors of the form:

$$\phi(x) = (\phi_1(x), \phi_2(x))$$

such that(similar to the proof for sum over kernels)

$$<\phi(x), \phi(z)> = <\phi_1(x), \phi_1(z)> + <\phi_2(x), \phi_2(z)>$$

In the euclidean space, thus $\phi(x)$ is the vector formed by appending the components of $\phi_2(x)$ to $\phi_1(x)$, and then:

$$<\phi(x), \phi(z)> = \sum_{i=1}^{dim(K_1)} \phi_{1,i}(x)\phi_{1,i}(z) + \sum_{j=1}^{dim(K_2)} \phi_{1,j}(x)\phi_{1,j}(z)$$

$$= \sum_{i=1}^{dim(K_2)+dim(K_2)} \phi_i(x)\phi_j(z)$$

Since K_1 and K_2 are infinite sum over kernel polynomials from equation(2), hence K can be written down as the inner product of a feature space with infinite dimension as shown above. Here, K represents the RBF, while K_1 and K_2 represents the infinite sums RBF is composed off as in equation(2).