Exercise 17

To show:

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be data in $\mathbb{R}^d \times \{-1.1\}$. Suppose that the data are linearly separable, that is, there exists a $w \in \mathbb{R}^d$ such that $y_i w^T x_i > 0$ for all $i = 1, \ldots, n$. The margin of such a vector is

$$\gamma(w) = \min_{i=1,\ldots,n} \frac{y_i w^T x_i}{||w||}.$$

Formulate a convex optimization problem whose solution is a vector w^* that classifies the data correctly (i.e., $y_i w^{*T} x_i > 0$ for all $i = 1, \ldots, n$). Show that the optimal solution w^* lies in the vector space spanned by the examples x_i for which the margin $\frac{y_i w^T x_i}{||w||}$ is minimal among all examples.

Solution:

The optimal classifier is the one that maximizes the margin. The margin is the distance from the closest point to the hyperplane among all points. Thus, w^* is the the parameter that is maximizing this distance. In the following let $\gamma = \gamma(w)$. We proceed by formulating the maximization problem for the margin:

$$\begin{aligned} \underset{w}{\text{maximize}} \quad & \gamma \\ \text{subject to} \quad & \frac{y_i w^T x_i}{||w||} \geq \gamma \; i = 1, \ldots, n. \end{aligned}$$

We choose $||w|| = \frac{1}{\gamma}$ and the maximization problem becomes:

$$\begin{aligned} \underset{w}{\text{maximize}} \quad & \frac{1}{||w||} \\ \text{subject to} \quad & y_i w^T x_i \geq 1 \; i = 1, \ldots, n. \end{aligned}$$

Notice that the last problem is equivalent to the following minimization problem, which is convex, since $||w||$ is convex. Hence we find the solution for the first part of the exercise

$$\begin{aligned} \underset{w}{\text{minmize}} \quad & ||w|| \\ \text{subject to} \quad & y_i w^T x_i \geq 1 \; i = 1, \ldots, n. \end{aligned}$$

In the second part of the exercise we want to show that the optimal choice of w lies in the vector space spanned by the x_i for which $\frac{y_i w^T x_i}{||w||}$ is minimal . Hence, it sufficies to show that w^* can be

expressed as a linear combination of those x_i. We find w^* by minizing $\frac{1}{2}||w||^2$. Since the functions are monotonically increasing this is legitimate. We compute w^* by solving the Lagrangian for this problem:

$$\mathcal{L} = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \lambda y_i w^T x_i - 1$$
$$\mathcal{L} = \frac{1}{2} w^T w - \sum_{i=1}^{n} \lambda y_i w^T x_i - 1$$

We compute the gradient of the Lagrangian with respect to w. By setting it to zero we solve for w to obtain the solution:

$$\nabla_w \mathcal{L} = \frac{1}{2} 2w - \sum_{i=1}^{n} \lambda_i y_i x_i = 0$$
$$w^* = \sum_{i=1}^{n} \lambda_i y_i x_i$$

Hence w^* is in fact a linear combination of those x_i on the margin, since the λ is zero for the other x_i. Hence w^* lies in the vector space.

Exercise 18

To show:

Suppose that data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d x \times \{-1.1\}$. are such that each $x_i \in 0, 1^d$ (i.e., the x_i have binary components) and for each $i = 1, \ldots, n$, $y_i = 1$ if and only if at least one component of x_i is 1. Show that the data are linearly separable. What is largest achievable margin (i.e., the smallest distance of all data points to the separating hyperplane) in the worst case? Repeat the exercise in the case when for each$i = 1, \ldots, n$, $y_i = 1$ if and only if the sum of the components of x_i is at least $d/2$ (assume here that d is odd).

Solution:

First, start by visualising the problem. The only point labeled as (-)1 is the zero vector and therefore the origin. Intuitively, we can conclude that the data is linearly separable in this setting, since all other data points lie to the "right" of the origin (labled as (+)1 if at least one entry is one).

For the worst case we want to find a hyper plane separating the data such that the margin we find is the minimum possible among all possible margins (hence, the worst). Thus, this separating hyplerplane is such that the closest points to the origin lie all on the margin. In this setting those points are obviously those lying on the axes and therefore have exactly one component equal to one.

