ROB 101 - Fall 2021

Hyperplanes in $\mathbb{R}^n$, Quadratic Program, and Maximum Margin Classifier

November 29, 2021

UNIVERSITY OF MICHIGAN

Learning Objectives

- Introduce material that is assumed in UofM Computer Science courses that have Math 214 as a prerequisite.

- Provide a resource for use after you leave ROB 101.

- Learn how to separate $\mathbb{R}^n$ into two halves via hyperplanes.
- The notion of signed distance to a hyperplane.
- An example of a max-margin classifier, a common tool in Machine Learning.

We want to study linear structures than can be used to divide $\mathbb{R}^n$ into pieces.

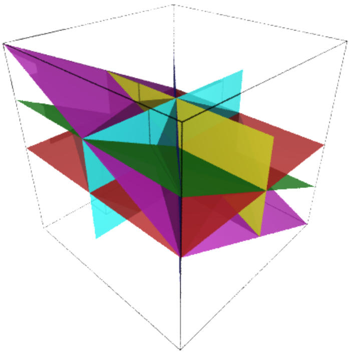
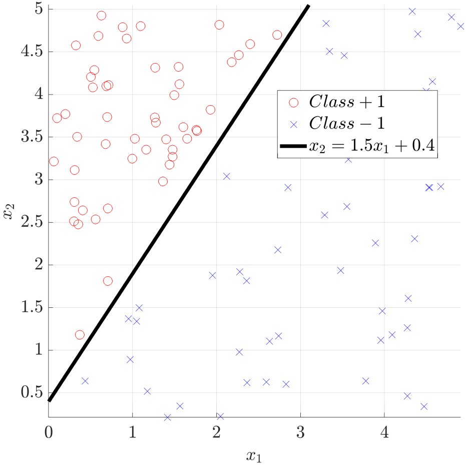
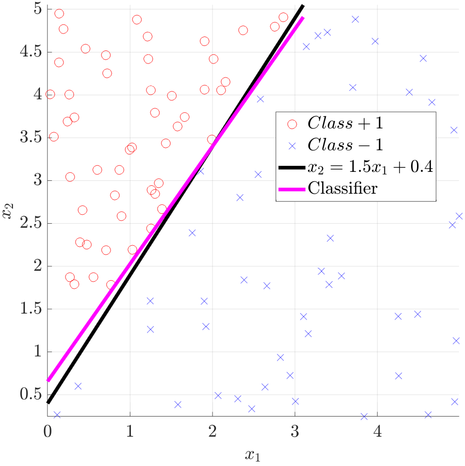


Figure: Dividing $\mathbb{R}^3$ into disjoint regions. Image from Wikimedia Commons.

Example: Classification

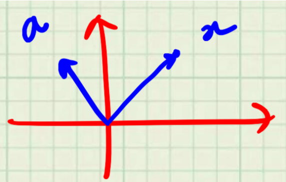


Example: Classification



- Consider the set of all points, $x \in \mathbb{R}^2$ such that
$$\langle a, x \rangle = 0, \quad a \in \mathbb{R}^2.$$

- $a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, and $\langle a, x \rangle = a_1 x_1 + a_2 x_2$.



Lines in $\mathbb{R}^2$ as Separating Hyperplanes

- Consider the set of all points, $x \in \mathbb{R}^2$ such that

$$\langle a, x \rangle = 0, \quad a \in \mathbb{R}^2.$$

- Writing it in the set notation:

$$L = \{x \in \mathbb{R}^2 \mid \langle a, x \rangle = 0,\ a \in \mathbb{R}^2\}.$$

- L is a line that passes through the origin.

- Consider the set of all points, $x \in \mathbb{R}^n$ such that

$$\langle a, x \rangle = 0, \quad a \in \mathbb{R}^n.$$

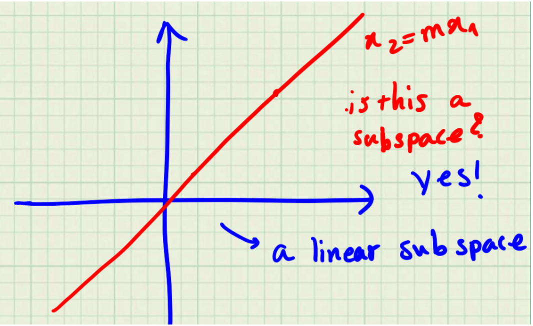
- Writing it in the set notation:

$$H = \{x \in \mathbb{R}^n \mid \langle a, x \rangle = 0,\ a \in \mathbb{R}^n\}.$$

- H is a hyperplane that passes through the origin.

- $H = \{x \in \mathbb{R}^n \mid \langle a, x\rangle = 0,\ a \in \mathbb{R}^n\}$.

