Machine Learning I: Foundations
Exercise Sheet 2

Prof. Marius Kloft TA:Billy Joe Franks

19.05.2021
Deadline: 18.05.2020

1) (MANDATORY) 10 Points
Interestingly the linear hard-margin SVM, given by

$$\max_{\gamma, b\in\mathbb{R}, \mathbf{w}\in\mathbb{R}^d} \ \gamma \qquad (1)$$
$$\text{s.t. } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq \|\mathbf{w}\|\gamma, \ \forall i \in \{1, \ldots, n\},$$

requires only two (non-equal) training points (with opposite labels) to find a separating hyperplane. Let $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $Y := \{y_1, \ldots, y_n\}$, with $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$, be a dataset. Let γ^*, $\mathbf{w}^*$, and b^* be the optimal solution to the above optimization problem (1) on X, Y. You may assume $w_1 \neq 0$.

a) Find a minimal dataset (X', Y') with $|X'| = |Y'| = 2$ (consisting of only two data points) with the same hard-margin SVM solution (Eq. (1)) as for the dataset (X, Y) , that is, γ^*, $\mathbf{w}^*$, and b^*.

If we think about this exercise a bit its easy to figure out that the points we need to choose need to be on either side of the separating hyperplane each with distance margin γ from the separating hyperplane. Additionally the line going through both points needs to be orthogonal to the seperating hyperplane. This is easy to achieve. First we find a point on the hyperplane $H = \{\mathbf{x} \in \mathbb{R}^d | w^{*T}\mathbf{x} + b^* = 0\}$. Since we may assume that $w_1 \neq 0$, we will look for a point $\mathbf{x}$, with $x_1 \in \mathbb{R}$ and $\forall i \neq 1 : x_i = 0$. Lets call this point $\mathbf{x}^*$.

$$0 = w^{*T}\mathbf{x}^* + b^* = w_1^* x_1^* + b^*$$
$$x_1^* = -\frac{b^*}{w_1^*}$$

Now that we have a point on H, we just need to add (subtract) $\gamma^* \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|}$. As such the points we choose are

- $\mathbf{x}_1^* := x^* + \gamma^* \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|}$ with $y_1 = 1$
- $\mathbf{x}_2^* := x^* - \gamma^* \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|}$ with $y_2 = -1$.

Just to check that $\mathbf{w}^*$ and b^* actually fulfill the constraints for $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$.

$$\begin{aligned} & y_1 \left(\mathbf{w}^{*T} \left(\mathbf{x}^* + \gamma^* \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|} \right) + b^* \right) \\ =& y_1 \left(\left(w_1^* \frac{-b^*}{w_1^*} + \gamma^* \frac{\|\mathbf{w}^*\|^2}{\|\mathbf{w}^*\|} \right) + b^* \right) \\ =& 1 - 1 + \gamma^* \|\mathbf{w}^*\| = \gamma^* \|\mathbf{w}^*\| \geq \gamma^* \|\mathbf{w}^*\| \end{aligned}$$

$$\begin{aligned} & y_2 \left(\mathbf{w}^{*T} \left(\mathbf{x}^* - \gamma^* \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|} \right) + b^* \right) \\ =& y_2 \left(\left(w_1^* \frac{-b^*}{w_1^*} - \gamma^* \frac{\|\mathbf{w}^*\|^2}{\|\mathbf{w}^*\|} \right) + b^* \right) \\ =& -1(1 - 1 - \gamma^* \|\mathbf{w}^*\|) = \gamma^* \|\mathbf{w}^*\| \geq \gamma^* \|\mathbf{w}^*\| \end{aligned}$$

Since we expect both of these vectors to be support vectors it makes sense that the constraints are fulfilled with equality.

b) Prove that, for your choice of X' and Y' in a), γ, $\mathbf{w}^*$, and b^* are optimal solutions of (1).

Lets assume the contrary, i.e. there exist γ', $\mathbf{w}'$, and b', such that the constraints of (1) are fulfilled for $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$, and $\gamma' > \gamma^*$. To this end first note the following

$$\left\|\mathbf{w}'\right\| \geq \mathbf{w}'^T \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|} \tag{2}$$

$$\gamma' \left\|\mathbf{w}'\right\| > \gamma^* \mathbf{w}'^T \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|} \tag{3}$$

