Prof. Dr. Jan Rolfes, Ehsan Waiezi, Lars Weidne Winter term 23/24

Mathematics of Learning – Worksheet 8 - – Discussion on Dec. 7/8th, 2023

- The exercise sheets will be uploaded every Monday. Solution sketches will be uploaded one week later.
- You can hand in your own solutions via StudOn and we correct them - this is not mandatory. Please hand in small groups of 2-3 students.
- For questions, please use the forum on StudOn since other students may have similar questions. If you have a more personal question about the exercises please send an email to ehsan.waiezi@fau.de or lars.weidner@fau.de respectively.

Basics [Hyperplanes].

We want to consider a few properties of sets of the kind

$$H := \{x \in \mathbb{R}^p : a^T x = b\}$$

for vectors $a \in \mathbb{R}^p$ and vectors $b \in \mathbb{R}$. These sets are called *hyperplanes*.
a) Prove: H is empty, if and only if $a = 0^p$ and $b \neq 0$.
b) Prove: H is convex, i.e., if $x, y \in H$, then for any real number $\lambda \in [0, 1]$, also $\lambda x + (1 - \lambda)y \in H$.
c) Prove: H is affine, i.e., if $x, y \in H$, then for any real numbers $\lambda_1, \lambda_2 \in \mathbb{R}$ with $\lambda_1 + \lambda_2 = 1$, also $\lambda_1 x + \lambda_2 y \in H$.

If we make a slight adaptations to the sets under consideration, i.e., we change $=$ by $\leq$, we get so-called halfspaces:

$$S := \{x \in \mathbb{R}^p : a^T x \leq b\}$$

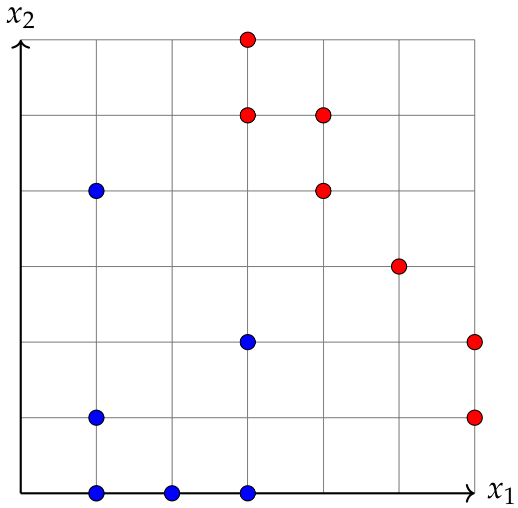
d) Prove: Halfspaces are convex, but (in general) not affine.

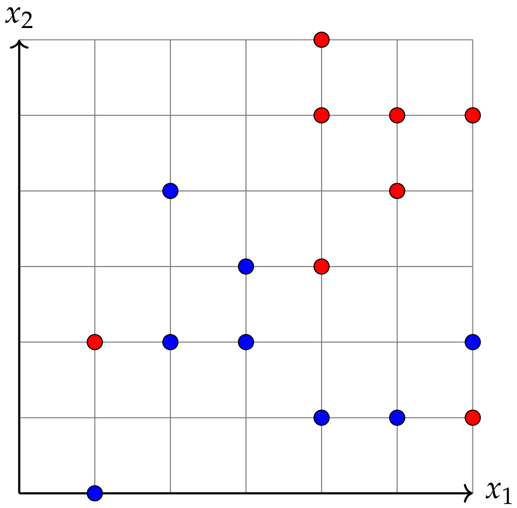
Solution.
a) Let H be empty. If $a = 0$ and $b = 0$, then H is obviously $\mathbb{R}^p$. If $a \neq 0$, then we have at least one component of a, wlog. $a_1, \neq 0$. Then, the vector $x := (\frac{b}{a_1}, 0, ..., 0)^T \in H$, since it holds that $a^T x = a_1 \frac{b}{a_1} + a_{2:p}^T (0, ..., 0)^T = b$. On the other side, let $a = 0$ and $b \neq 0$, then $a^T x = (0)^T x = 0 \neq b$.
b) Let $x, y \in H$, i.e., $a^T x = b$ and $a^T y = b$. We can check that
$a^T(\lambda x + (1 - \lambda)y) = \lambda(a^T x) + (1 - \lambda)a^T y = \lambda b + (1 - \lambda)b = b.$
c) This is analogous to b) (we note that $\lambda_1 + \lambda_2 = 1$ suffices, we do not need $\lambda \in [0, 1]$).
d) Half spaces are convex: Let $x, y \in S$, i.e., $a^T x \leq b$ and $a^T y \leq b$. We can check that
$a^T(\lambda x + (1 - \lambda)y) = \lambda(a^T x) + (1 - \lambda)a^T y \leq \lambda b + (1 - \lambda)b = b.$
The $\leq$ works because both λ and $(1 - \lambda)$ are nonnegative.

