# Machine Learning I: Foundations
# Exercise Sheet 2

Prof. Marius Kloft TA:Billy Joe Franks

16.05.2020
Deadline: 12.05.2020

**1) (MANDATORY) 10 Points**

In this exercise we will consider the soft-margin SVM

$$\begin{aligned} &\min \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{s.t. } 1-\xi_i - y_i(\mathbf{w}^T\mathbf{x}_i + b) \le 0, \ \ -\xi_i \le 0, \ \forall i \in \{1,\dots,n\} \end{aligned}$$

Let $a \in (-1, 2)$, consider the one-dimensional datapoints $(2,1), (a,1), (-2,-1)$. For each datapoint the second value is the label. This is a binary classification dataset. We will investigate the behavior of $C$.

a) Determine for (1) $w = \frac{1}{2}, b = 0$ and (2) $w = \frac{2}{a+2}, b = \frac{2-a}{a+2}$ the objective function depending on $a$ and $C$.

Let $o_{\mathbf{w},b}$ denote the optimal objective value of the soft-margin SVM given $\mathbf{w}$ and $b$. Then we claim the functions are as follows:

| $\mathbf{w}$ | $b$ | $o_{\mathbf{w},b}$ |
|---|---|---|
| $\frac{1}{2}$ | 0 | $\frac{1}{8} + C\frac{2-a}{2}$ |
| $\frac{2}{a+2}$ | $\frac{2-a}{a+2}$ | $\frac{2}{(a+2)^2}$ |

The first objective is simply $\frac{1}{2}\left\|\mathbf{w}\right\|^2$ plus a single $\xi_{(a,1)}$, all others are 0. The second is just $\frac{1}{2}\left\|\mathbf{w}\right\|^2$, all $\xi_i$ are 0. This can be checked by substituting all relevant values.

Consider $\mathbf{w} = \frac{1}{2}$ and $b = 0$:

$$\begin{aligned}
1 - \frac{1}{2}2 &= 0 && \leq \xi_{(2,1)} \\
1 - \frac{1}{2}a &= \frac{2-a}{2} && \leq \xi_{(a,1)} \\
1 + \frac{1}{2}(-2) &= 0 && \leq \xi_{(-2,-1)}
\end{aligned}$$

Consider $\mathbf{w} = \frac{2}{a+2}$ and $b = \frac{2-a}{a+2}$:

$$\begin{aligned}
1 - \left(\frac{2}{a+2}2 + \frac{2-a}{a+2}\right) &= 1 - \frac{6-a}{2+a} && \leq \xi_{(2,1)} \\
1 - \left(\frac{2}{a+2}a + \frac{2-a}{a+2}\right) &= 0 && \leq \xi_{(a,1)} \\
1 + \left(\frac{2}{a+2}(-2) + \frac{2-a}{a+2}\right) &= 0 && \leq \xi_{(-2,-1)}
\end{aligned}$$

Where $1 - \frac{6-a}{2+a} = \frac{-4+2a}{a+2} \leq 0$ for $a \in (-1, 2)$.

b) For which value of $C$ is (1) uniformly better than (2), i.e. $\forall a \in (2, -1)$. For which value of $C$ is (2) uniformly better than (1)? **Hint:** Explicitly evaluating the intersections of functions is quite involved in this case. Consider using a plotting tool to compare the functions. Prove or argue whatever you find post hoc.

We will consider the intersections between (1) and (2). To this end consider (1)=(2)

$$\begin{aligned}
\frac{1}{8} + C\frac{2-a}{2} &= \frac{2}{(a+2)^2} \\
C &= \left(\frac{2}{(a+2)^2} - \frac{1}{8}\right)\frac{2}{2-a} \\
C &= \left(\frac{16-(a+2)^2}{8(a+2)^2}\right)\frac{2}{2-a} \\
C &= \left(\frac{(a+6)(2-a)}{8(a+2)^2}\right)\frac{2}{2-a} \\
C &= \frac{(a+6)}{4(a+2)^2}
\end{aligned}$$

