Machine Learning

Group 16 - Assigment 4

Luca Pfrang	5332329
Tran Luong Bang	5365615
Louis Ngo	4506205
Vu Thi Hoang Anh	5367008
Tidiane Ndir	5364532

Freiburg, November 25 2021

1. Support Vector Machine

a. Explain the “kernel trick” and why use it in SVMs.

A kernel trick is a simple method where a non-linear data is projected onto a higher dimension space so as to make it easier to classify the data where it could be linearly divided by a plane. The kernel trick replaces the mapping and following dot product operations by a simple calculation in the input space.

Some kernel functions:

- Polynomial kernel
- Gaussian Radial basis function kernels
- Sigmoid kernel

Kernel trick is widely used in the SVM model to bridge linearity and non-linearity.

b. What is the difference between hard-and soft-margin SVM?

The difference between a hard margin and a soft margin in SVMs lies in the separability of the data. If our data is linearly separable, we go for a hard margin. However, if this is not the case, it won’t be feasible to do that. In the presence of the data points that make it impossible to find a linear classifier, we would have to be more lenient and let some of the data points be misclassified. In this case, a soft margin SVM is appropriate.

Sometimes, the data is linearly separable, but the margin is so small that the model becomes prone to overfitting or being too sensitive to outliers. Also, in this case, we can opt for a larger margin by using soft margin SVM in order to help the model generalize better.

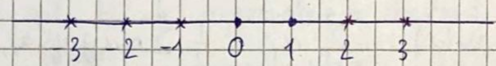
2. Linear Separability

ML Exercise 4:

2 Linear Seperability:

1. A sketch of the data: Our input space is 1D.

   (Here we denote output 1 as • and output -1 as *)

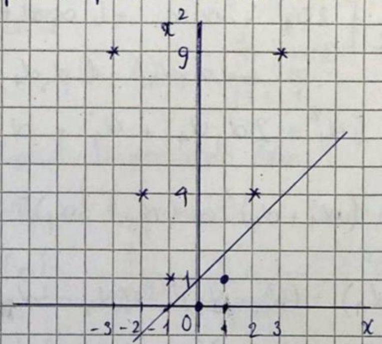


⟨Figure 1⟩

It is not linearly separable.

2. Using the mapping $g: \mathbb{R} \to \mathbb{R}^2$ where $g(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$

   Our feature space becomes 2D and a plot of it is provided below



⟨Figure 2⟩

In this feature space, we can draw a separating hyperplane. Its parameter are calculated as below:

First of all, we assume that the hyperplane is only dependent on the border points, which are $x = [-1 \quad 0 \quad 1 \quad 2]$, $x^2 = [1 \quad 0 \quad 1 \quad 4]$

Let's denote them as $[x_1 \quad x_2 \quad x_3 \quad x_4]$

In order to solve the optimization problem, let's first calculate the pairwise dot products in feature space:

$\langle x_1, x_1 \rangle = (-1).(-1) + 1.1 = 2$

$\langle x_1, x_2 \rangle = (-1).0 + 1.0 = 0$

$\langle x_1, x_3 \rangle = (-1).1 + 1.1 = 0$

$\langle x_1, x_4 \rangle = (-1).2 + 1.4 = 2$

Similarly, we have: $\langle x_2, x_3 \rangle = 0$; $\langle x_2, x_4 \rangle = 0$; $\langle x_2, x_2 \rangle = 0$

$\langle x_3, x_3 \rangle = 2$; $\langle x_3, x_4 \rangle = 6$

$\langle x_4, x_4 \rangle = 20$

We want to find w, b such that: $\frac{1}{2}\|w\|^2$ is min

subject to $y_i(\langle w, x_i\rangle + b) \geq 1 \quad \forall i = 1,2,3,4$

This can be rewritten as

$$\text{minimize } L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{4} \alpha_i \left(y_i(\langle w, x_i\rangle + b) - 1\right)$$

and further represented as an optimization problem in α_i

$$\underset{\alpha}{\text{argmax}}\ W(\alpha) = \sum_{i=1}^{4} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{4} \alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle \quad (1)$$

subject to $\sum_{i=1}^{4} \alpha_i y_i = 0$ or $-\alpha_1 + \alpha_2 + \alpha_3 - \alpha_4 = 0$