Half spaces are not affine: Consider for example the half space $\{x \in \mathbb{R} : -x \leq 0\}$. We check for the elements 1 and 0 in the half space, that $-1 \cdot 1 + 2 \cdot 0$ is not in the half space.

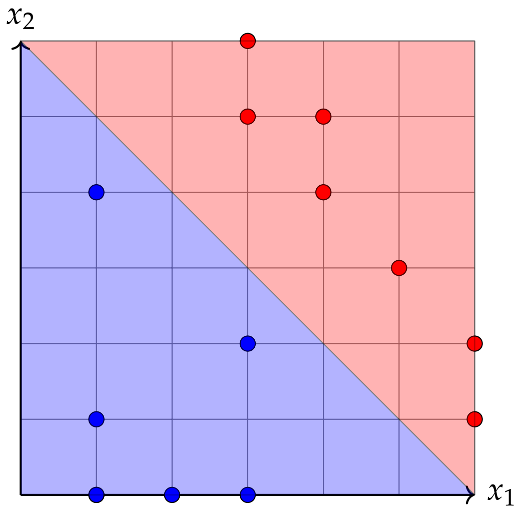
Exercise 1 [Find vectors for 2d point classification].

Consider these two pictures. Find lines which separate the two classes (red points and blue points) as good as possible for every figure. You are allowed to be creative, but please explain your understanding of "as good as possible". Write down the definition of your lines formally correct as hyperplanes.

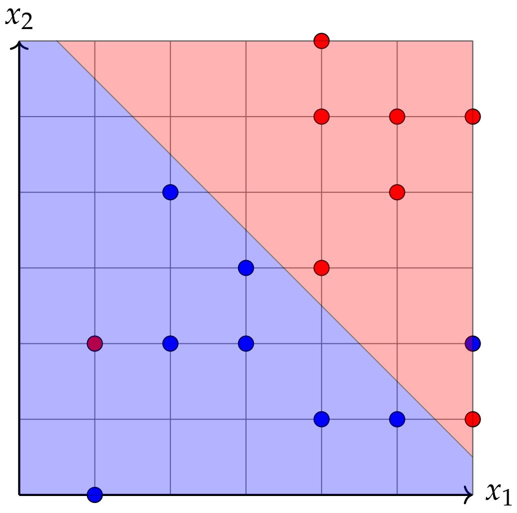




Solution. Lets be modest at the beginning, and lets define "as good as possible" as "as few as possible fail classifications".



$x_1 + x_2 = 6$



$x_1 + x_2 = 6.5$

Alternatively, we could for example maximize the difference between the border hyperplane and the correctly classified points, i.e., $y_i(a^T x_i - b) > 0$. This would lead to the optimization problem

$$\max \sum_{i=1}^{n} y_i(a^T x_i - b)$$
$$\text{s.t. } ||a|| = 1.$$

This problem is (after a short sequence of clever thoughts) defined on a one dimensional subset of $\mathbb{R}^p$ and the objective function is Lipschitz continuous (the constant depends on the data), hence it can be solved rather efficiently up to a certain accuracy.

Exercise 2 [The path to support vector machines].

Let $\{(x_1, y_1), ..., (x_n, y_n)\} \subset \mathbb{R}^p \times \{-1, 1\}$ be n pairs of labeled data (we exclude the border case where only one label exists since it is not interesting) within $\mathbb{R}^p$. Consider the following two optimization problems:

$$\begin{aligned} \max_{\beta\in\mathbb{R}^p,\beta_0\in\mathbb{R},M\in\mathbb{R}} \quad & M \\ \text{s.t. } y_i(x_i^T\beta + \beta_0) \geq M \quad & \text{for all } i = 1, ..., n \\ ||\beta|| = 1 & \end{aligned} \qquad \begin{aligned} \min_{\beta\in\mathbb{R}^p,\beta_0\in\mathbb{R}} \quad & ||\beta||^{(2)} \\ \text{s.t. } y_i(x_i^T\beta + \beta_0) \geq 1 \quad & \text{for all } i = 1, ..., n \end{aligned}$$

a) Try to describe in words what the optimization problems do.