- $\langle a, x\rangle = 0 \iff a$ (normal vector) is orthogonal to all vectors that lie on the hyperplane.



- We can divide $\mathbb{R}^2$ into two halves.
- Indeed, we define the following half-planes.

$$H^+ := \{x \in \mathbb{R}^2 \mid \langle a, x \rangle > 0\},$$
$$H^- := \{x \in \mathbb{R}^2 \mid \langle a, x \rangle < 0\}.$$

Remark

Using the angle between a *and* x, $\cos\theta = \frac{\langle a,x \rangle}{\|a\| \cdot \|x\|}$, *we see that* $\langle a, x \rangle > 0$ *is the side* (H^+) *where the angle is less than* $90\,\mathrm{deg}$, *and* $\langle a, x \rangle < 0$ *is the side* (H^-) *where the angle is greater than* $90\,\mathrm{deg}$.

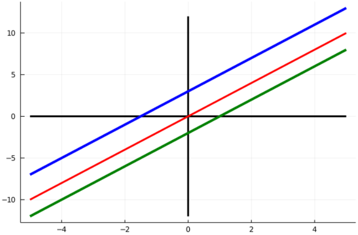
The same observation holds in $\mathbb{R}^n$.

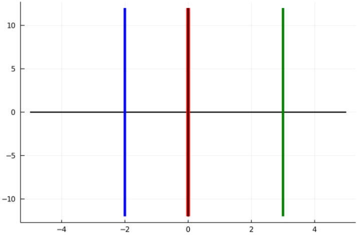
- We can divide $\mathbb{R}^n$ into two halves.
- Indeed, we define the following half-planes.

$$H^+ := \{x \in \mathbb{R}^n \mid \langle a, x \rangle > 0\},$$
$$H^- := \{x \in \mathbb{R}^n \mid \langle a, x \rangle < 0\}.$$

If we take a subspace and translate it by x_c, then we get an affine subspace.

- $M = x_c + H = \{x \in \mathbb{R}^n \mid \langle a, x - x_c \rangle = 0,\ a, x_c \in \mathbb{R}^n\}$.





If we take a subspace and translate it by x_c, then we get an affine subspace.

- $M = x_c + H = \{x \in \mathbb{R}^n \mid \langle a, x - x_c \rangle = 0,\ a, x_c \in \mathbb{R}^n\}$.

Remark

Previously, we had $x_c = 0$. If $x_c \neq 0$, then the vector $x - x_c$ lie on the hyperplane. Hence, $\langle a, x - x_c \rangle = 0$.

We can look into the translated subspace as follows.

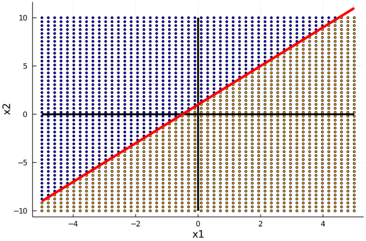
$$\langle a, x - x_c \rangle = a^{\mathsf{T}}(x - x_c) = 0$$
$$a^{\mathsf{T}} x = a^{\mathsf{T}} x_c =: d, \quad d \in \mathbb{R}.$$

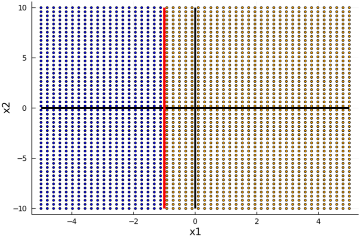
Remark

If a is normalized, then $d = a^{\mathsf{T}} x_c$ is the distance from the origin to the hyperplane, i.e., the length of a vector parallel to a that starts from the origin and ends at the hyperplane.

We can divide $\mathbb{R}^n$ into two halves.

$$H^+ := \{x \in \mathbb{R}^n \mid \langle a, x - x_c \rangle > 0\},$$
$$H^- := \{x \in \mathbb{R}^n \mid \langle a, x - x_c \rangle < 0\}.$$





- For any point that does not lie on the hyperplane, we have $\langle a, x - x_c \rangle \neq 0$ or $a^\mathsf{T} x \neq a^\mathsf{T} x_c$.

- We define the signed distance of a point to the hyperplane by the amount of deviation from the hyperplane equation.

$$y(x) = \frac{\langle a, x - x_c \rangle}{\|a\|}$$

- We normalize by $\|a\|$ to avoid scaling the space.

Supervised machine learning can be divided into two categories:

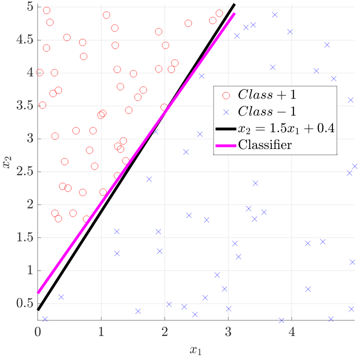
1. Regression; in this case, the outputs (also called target values) are continuous (real numbers).