We can find the relevant intersections by considering $a \nearrow 2$ and $a \searrow -1$.

$$\lim_{a \nearrow 2} \frac{(a+6)}{4(a+2)^2} = \frac{1}{8}$$

$$\lim_{a \searrow -1} \frac{(a+6)}{4(a+2)^2} = \frac{5}{4}$$

Lastly there is an intersection at $a = 2$ for any $C$.

$$\frac{1}{8} + C\frac{2-2}{2} = \frac{1}{8} = \frac{2}{(2+2)^2}.$$

This intersection is seperate, because in the derivation above we canceled out $2-a$ and it cannot be found as it is an intersection for any $C$.
For the choices of C above we can now figure out which function is uniformly better.
For $0 \leq C \leq \frac{1}{8}$ ($C < 0$ is not permitted) (1) is uniformly better than (2). $\frac{1}{8} + \frac{2-a}{8 \cdot 2} = \frac{4-a}{16} < \frac{2}{(a+2)^2}$ for $a \in (-1, 2)$.
For $C >= \frac{5}{4}$ (2) is better than (1). If $C > \frac{5}{4}$, $a$ has to be chosen smaller than $-1$ for there to be an intersection. Consider $a = 0$:

$$\frac{11}{8} = \frac{1}{8} + \frac{5}{4}\frac{2-0}{2} \geq \frac{2}{(0+2)^2} = 1,$$

as such (2) is smaller than (1) for $a \in (-1, 2)$.
For $\frac{1}{8} < C < \frac{5}{4}$ none of the two is better. This can be verifyed by finding the intersection for $a \in (-1, 2)$, however this is quite involved and will be omitted here. (This is not relevant for scoring points.)

c) Conclude how $C$ influences the optimization problem. Justify your conclusion using at most 5 sentences.

This might not be obvious, however this should have been figured out with the above questions. (1) and (2) are extremes of classifiers. (2) is the hard-margin SVM solution. (1) is the hard-margin SVM solution assuming $(a, 1)$ to be an outlier, as such (1) is the expected solution of the soft-margin SVM for specific $C$.
Using the above observations it is easy to conclude that the lower $C$ the more we strive for hard-margin SVM classification, we assume there are few outliers, or little jitter in the data. On the other hand as $C$ increases we move toward only considering the two most extreme points and choosing the hard-margin SVM that seperates those.
From this question you should follow that there is a cut-off point at which increasing $C$ does not impact the optimization anymore.

