Exercise set 2

Advanced Course in Machine Learning

Enrico Buratto

Exercise 1

Task a

First of all, we can write the sigmoid function with respect to $\mathbf{w}^T\mathbf{x}$:

$$\text{sigm}(\mathbf{w}^\top \mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}}}$$

Now, we can say that

$$\text{sigm}(-z) = 1 - \text{sigm}(z)$$

This is true, in fact:

$$\begin{aligned}
&1 - \frac{1}{1+e^{-z}} = \frac{1}{1+e^{z}} \\
\Rightarrow &\frac{1+e^{-z}-1}{1+e^{-z}} = \frac{1}{1+e^{z}} \\
\Rightarrow &\frac{e^{-z}}{1+e^{-z}} = \frac{1}{1+e^{z}} \\
\Rightarrow &\frac{1/e^{z}}{1+\frac{1}{e^{z}}} = \frac{1}{1+e^{z}} \\
\Rightarrow &\frac{1}{e^{z}} = \frac{1+e^{-z}}{1+e^{z}} \\
\Rightarrow &\frac{e^{z}+e^{-z}e^{z}}{1+e^{z}} = 1 \\
\Rightarrow &\frac{1+e^{z}}{1+e^{z}} = 1
\end{aligned}$$

We then have

$$P(y \mid \mathbf{w}, \mathbf{x}) = \begin{cases} \text{sigm}\left(\mathbf{w}^\top \mathbf{x}\right) & \text{for } y = 1 \\ 1 - \text{sigm}\left(\mathbf{w}^\top \mathbf{x}\right) = \text{sigm}\left(-\mathbf{w}^\top \mathbf{x}\right) & \text{for } y = -1 \end{cases}$$

We can write the formula above in a compact way:

$$P(y \mid \mathbf{w}, \mathbf{x}) = \text{sigm}(y\mathbf{w}^\top \mathbf{x})$$

So now we have:

$$\begin{aligned}
&-\sum_n \log\left(y_n \mid \mathbf{w}, \mathbf{x}_n\right) \\
=&-\sum_n \log\left(\operatorname{sigm}\left(y_n \mathbf{w}^\top \mathbf{x}_n\right)\right). \\
=&-\sum_n \log\left(\frac{1}{1+e^{-y_n \mathbf{w}^\top \mathbf{x}_n}}\right) \\
=&\sum_n \log\left(1+e^{-y_n \mathbf{w}^\top \mathbf{x}_n}\right)
\end{aligned}$$

Exercise 2

Task a

The results of this task are illustrated in the picture below. The legend and explanation are as follows:

- The decision surface that separate the two classes, *i.e.* the optimal hyperplane, is represented by the blue continuous line. This hyperplane is characterized by $\mathbf{w}^\top \mathbf{x} + \mathbf{b} = 0$. The two orange dashed lines are, respectively, the negative hyperplane (on the left of the optimal) and the positive hyperplane (on the right of the optimal). These two hyperplanes are characterized by, respectively, $\mathbf{w}^\top \mathbf{x} + \mathbf{b} = -1$ and $\mathbf{w}^\top \mathbf{x} + \mathbf{b} = 1$;
- The weight vector is represented by the green line. Note that the direction of the vector is the vector itself, while the lenght is dependent on the bias;
- The support vectors corresponding to $\alpha_i > 0$ are pointed by the two yellow circles;
- The margin of the classifier is represented by the brown arrow. Since this is a maximum margin, this is equal to $\frac{2}{\|\mathbf{w}\|}$;
- The samples that violate the margin are represented by the two additional crosses:
  - For the grey cross the slack variable ϵ_i is between 0 and 1;
  - For the yellow cross the slack variable ϵ_i is bigger than one.

Task b

We have the SVM primal problem given by

$$\begin{gathered}
\min \tfrac{1}{2}\|w\|^2 \\
\text{s.t.} \quad \left(\mathbf{w}^\top \mathbf{x}_n + b\right) y_n \geqslant 1 \\
\Rightarrow \left(\mathbf{w}^\top \mathbf{x}_n + b\right) y_n - 1 \geqslant 0
\end{gathered}$$

In order to prove the equivalence to its dual problem, we first write the Lagrangian for it; this is

$$\Lambda(\mathbf{w}, b, \alpha) = \frac{1}{2}\mathbf{w}^\top \mathbf{w} - \sum_n \alpha_n \left[\left(\mathbf{w}^\top \mathbf{x}_n + b\right) y_n - 1\right]$$

Now, in order to find the optimal solution we compute the gradient of Λ with respect to $\mathbf{w}$ and b, and we get

$$\nabla\Lambda(\mathbf{w}, b, \alpha) = \begin{pmatrix} \sum_n \alpha_n x_n y_n - \mathbf{w} \\ -\sum_n \alpha_n y_n \end{pmatrix}$$

