

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 $\mathbf{w}^T \mathbf{x}$:

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

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

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

We can write the formula above in a compact way:

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

So now we have

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

which was the result we were searching for.

Task b

In order to compute the gradient, as suggested we analyze the negative log-likelihood on different cases.

Case $y = -1$

We have that

$$\mathcal{L} = \sum_n \log(1 + e^{\mathbf{w}^\top \mathbf{x}_n})$$

and we want to show that

$$\nabla \mathcal{L} = \mathbf{g} = \mathbf{X}^T (\mu - \mathbf{0}) = \mathbf{X}^T \mu$$

with $\mu_n = \text{sigm}(\mathbf{w}^\top \mathbf{x}_n)$. First of all, we can rewrite \mathcal{L} as a sigmoid for convenience:

$$\begin{aligned}
\mathcal{L} &= - \sum_n \log\left(\frac{1}{1 + e^{\mathbf{w}^\top \mathbf{x}_n}}\right) \\
&= - \sum_n \log[\text{sigm}(-\mathbf{w}^\top \mathbf{x}_n)] \\
&= - \sum_n \log[1 - \text{sigm}(\mathbf{w}^\top \mathbf{x}_n)]
\end{aligned}$$

Let's now compute the partial derivative with respects to \mathbf{w}^\top :

$$\begin{aligned}
& \frac{\partial}{\partial \mathbf{w}^\top} \log[1 - \text{sigm}(\mathbf{w}^\top \mathbf{x}_n)] \\
&= \frac{1}{1 - \text{sigm}(\mathbf{w}^\top \mathbf{x}_n)} \cdot \frac{\partial}{\partial \mathbf{w}^\top} (1 - \text{sigm}(\mathbf{w}^\top \mathbf{x}_n)) \\
&= \frac{1}{1 - \text{sigm}(\mathbf{w}^\top \mathbf{x}_n)} \cdot \frac{\partial}{\partial \mathbf{w}^\top \mathbf{x}_n} \text{sigm}(\mathbf{w}^\top \mathbf{x}_n) \cdot \frac{\partial}{\partial \mathbf{w}^\top} \mathbf{w}^\top \mathbf{x}_n
\end{aligned}$$

Now, we know that

$$\begin{aligned}
& \frac{\partial}{\partial \mathbf{w}^\top \mathbf{x}} \text{sigm}(\mathbf{w}^\top \mathbf{x}) \\
&= \frac{\partial}{\partial \mathbf{w}^\top \mathbf{x}} (1 + e^{-\mathbf{w}^\top \mathbf{x}})^{-1} = e^{-\mathbf{w}^\top \mathbf{x}} (1 + e^{-\mathbf{w}^\top \mathbf{x}})^{-2} \\
&= \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}}} \cdot \frac{e^{-\mathbf{w}^\top \mathbf{x}}}{1 + e^{-\mathbf{w}^\top \mathbf{x}}} \\
&= \text{sigm}(\mathbf{w}^\top \mathbf{x}) (1 - \text{sigm}(\mathbf{w}^\top \mathbf{x}))
\end{aligned}$$

Therefore, the previous partial derivative becomes the following:

$$\begin{aligned} \frac{1}{1 - \text{sigm}(\mathbf{w}^\top \mathbf{x}_n)} \cdot (1 - \text{sigm}(\mathbf{w}^\top \mathbf{x}_n)) \cdot \text{sigm}(\mathbf{w}^\top \mathbf{x}_n) \mathbf{x}_n \\ = \text{sigm}(\mathbf{w}^\top \mathbf{x}_n) \mathbf{x}_n \end{aligned}$$

We can now compute the gradient and therefore prove the equivalence:

$$\begin{aligned} \nabla \mathcal{L} &= \frac{\partial}{\partial \mathbf{w}^\top} \mathcal{L} = \text{sigm}(\mathbf{w}^\top \mathbf{x}_n) \mathbf{x}_n \\ \Rightarrow - \sum_n \text{sigm}(\mathbf{w}^\top \mathbf{x}_n) \mathbf{x}_n &= \mathbf{X}^\top \text{sigm}(\mathbf{w}^\top \mathbf{x}_n) = \mathbf{X}^\top \mu \end{aligned}$$