b) Prove or disprove: Whenever $\hat{\beta}, \hat{\beta}_0$ solves the right problem, then $\frac{\hat{\beta}}{||\hat{\beta}||}, \frac{\hat{\beta}_0}{||\hat{\beta}||}$ solves the left problem.

c) Interpret the points of the previous task as labeled input data ($p = 2$, $n = 13/16$, red is 1 and blue is -1). Try to solve the left as well as the right optimization problem for both data sets (*Remark: This looks quite simple, but it is not, since you have to solve a constrained quadratic optimization problem. There are two ways to go which I can see: 1. set up the KKT-System and try to solve it explicitly, maybe with the help of some nonlinear equation system solver, 2. Try to get in touch with optimization software: search e.g., for pyomo, neos server. Try to do either 1 or 2.*).

Solution.

a) The optimization problems are a simplified version of the SVM problem which has been part of the lecture. The left one maximizes a variable M, which is the space between a separating hyperplane (which is defined by $\beta^T x + \beta_0 = 0$) and a certain data point set. The first class (1) of points should be on the one side of the plane, the second (class -1) on the other. Note that M can be chosen negative if the points cannot be separated perfectly as in the second data set in Exercise 1. In this case the hyperplane should be as close as possible to misclassified points

The right optimization problem does the same, with the slight difference, that the space between hyperplane and point sets now is measured in "how large is the norm of β". Therefore, the right optimization problem gets infeasible if the two classes cannot be separated without misclassification. This is also the reason why exercise b) is not an equivalence proof.

b) Let $(\hat{\beta}, \hat{\beta}_0)$ be an optimal solution to the right problem. We note that β is not the 0-vector, whenever there are points in each class. Then, $\frac{\hat{\beta}}{||\hat{\beta}||}, \frac{\hat{\beta}_0}{||\hat{\beta}||}$ is feasible for the left problem, since we know that the norm of $\frac{\hat{\beta}}{||\hat{\beta}||}$ is 1 (dividing a vector through its own norm results in a vector of norm 1). It remains to set M to a value as large as possible. Since we know that $\hat{\beta}, \hat{\beta}_0$ is feasible for the right problem, i.e.,

$$y_i(x_i^T\hat{\beta} + \hat{\beta}_0) \geq 1 \quad \text{for all } i,$$

we can (dividing both sides by $||\hat{\beta}||$) deduce that

$$y_i(x_i^T \frac{\hat{\beta}}{||\hat{\beta}||} + \frac{\hat{\beta}_0}{||\hat{\beta}||}) \geq \frac{1}{||\hat{\beta}||} \quad \text{for all } i$$

holds, i.e., we can set $\hat{M} := \frac{1}{||\hat{\beta}||}$ (and therefore $||\hat{\beta}|| = \frac{1}{\hat{M}}$).
Assume for contradiction that there is $\bar{\beta}, \bar{\beta}_0, \bar{M}$ solving left with $\bar{M} > \hat{M}$ (implying $||\bar{\beta}|| = 1$). We observe that $\frac{\bar{\beta}}{\bar{M}}, \frac{\bar{\beta}_0}{\bar{M}}$ solve the right problem. However, the norm of $\frac{\bar{\beta}}{\bar{M}}$ would be $\frac{1}{\bar{M}} < \frac{1}{\hat{M}}$, which is a contradiction to the initial assumption that $\hat{\beta}, \hat{\beta}_0$ are optimal.
c) See also the implementation files. We get as a solution the following values:

- Data set one
  - Left:
    $\beta = \begin{pmatrix} 0.8 \\ 0.6 \end{pmatrix}$, $\beta_0 = -4.5$, $M = 0.9$.
  - Right:
    $\beta = \begin{pmatrix} 0.889 \\ 0.667 \end{pmatrix}$, $\beta_0 = -5.0$, $\frac{1}{||\beta||} = 0.9$.
- Data set two
  - Left:
    $\beta = \begin{pmatrix} 0.196 \\ 0.981 \end{pmatrix}$, $\beta_0 = -3.24$, $M = -1.08$.
  - Right: infeasible.