To show that we propose that the lowest possible margin is exactly half the distance from the origin to the hyperplane going through the d closest points to the origin. Formally we define a hyperplane as a set of points satisfying:

$$\begin{aligned} w^T x_i &= b \\ &= \sum_{j=1}^{d} w_j x_i^{(j)} = b \end{aligned} \tag{1}$$

The distance from a hyperplane to the origin is defined as:

$$d(O, H) = \frac{b}{||w||} \tag{2}$$

Using (1), the hyperplane going exactly through the closest points is such that we have exactly d points with one entry equals 1 implying that w has exactly d elements equal to b. We want to find the distance to the orgin and therefore we compute the norm of w

$$\begin{aligned} ||w|| &= w^T w \\ &= \sqrt{b_1^2 + b_2^2 + \cdots + b_d^2} \\ &= \sqrt{db^2} \\ &= \sqrt{d}b \end{aligned} \tag{3}$$

Let H_R be the plane going through the points with only one entry equal to one. Substitute (3) into (2) to obtain

$$d(O, H_R) = \frac{b}{||w||} = \frac{b}{\sqrt{d}b} = \frac{1}{\sqrt{d}} \tag{4}$$

Using that we find the margin as half the distance between the plane and the origin.

$$\gamma = \frac{1}{2\sqrt{d}}$$

Since in this case the distance is greater than zero implies that the data are linearly separable. Comming to the second part of the exercise. In this case the points get labeled as one if the sum of components is at least $\frac{d}{2}$. Since the number of data points is odd we find:

$$y = +1 \quad \overset{\text{iff}}{\Longrightarrow} \quad \sum_{j=1}^{d} x_i^{(j)} \geq \frac{d+1}{2}$$

$$y = -1 \quad \overset{\text{iff}}{\Longrightarrow} \quad \sum_{j=1}^{d} x_i^{(j)} < \frac{d-1}{2}$$

For the points with label minus one we find w for those points with exactly one component equal to one:

$$\begin{aligned} w^T x_i &= b \\ &= \sum_{j=1}^{d} w_j x_i^{(j)} = b \\ &= \frac{2b}{(d-1)} \end{aligned}$$

In the following the procedure is quite similiar:

$$\begin{aligned} ||w|| &= w^T w \\ &= \sqrt{\left(\frac{2b}{(d-1)}\right)^2 + \cdots + \left(\frac{2b}{(d-1)}\right)^2} \\ &= \sqrt{d\left(\frac{2b}{(d-1)}\right)^2} \\ &= \sqrt{d}\frac{2b}{(d-1)} \end{aligned} \tag{5}$$

We use that norm to compute the distance to the origin again. Denote H_L as the plane going through the points with exactly one component equals to one. Recall (1) and substitute (5):

$$d(O, H_L) = \frac{b}{||w||} = \frac{b}{\left(\sqrt{d}\frac{2b}{(d-1)}\right)} = \frac{(d-1)}{2\sqrt{d}}$$

Now consider the plane going through the point having exactly all components equal to one. Let that plane be H_1. The distance to the origin is $d(0, H_1) = \sqrt{d}$. Let H_R be the hyperplane, which is paralell to the separating hyperplane to the right. We want to find the distance between H_L and H_R. Hence, we compute:

$$d(H_L, H_R) = d(0, H_1) - d(0, H_L) - d(H_R, H_1) \tag{6}$$

We found already $d(O, H_L)$. Since this is symetric we also found $d(H_1, H_R)$ and hence we can solve the last expression

$$\begin{aligned} d(H_L, H_R) &= \sqrt{d} - 2\frac{(d-1)}{2\sqrt{d}} \\ &= \frac{1}{\sqrt{d}} \end{aligned}$$

Finally, we notice that the margin is half of the distance between H_L and H_R and we find the margin for the second part of the exercise:

$$= \frac{1}{2\sqrt{d}}$$

Exercise 19

To show:

Let $\mathcal{H}$ be the Hilbert space of all sequences $s = \{s_n\}_{n=0}^{\infty}$ satisfying $\sum_{n=0}^{\infty} s_n^2 < \infty$ with inner product $\langle s, t\rangle = \sum_{n=0}^{\infty} s_n t_n$. Consider the feature map $\Phi : \mathbb{R} \to \mathcal{H}$ that assigns, to each real number x, the sequence $\Phi(x)$ whose n-th element equals

$$(\Phi(x))_n = \frac{1}{\sqrt{n!}} x^n e^{-x^2/n},\ n = 0, 1, 2, \ldots$$

Determine the kernel function $K(x, y) = \langle \Phi(x), \Phi(y)\rangle$ for $x, y \in \mathbb{R}$ (You max use the fact that $\sum_{n=0}^{\infty} x^n/n! = e^x$.) Can you generalize the kernel so that it is defined on $\mathbb{R}^d \times \mathbb{R}^d$ instead of $\mathbb{R} \times \mathbb{R}$? What is the corresponding feature map?