Substituting the pre-calculated dot products into (1) gives:

$$W(\alpha) = \sum_{i=1}^{4}\alpha_i - \frac{1}{2}\Big[2\alpha_1^2 + 2\alpha_1\alpha_4 + 6\,\alpha_3\alpha_4(-1) + 2\alpha_3^2 + 2\alpha_1\alpha_4 + 6\,\alpha_3\alpha_4(-1) + 20\,\alpha_4^2\Big]$$

$$= \sum_{i=1}^{4}\alpha_i - \left[\alpha_1^2 + 2\alpha_1\alpha_4 + \alpha_4^2 + \alpha_3^2 - 6\alpha_3\alpha_4 + 9\alpha_4^2\right]$$

$$= \sum_{i=1}^{4}\alpha_i - (\alpha_1+\alpha_4)^2 - (\alpha_3 - 3\alpha_4)^2$$

$$= 2(\alpha_1+\alpha_4) - (\alpha_1+\alpha_4)^2 - (\alpha_3 - 3\alpha_4)^2$$

$$= 1 - (\alpha_1+\alpha_4-1)^2 - (\alpha_3-3\alpha_4)^2$$

$$W(\alpha)_{\max} = 1 \text{ when } \begin{cases} \alpha_1 + \alpha_4 = \alpha_2 + \alpha_3 = 1 \\ \alpha_3 - 3\alpha_4 = 0 \end{cases}$$

$$w = \sum_{i=1}^{4} \alpha_i y_i x_i = \alpha_1(-1)\begin{bmatrix}-1\\1\end{bmatrix} + \alpha_2(1)\begin{bmatrix}0\\0\end{bmatrix} + \alpha_3(1)\begin{bmatrix}1\\1\end{bmatrix} + \alpha_4(-1)\begin{bmatrix}2\\4\end{bmatrix}$$

$$= \begin{bmatrix}\alpha_1 + \alpha_3 - 2\alpha_4\\ -\alpha_1 + \alpha_3 - 4\alpha_4\end{bmatrix}$$

$$= \begin{bmatrix}(\alpha_1+\alpha_4) + (\alpha_3 - 3\alpha_4)\\ (-\alpha_1 - \alpha_4) + (\alpha_3 - 3\alpha_4)\end{bmatrix} = \begin{bmatrix}1\\-1\end{bmatrix}$$

Substituting w to any of the border points gives $b = 1$

Hyperplane: $\langle [1 \ \ -1]^T, x\rangle + 1 = 0$

3. For $x = \frac{1+\sqrt{5}}{2}$, we have $\mathbf{x} = [x \ \ x^2]^T = \left[\frac{1+\sqrt{5}}{2} \quad \frac{3+\sqrt{5}}{2}\right]^T$

$$\left\langle \begin{bmatrix}1\\-1\end{bmatrix}, \begin{bmatrix}\frac{1+\sqrt{5}}{2}\\ \frac{3+\sqrt{5}}{2}\end{bmatrix}\right\rangle + 1 = -1 + 1 = 0$$

Hence the point lies on the separating hyperplane.

3. Soft-Margins and Regression in SVM:

1. To accommodate soft margins, the objective function of SVM is modified by adding the sum of slack variables, controlled by a certain factor C:

   $$\tau(w, \xi) = \frac{1}{2}|w|^2 + C\sum_{i=1}^{m} \xi_i$$

   subject to $y_i(< w, x_i > + b) \geq 1 - \xi_i$ for all i = 1,...,m

   The hyper parameter C controls the strength of this regularization. When C is large, it allows for very few margin "violations" and hence less points will become support vectors.

2. To use support vector in regression, we introduce precision term ϵ which determines how much our predictions can deviate from the ground truth. Any point within the $[-\epsilon, +\epsilon]$ tube is considered as having 0 loss and any point outside have a loss of $|\xi| - \epsilon$. This is called the ϵ intensive loss function:

   $$\xi = 0 \; if \; |\xi| \leq \epsilon$$
   $$\xi = |\xi| - \epsilon \; otherwise$$