Case $y = 1$

The procedure for this case is almost the same. This time we have that

$$\mathcal{L} = \sum_n \log(1 + e^{-\mathbf{w}^\top \mathbf{x}_n})$$

and we want to show that

$$\nabla \mathcal{L} = \mathbf{g} = \mathbf{X}^\top (\mu - \mathbf{1})$$

with $\mu = \text{sigm}(\mathbf{w}^\top \mathbf{x}_n)$ and $\mathbf{1}$ a vector of ones. As we did before, we can rewrite \mathcal{L} as a sigmoid for convenience:

$$\begin{aligned} \mathcal{L} &= - \sum_n \log\left(\frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}_n}}\right) \\ &= - \sum_n \log[\text{sigm}(\mathbf{w}^\top \mathbf{x}_n)] \end{aligned}$$

Let's now compute the partial derivative with respects to \mathbf{w}^\top :

$$\begin{aligned} \frac{\partial}{\partial \mathbf{w}^\top} \log[\text{sigm}(\mathbf{w}^\top \mathbf{x}_n)] \\ = \frac{1}{\text{sigm}(\mathbf{w}^\top \mathbf{x}_n)} \cdot \frac{\partial}{\partial \mathbf{w}^\top} (\text{sigm}(\mathbf{w}^\top \mathbf{x}_n)) \\ = \frac{1}{\text{sigm}(\mathbf{w}^\top \mathbf{x}_n)} \cdot \frac{\partial}{\partial \mathbf{w}^\top \mathbf{x}_n} \text{sigm}(\mathbf{w}^\top \mathbf{x}_n) \cdot \frac{\partial}{\partial \mathbf{w}^\top} \mathbf{w}^\top \mathbf{x}_n \\ = \frac{1}{\text{sigm}(\mathbf{w}^\top \mathbf{x}_n)} \cdot (1 - \text{sigm}(\mathbf{w}^\top \mathbf{x}_n)) \cdot \text{sigm}(\mathbf{w}^\top \mathbf{x}_n) \mathbf{x}_n \\ = (1 - \text{sigm}(\mathbf{w}^\top \mathbf{x}_n)) \mathbf{x}_n \end{aligned}$$

We can now compute the gradient and therefore prove the equivalence in the same way. In the end we get

$$\nabla \mathcal{L} = \mathbf{X}^\top (\mu - \mathbf{1})$$

which is the result we had to find.

Now we can prove also that the Hessian matrix \mathbf{H} of the loss is equal to

$$\mathbf{H} = \mathbf{X}^\top \mathbf{S} \mathbf{X}$$

We can compute the Hessian starting from the gradient we calculated in the previous step. For convenience, let's first put the two cases together into a general result:

$$\nabla \mathcal{L} = (\text{sigm}(\mathbf{w}^\top \mathbf{x}_n) - y_n) \mathbf{x}_n$$

(Note: y_n can be either 0 or 1 in this case, but this is a more general result, thus without loss of generality). The Hessian is then

$$\nabla^2 \mathcal{L} = \frac{\partial}{\partial \mathbf{w} \partial \mathbf{w}^\top} \mathcal{L} = \mathbf{x}_n^\top \text{sigm}(\mathbf{w}^\top \mathbf{x}_n) (1 - \text{sigm}(\mathbf{w}^\top \mathbf{x}_n)) \mathbf{x}_n$$

And therefore we have that

$$\begin{aligned} \sum_n \mathbf{x}_n^\top \text{sigm}(\mathbf{w}^\top \mathbf{x}_n) (1 - \text{sigm}(\mathbf{w}^\top \mathbf{x}_n)) \mathbf{x}_n \\ = \mathbf{X}^\top \text{sigm}(\mathbf{w}^\top \mathbf{x}) (1 - \text{sigm}(\mathbf{w}^\top \mathbf{x})) \mathbf{X} \\ = \mathbf{X}^\top \mathbf{S} \mathbf{X} \end{aligned}$$

Task c

In order to prove that the Hessian matrix \mathbf{H} is positive semidefinite, we have to prove that $\mathbf{v}^\top \mathbf{H} \mathbf{v} \geq 0$ for a generic vector $\mathbf{v} \in \mathbb{R}^d$.

Therefore, if we are able to prove that $\mathbf{X}^\top \mathbf{S} \mathbf{X} \geq 0$, *i.e.* \mathbf{S} is positive semidefinite, we have also proved that \mathbf{H} is positive semidefinite.

Now, \mathbf{S} is positive semidefinite because its eigenvalues are positive; in fact, \mathbf{S} has only one eigenvalue

$$\lambda = -[\mu(1 - \mu)]^2 + \mu(1 - \mu)$$

and this eigenvalue is always greater than 0 because of the definition of μ , which is $0 < \text{sigm}(\mathbf{w}^\top \mathbf{x}) \leq 1$.

Wrapping up, \mathbf{S} is positive semidefinite, therefore \mathbf{H} is positive semidefinite, therefore the loss is convex.