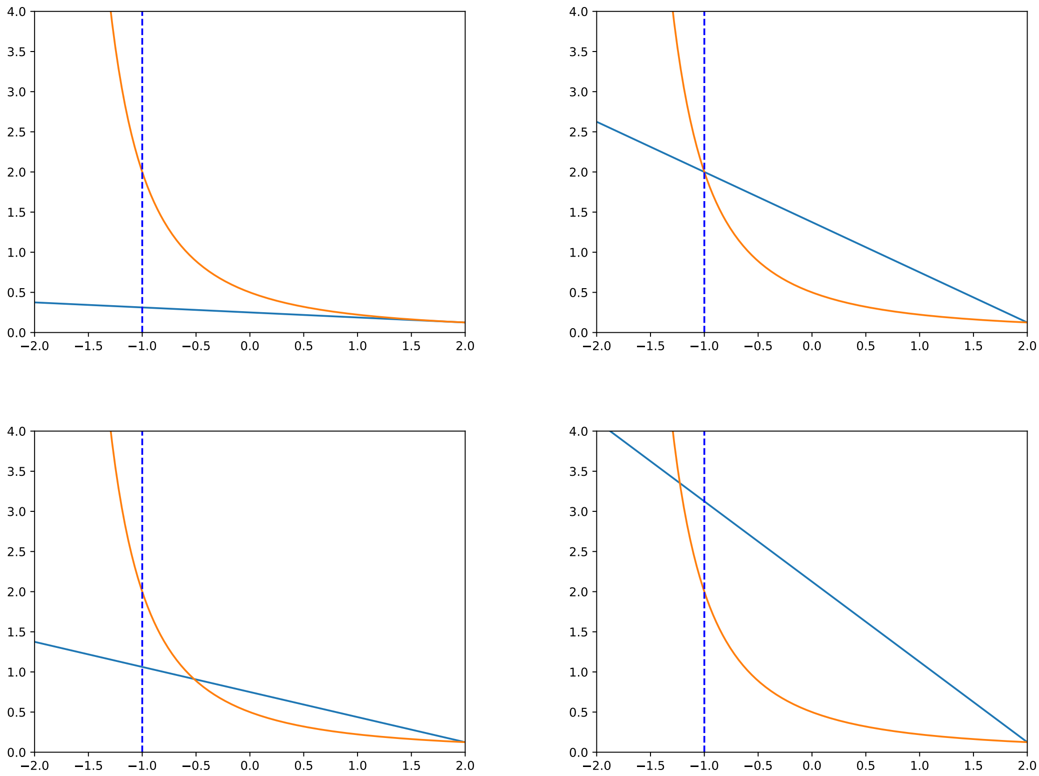


Figure 1: Shows the plots of (1), in orange, and (2), in blue, for differing $C$, $a = -1$ is marked with a blue dotted line. For $C = \frac{1}{8}$ (top-left). For $C = 1.25$, (2) is below (1) (top-right). $C = 0.5$(bottom-left). $C = 2$ (bottom-right).

**2)** Let $f(a_1, a_2, \cdots, a_n) = \ln(e^{a_1} + e^{a_2} + \cdots + e^{a_n})$. Show that $f$ is convex.

We starting calculating the $\nabla f$. Assume that $u = e^{a_1} + e^{a_2} + \cdots + e^{a_n}$

$$\frac{\partial f}{\partial a_i} = \frac{\partial \ln(u)}{\partial u}\frac{\partial u}{\partial a_i} = \frac{1}{u}e^{a_i} = \frac{e^{a_i}}{e^{a_1} + e^{a_2} + \cdots + e^{a_n}}$$

Now we need to find the hessians matrix $\mathbf{H}^f$. Observe that we have different patterns of derivatives: the elements in and outside of the diagonal. Therefore:

$$\begin{aligned}
\mathbf{H}^f_{(i,i)} = \frac{\partial f}{\partial a_i \partial a_i} &= \frac{\frac{\partial e^{a_i}}{\partial a_i}(e^{a_1} + e^{a_2} + \cdots + e^{a_n}) - e^{a_i}\frac{\partial e^{a_1 + e^{a_2} + \cdots + e^{a_n}}}{\partial a_i}}{(e^{a_1} + e^{a_2} + \cdots + e^{a_n})^2} \\
&= \frac{e^{a_i}(e^{a_1} + e^{a_2} + \cdots + e_{a_i} + \cdots + e^{a_n}) - e^{a_i}e^{a_i}}{(e^{a_1} + e^{a_2} + \cdots + e^{a_n})^2} \\
&= \frac{\sum_{k \neq i} e^{a_i}e^{a_k}}{(e^{a_1} + e^{a_2} + \cdots + e^{a_n})^2}
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{H}^f_{(i,j)} = \frac{\partial f}{\partial a_i \partial a_j} &= \frac{\frac{\partial e^{a_i}}{\partial a_j}(e^{a_1} + e^{a_2} + \cdots + e^{a_n}) - e^{a_i}\frac{\partial e^{a_1 + e^{a_2} + \cdots + e^{a_n}}}{\partial a_j}}{(e^{a_1} + e^{a_2} + \cdots + e^{a_n})^2} \\
&= \frac{0 \times (e^{a_1} + e^{a_2} + \cdots + e_{a_i} + \cdots + e^{a_n}) - e^{a_i}e^{a_j}}{(e^{a_1} + e^{a_2} + \cdots + e^{a_n})^2} \\
&= \frac{-e^{a_i}e^{a_j}}{(e^{a_1} + e^{a_2} + \cdots + e^{a_n})^2}
\end{aligned}$$