Solution:

Start by writing down the inner product of the features maps:

$$\begin{aligned} K(x, y) &= \langle \Phi(x), \Phi(y)\rangle \\ &= \sum_{n=0}^{\infty} \frac{1}{\sqrt{n!}} x^n e^{-\frac{x^2}{2}} \frac{1}{\sqrt{n!}} y^n e^{-\frac{y^2}{2}} \\ &= \sum_{n=0}^{\infty} \frac{1}{n!} x^n e^{-\frac{x^2}{2}} y^n e^{-\frac{y^2}{2}} \end{aligned}$$

Notice that both exponential terms don't depend on nm so we can pull it out of the sum

$$\begin{aligned} &= e^{-\frac{x^2}{2}} e^{-\frac{y^2}{2}} \sum_{n=0}^{\infty} \frac{1}{n!} x^n y^n \\ &= e^{-\frac{(x^2+y^2)}{2}} \sum_{n=0}^{\infty} \frac{1}{n!} x^n y^n \end{aligned}$$

In the next step use the fact that the series is convergence to the exponential and hence we get:

$$\begin{aligned} &= e^{-\frac{(x^2+y^2)}{2}} e^{xy} \\ &= e^{-\frac{(x-y)^2}{2}} \end{aligned}$$

Finally notice that in $\mathbb{R}$ the euclidian distance between two points is just the difference and hence we get the final result:

$$= e^{-\frac{||x-y||^2}{2}}$$

For the second part of the exercise it is kind of intuitive that the last expression also holds in $\mathbb{R}^d$, however I could not find an exact feature map feature, that secures that.

Exercise 20

To show:

Let $K_1, K_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be kernels. Prove that $K_1 + K_2$ and $K_1 K_2$ are also kernels.

Solution:

A)

We want to show that sum of two kernels is a kernel again. Since a kernel is positive semidefinite (that ist $x^T K x \geq 0$) we show that the sum of two kernel is again positiv semidefinite. In doing so Let $K_1 = K_1(x, y)$ and $K_2 = K_2(x, y)$.

$$\begin{aligned} & y^T (K_1 + K_2) y \\ & = y^T K_1 y + y^T K_2 y \geq 0' \end{aligned}$$

The sum of two at least zero valued number is obviously again at least zero

B)

In the second part of the exercise we show that the product of two valid kernels is again a valid kernel. In doing define the product of the kernels as follows:

$$K_p(x, y) = K_1(x, y) K_2(x, y)$$

Let Φ be the feature map for K_1 and Ψ the feature map for K_2. Using that we get:

$$K_p(x, y) = \big(\Phi(x)^T \Phi(y)\big)\big(\Psi(x)^T \Psi(y)\big)$$

Applying the definition of the inner product we write out the last expression explicitly:

$$\begin{aligned} & = \sum_{i=1}^{N} \Phi_i(x)\Phi_i(y) \sum_{j=1}^{M} \Psi_j(x)\Psi_i(y) \\ & = \sum_{i=1}^{N} \sum_{j=1}^{M} \big[\Phi_i(x)\Psi_j(x)\big]\big[\Phi_i(y)\Psi_j(y)\big] \\ & = \sum_{k=1}^{K} \varphi_k(x)\varphi_k(y) \\ & = \varphi(x)^T \varphi(y) \end{aligned}$$

Question 21

To show:

Let $\mathcal{X}_n = \{0,1\}^n$ be the set of binary strings of length n. Let $m < n$. We say that $s \in \{0,1\}^m$ is a substring of $x = (x_1, \ldots, x_n) \in \mathcal{X}_n$ if for some $i \in \{1, \ldots, n-m+1\}, s = (x_1, \ldots, x_{i+m-1})$. Define the function $K : \mathcal{X}_n \times \mathcal{X}_n \to \mathbb{R}$ as the number of common substrings of its aguments, that is

$$K(x,y) = \sum_{s \in \{0,1\}^n} \mathbf{1}_{\{s \text{ is a substring of both } x \text{ and } y\}}, \text{ for } x, y, \in \mathcal{X}_n$$

.

Prove that K is a kernel function. Determine a feature map Φ defined of $\mathcal{X}_n$, mapping to some Hilbert space $\mathcal{H}$ for which $K(x,y)$ is the inner product of $\Phi(x)$ and $\Phi(y)$. What is the dimension of $\mathcal{H}$?

Solution:

Let x a be vector with binary components. In this case there are in general at most 2^m possible substrings. Using that information we define a feature map for all possible combinations of substrings $\Phi : \mathcal{X}_n \mapsto \{0,1\}^{2^m}$. In particular we define Φ such that it is one if there is a substring in x:

$$\Phi(x)^{(i)} = \begin{cases} 1, \text{ if } s_i \in x \\ 0, \text{ otherwise} \end{cases}$$

The resulting vector has 2^m components. Now applying this feature map to two vectors we see that it is a kernel:

$$\begin{aligned} \langle \Phi(x), \Phi(y) \rangle &= \sum_{s \in \{0,1\}^m} \mathbf{1}_{\{s \text{ is a substring of } x\}} \sum_{s \in \{0,1\}^m} \mathbf{1}_{\{s \text{ is a substring of } y\}} \\ &= \sum_{s \in \{0,1\}^m} \mathbf{1}_{\{s \text{ is a substring of both } x \text{ and } y\}} \\ &= K(x,y) \end{aligned}$$

Now finally consider the dimension of the Hilbert space. Say s to be the set of the substrings length m (less than n) . The cardinality is 2^m and so is the dimension of the hilbert space