2. Classification; in this case, the outputs (targets) are discrete categories (called labels).

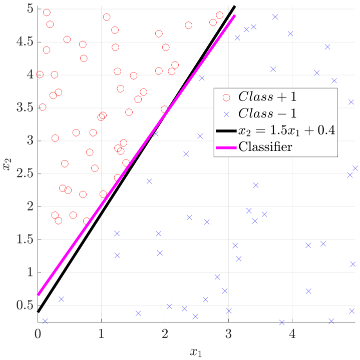
You have seen regression problems in ROB 101. In this example, we will formulate a classification problem.

- We wish to find a classifier (here a hyperplane) that separates $\times$ and $\circ$ categories.

- Furthermore, we want to predict the label for a new input (called query or test point).



- We wish to find a classifier (here a hyperplane) that separates $\times$ and $\circ$ categories.
- Furthermore, we want to predict the label for a new input (called query or test point).



- We are given a data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where the inputs are $x_i \in \mathbb{R}^2$ and targets are $y_i \in \{+1, -1\}$.
- Our model is a hyperplane $a^\mathsf{T} x + a_0 = 0$. a_0 is called the bias term.
- Define $w := \begin{bmatrix} a \\ a_0 \end{bmatrix}$ and $\bar{x} := \begin{bmatrix} x \\ 1 \end{bmatrix}$. Then $w^\mathsf{T} \bar{x} = 0$.

We define the following *hard margins*.

- $w^\mathsf{T}\bar{x} = 1$, anything on or above this boundary belongs to class $+1$.
- $w^\mathsf{T}\bar{x} = -1$, anything on or below this boundary belongs to class -1.
- We get the following constraints:

$$
\begin{aligned}
w^\mathsf{T}\bar{x}_i &\geq 1, \quad \text{if } y_i = 1, \\
w^\mathsf{T}\bar{x}_i &\leq -1, \quad \text{if } y_i = -1.
\end{aligned}
$$

- We get the following constraints:

$$w^\mathsf{T}\bar{x}_i \geq 1, \quad \text{if } y_i = 1,$$
$$w^\mathsf{T}\bar{x}_i \leq -1, \quad \text{if } y_i = -1.$$

- We can combine both constraints into one as

$$y_i \cdot w^\mathsf{T}\bar{x}_i \geq 1, \quad \text{for } i = 1, \ldots, n.$$

We now formulate the following *constrained* optimization problem.

$$
\begin{aligned}
\min_{w \in \mathbb{R}^3} \quad & \frac{1}{2} w^\mathsf{T} w \\
\text{subject to} \quad & y_i \cdot w^\mathsf{T} \bar{x}_i \geq 1, \quad \text{for } i = 1, \ldots, n.
\end{aligned}
$$

Remark

Training is the process of finding (called estimating or learning depending on the context) "optimal" $w^\star$.

Remark

Testing is the process of evaluating the trained model for a new input (an example that was not seen before).

Given a query point (new input) x_*, we can evaluate the signed distance and pass it through the sign function. This is called the decision or response function.

$$y_* = \operatorname{sgn}({w^\star}^\mathsf{T} \bar{x}_*),$$

where sgn is the sign function.

$$\operatorname{sgn}(x) = \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

A *Quadratic Program* (QP) is a special kind of optimization problem with *constraints*. The cost to be minimized is supposed to be quadratic, meaning that $f : \mathbb{R}^m \to \mathbb{R}$ has the form

$$f(x) = \frac{1}{2} x^\mathsf{T} Q x + q x,$$

where Q is an $m \times m$ symmetric matrix, meaning that $Q^\mathsf{T} = Q$, and where q is a $1 \times m$ row vector.

We consider the QP

$$
\begin{aligned}
x^* = \quad & \underset{x \in \mathbb{R}^m}{\arg\min} \quad \frac{1}{2} x^\mathsf{T} Q x + q x \\
& A_{in} x \preceq b_{in} \\
& A_{eq} x = b_{eq} \\
& lb \preceq \ x \preceq ub
\end{aligned}
$$

and assume that Q is symmetric ($Q^\mathsf{T} = Q$) and *positive definite* ($x \neq 0 \implies x^\mathsf{T} Q x > 0$), and that the subset of $\mathbb{R}^m$ defined by the constraints is non empty, that is

$$C := \{x \in \mathbb{R}^m \mid A_{in} x \preceq b_{in}, A_{eq} x = b_{eq}, lb \preceq x \preceq ub\} \neq \emptyset.$$

Then x^* exists and is unique.

Example

Let's switch to the Julia notebook.

Next Time

- Soft Margin Classifier, Gaussian Support Vector Machine
- Read Chapter 13 of ROB 101 Book. QP is in Chapter 12.