Summarizing:

$$\mathbf{H}^f_{(i,j)} = \frac{1}{(e^{a_1} + e^{a_2} + \cdots + e^{a_n})^2} \times \begin{cases} \sum_{l \neq i} e^{a_i}e^{a_l} & i = j \\ -e^{a_i}e^{a_j} & i \neq j \end{cases}$$

Now we need prove $\mathbf{x}^\top \mathbf{H}\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{x} \neq \mathbf{0}$. Consider $\mathbf{z} = \mathbf{x}^\top \mathbf{H}$, then:

$$\begin{aligned}
z_i &= \sum_k \mathbf{x}_k \times H^f_{(k,i)} \\
&= x_i H^f_{(i,i)} + \sum_{k \neq i} x_j \times H^f_{(k,i)} \\
&= \frac{x_i \sum_{k \neq i} e^{a_i}e^{a_k} + \sum_{j \neq i} -x_j e^{a_i}e^{a_j}}{(e^{a_1} + e^{a_2} + \cdots + e^{a_n})^2}.
\end{aligned}$$

Now we calculate $\mathbf{z}^T\mathbf{x} \geq 0$. Proceeding the multiplication:

$$\begin{aligned}
\mathbf{z}^T\mathbf{x} &= \sum_i z_i x_i \\
&= \sum_i \frac{{x_i}^2 \sum_{k\neq i} e^{a_i}e^{a_k} + x_i \times \sum_{j\neq i} -x_j e^{a_i}e^{a_j}}{(e^{a_1}+e^{a_2}+\cdots+e^{a_n})^2} \\
&= \frac{1}{(e^{a_1}+e^{a_2}+\cdots+e^{a_n})^2} \sum_i \left( {x_i}^2 \sum_{k\neq i} e^{a_i}e^{a_k} + \sum_{j\neq i} -x_i x_j e^{a_i}e^{a_j} \right).
\end{aligned}$$

Let's consider $e^{a_k}e^{a_l}$. Once $e^{a_k}e^{a_l} = e^{a_l}e^{a_k}$, from the first inner sum we have:

$$ {x_k}^2 \times e^{a_k}e^{a_l} + {x_l}^2 \times e^{a_l}e^{a_k} = e^{a_k}e^{a_l} \times ({x_k}^2 + {x_l}^2). $$

From the second inner sum:

$$ -x_k x_l \times e^{a_k}e^{a_l} - x_l x_k \times e^{a_l}e^{a_k} = -2x_k x_l e^{a_k}e^{a_l}. $$

Therefore:

$$\begin{aligned}
e^{a_k}e^{a_l} \times ({x_k}^2 + {x_l}^2) - 2x_k x_l e^{a_k}e^{a_l} &= e^{a_k}e^{a_l}({x_k}^2 - 2x_k x_l + {x_l}^2) \\
&= e^{a_k}e^{a_l}(x_k - x_l)^2
\end{aligned}$$

Finally

$$ \mathbf{x}^\top \mathbf{H}\mathbf{x} = \frac{1}{(e^{a_1}+e^{a_2}+\cdots+e^{a_n})^2} \sum_i^{n-1} \sum_{j=i+1}^{n} e^{a_i}e^{a_j}(x_i - x_j)^2 \geq 0. $$

**3)** Using any techniques you have learned from class, determine the domain on which the function $f(x, y) = xy^3$ is convex, i.e. for which combination of $x$ and $y$ is $f(x, y)$ convex?

Using basic calculus we see that the Hessian of $f$ is

$$H(x, y) = \begin{bmatrix} 0 & 3y^2 \\ 3y^2 & 6xy \end{bmatrix}.$$

From this we see that

$$\det(H) = -9y^4,$$

which is negative everywhere except where $y = 0$. Since the determinant is the product of the eigenvalues and since every symmetric matrix has real eigenvalues we see that there must be one positive and one negative eigenvalue whenever $y \neq 0$ and therefore the Hessian is not positive semidefinte when $y \neq 0$ and thus the function is not convex whenever $y \neq 0$. Thus there is no point which contains an open ball on which the function is convex, so the function is nowhere convex.

**4)** Solve programming task 2.