2 follows directly from the Cauchy-Schwarz inequality and 3 follows from 2 and $\gamma' > \gamma^*$ and $\|\mathbf{w}'\| > 0$, otherwise one constraint wont be fulfilled, (we also need all $\gamma > 0$, otherwise this statement is not true, since the hyperplane would flip if you chose the points on opposite sides, we will ignore this case). Now consider the constraints for $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$

$$\begin{aligned}
& y_1 \left(\mathbf{w}'^T \left(\mathbf{x}^* + \gamma^* \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|} \right) + b' \right) \\
=& y_1 \left(\left(w_1' \frac{-b^*}{w_1^*} + \gamma^* \frac{\mathbf{w}'^T \mathbf{w}^*}{\|\mathbf{w}^*\|} \right) + b' \right) \\
=& \left(\gamma^* \frac{\mathbf{w}'^T \mathbf{w}^*}{\|\mathbf{w}^*\|} \right) + \left(w_1' \frac{-b^*}{w_1^*} + b' \right) \geq \gamma' \left\|\mathbf{w}'\right\| \\
& \left(w_1' \frac{-b^*}{w_1^*} + b' \right) \geq \gamma' \left\|\mathbf{w}'\right\| - \left(\gamma^* \frac{\mathbf{w}'^T \mathbf{w}^*}{\|\mathbf{w}^*\|} \right)
\end{aligned}$$

$$\begin{aligned}
& y_2 \left(\mathbf{w}'^T \left(\mathbf{x}^* - \gamma^* \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|} \right) + b' \right) \\
=& y_2 \left(\left(w_1' \frac{-b^*}{w_1^*} - \gamma^* \frac{\mathbf{w}'^T \mathbf{w}^*}{\|\mathbf{w}^*\|} \right) + b' \right) \\
=& \left(\gamma^* \frac{\mathbf{w}'^T \mathbf{w}^*}{\|\mathbf{w}^*\|} \right) + \left(w_1' \frac{b^*}{w_1^*} - b' \right) \geq \gamma' \left\|\mathbf{w}'\right\| \\
& \left(w_1' \frac{b^*}{w_1^*} - b' \right) \geq \gamma' \left\|\mathbf{w}'\right\| - \left(\gamma^* \frac{\mathbf{w}'^T \mathbf{w}^*}{\|\mathbf{w}^*\|} \right)
\end{aligned}$$

Now using (3) we know that $\gamma' \|\mathbf{w}'\| - \left(\gamma^* \frac{\mathbf{w}'^T \mathbf{w}^*}{\|\mathbf{w}^*\|} \right) > 0$, we will call this quantity ϵ. Also, we will call $c := w_1' \frac{b^*}{w_1^*} - b'$. Now substituting the above inequalities we get

$$-c \geq \epsilon, c \geq \epsilon$$

Since $\epsilon > 0$ this is a contradiction. As such γ^*, $\mathbf{w}^*$, and b^* must be optimal for $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$.

c) How is this choice of X' and Y' related to the nearest centroid classifier (NCC)? (Answer this question with at most 5 sentences.)

The NCC calculates its hyperplane equivalently to the idea used above. In this analogy $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ are the centroids of their respective classes in some NCC problem. However, $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ cannot simply be chosen as the centroids of X and Y, this is only the case in some special instances.

2) Consider the soft-margin SVM as in the lecture. Now assume we do not optimize over b and it is fixed to $b = 0$. Construct a dataset for which any classifier learned (with $b = 0$) performs poorly. Does any classifier with b fixed to a different constant like $b = 1$ still perform poorly?

If $b = 0$, then the hyperplane has to intersect $\mathbf{0}$. From this we can easily follow that the dataset needs to be sufficiently far away from the origin, and the optimal separating hyperplane (according to SVM with b learned) would need to be orthogonal to the line from the origin to the centroid of all data. If we then carefully construct the dataset it is easy to achieve a best performance of 50% accuracy. In fact just choosing the dataset to live entirely on a line intersecting the origin would suffice, since the learned margin would be 0 and the prediction would always be sign(0), essentially a coin flip.

3) Construct a worst-case dataset for the nearest centroid classifier (NCC). This dataset should be easily (not necessarily linearly) separable and the NCC should behave as poorly as possible on this training dataset. **Hint:** To this end you will have to figure out for yourself how poorly the NCC can perform. Is it possible for the NCC to have 0% accuracy?

Figuring out the worst-case for NCC is a bit hard. Notice that, if we take some dataset that performs arbitrarily, then moving a single point from one of the two classes will move the hyperplane, if we use one point from each class then we can move the centroids arbitrarily and thus the hyperplane arbitrarily. Thus, we can flip any NCC hyperplane and thus its prediction, by moving just 2 points. In the worst case, these two points will be classified correctly, however, all other will not. Thus it is possible to achieve arbitrarily bad performance, however, only if the dataset can have an arbitrary size. The worst accuracy will always be $\frac{2}{n}$, where n is the dataset size.

4) Solve programming task 2.