Then we can solve the linear problem $\nabla\Lambda(\mathbf{w}, b, \alpha) = \mathbf{0}$, and we get

$$\mathbf{w} = \sum_n \alpha_n x_n y_n$$

$$\sum_n \alpha_n y_n = 0$$

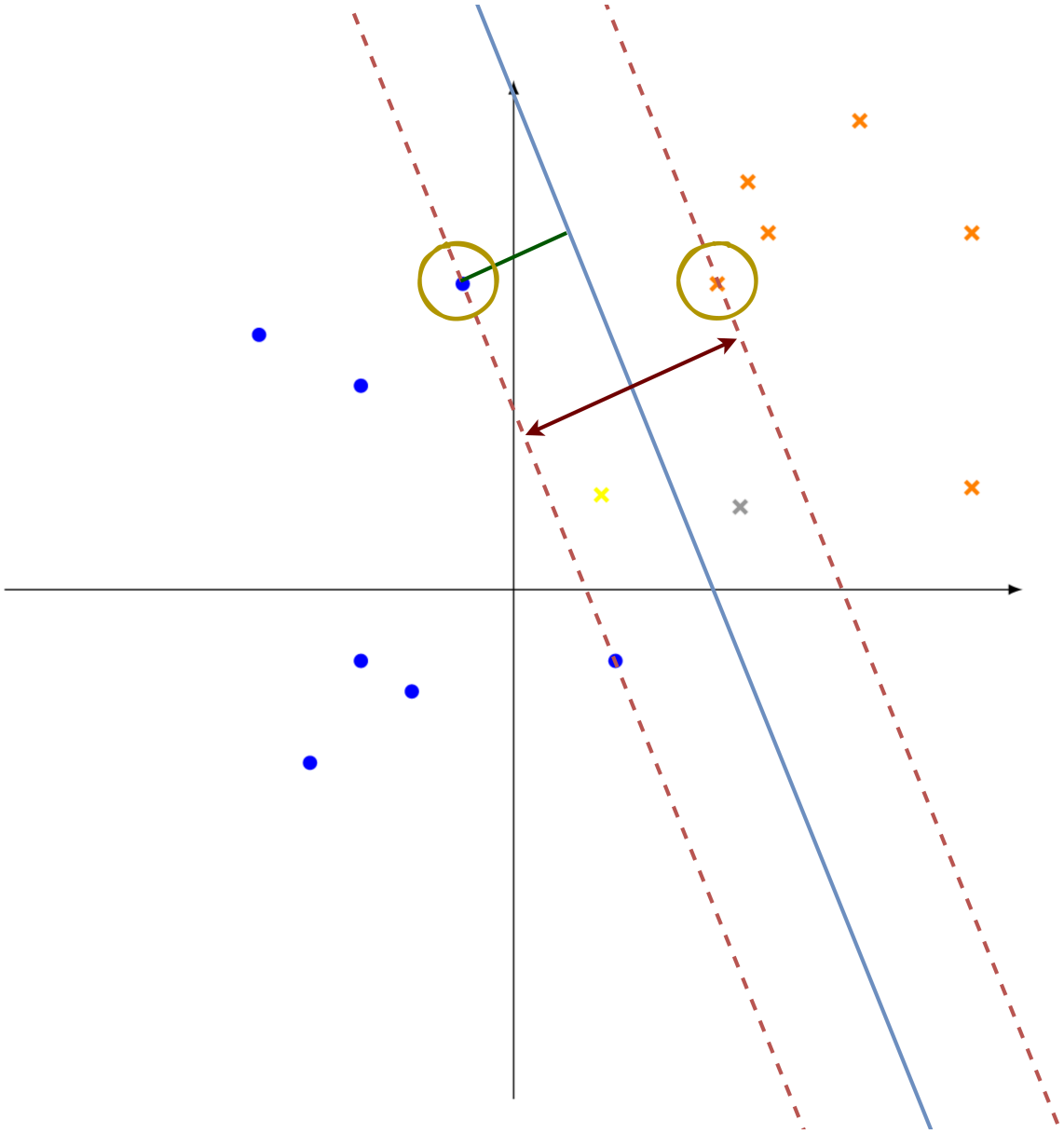

Figure 1: Support Vector Machines components.

We can then plug these results into the Lagrangian for the primal problem and simplify it to check if we get the expected dual problem formulation. The calculations follow:

$$
\begin{aligned}
\Lambda(\mathbf{w}, b, \alpha) = & \frac{1}{2}\left(\sum_i \alpha_i y_i \mathbf{x}_i^\top\right)\left(\sum_j \alpha_j \mathbf{x}_j y_j\right) \\
& + \sum_i \alpha_i y_i b \\
& + \sum_i \alpha_i \\
& - \sum_i \alpha_i y_i \left(\sum_i \alpha_i y_i \mathbf{x}_i^\top\right) \mathbf{x}^\top \\
= & \frac{1}{2} \sum_i \sum_j \alpha_j y_j \alpha_i y_i \mathbf{x}_i^\top \mathbf{x}_j \\
& + \sum_i \alpha_i \\
& - \sum_i \sum_j \alpha_j y_j \alpha_i y_i \mathbf{x}_i^\top \mathbf{x}_j \\
= & \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_j y_j \alpha_i y_i \mathbf{x}_i^\top \mathbf{x}_j
\end{aligned}
$$

Maximing this final problem is exactly the same of minimizing the primal problem; therefore, this is proved.