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 length 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.

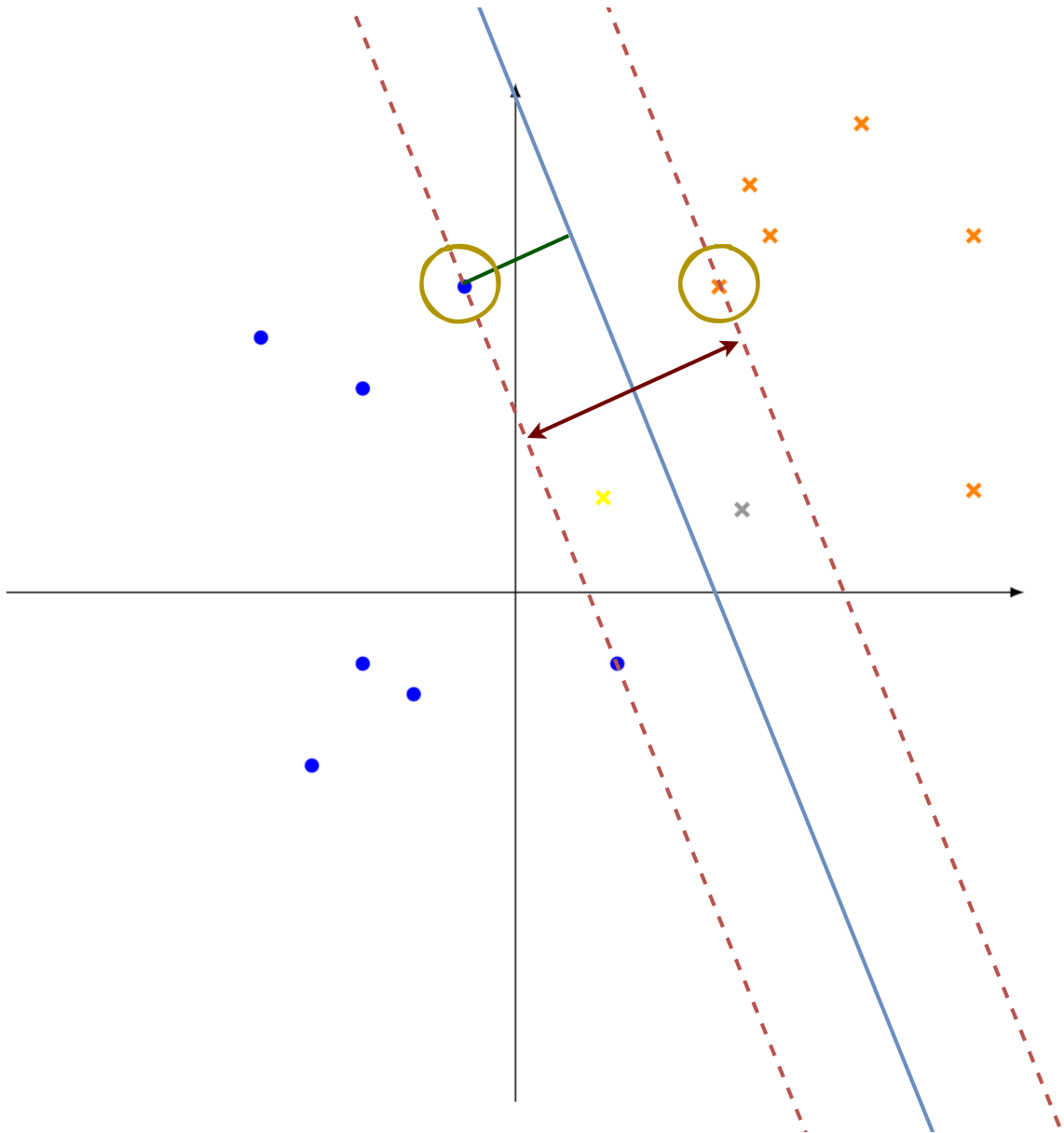


Figure 1: Support Vector Machines components.

Task b

We have the SVM primal problem given by

$$\begin{aligned} & \min \frac{1}{2} \|w\|^2 \\ & \text{s.t. } (\mathbf{w}^\top \mathbf{x}_n + b) y_n \geq 1 \\ & \Rightarrow (\mathbf{w}^\top \mathbf{x}_n + b) y_n - 1 \geq 0 \end{aligned}$$

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

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

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

$$\begin{aligned} \mathbf{w} &= \sum_n \alpha_n x_n y_n \\ \sum_n \alpha_n y_n &= 0 \end{aligned}$$

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:

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

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