Linear separation and feature maps

DATA 607 — Session 5 — 11/03/2019

Lines in $\mathbb{R}^2$: Direction vectors

A line in $\mathbb{R}^2$ is a set of points of the form

$$L_{\boldsymbol{u},\boldsymbol{v}} := \{\boldsymbol{u} + t\boldsymbol{v} : t \in \mathbb{R}\},$$

where

$$\boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \quad \boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \in \mathbb{R}^2, \quad \boldsymbol{v} \neq \boldsymbol{0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Note that $\boldsymbol{u} = \boldsymbol{u} + 0\boldsymbol{v} \in L_{\boldsymbol{u},\boldsymbol{v}}$.

$L_{\boldsymbol{u},\boldsymbol{v}}$ is called the line through $\boldsymbol{u}$ with direction vector $\boldsymbol{v}$.

- $\boldsymbol{u}' \in L_{\boldsymbol{u},\boldsymbol{v}} \Longrightarrow L_{\boldsymbol{u}',\boldsymbol{v}} = L_{\boldsymbol{u}',\boldsymbol{v}}$
- $\boldsymbol{v}' = c\boldsymbol{v},\ c \in \mathbb{R},\ c \neq 0 \Longrightarrow L_{\boldsymbol{u},\boldsymbol{v}'} = L_{\boldsymbol{u},\boldsymbol{v}}$

Lines in $\mathbb{R}^2$: Half-planes; sides; normal vectors

Dot product:

$$\boldsymbol{u} \cdot \boldsymbol{v} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = u_1 v_1 + u_2 v_2 \in \mathbb{R}$$

$\boldsymbol{u}$ and $\boldsymbol{v}$ are orthogonal or perpendicular.

$\boldsymbol{w}$ is a normal vector to $L_{\boldsymbol{u},\boldsymbol{v}}$ if $\boldsymbol{v} \cdot \boldsymbol{w} = 0$.

Lines in $\mathbb{R}^2$: Half-planes and sides

If you delete a line, L, from $\mathbb{R}^2$, you're left with two half-planes called the sides of L.

If $\boldsymbol{w}$ is a nonzero normal vector to $L_{\boldsymbol{u},\boldsymbol{v}}$, then the sets

$$H^{-}_{\boldsymbol{u},\boldsymbol{w}} = \{\boldsymbol{x} : \boldsymbol{w} \cdot (\boldsymbol{x} - \boldsymbol{u}) < 0\} \quad \text{and} \quad H^{+}_{\boldsymbol{u},\boldsymbol{w}} = \{\boldsymbol{x} : \boldsymbol{w} \cdot (\boldsymbol{x} - \boldsymbol{u}) > 0\}$$

are the sides of $L_{\boldsymbol{u},\boldsymbol{v}}$.

The normal vector $\boldsymbol{w}$, plotted with its tail on $L_{\boldsymbol{u},\boldsymbol{v}}$, points into $H^{+}_{\boldsymbol{u},\boldsymbol{w}}$.

Linear separation

Let

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$$

be a dataset, where $\boldsymbol{x}_i \in \mathbb{R}^2$ and $y_i \in \{-1, 1\}$.

Let

$$D^- = \{(\boldsymbol{x}_i, y_i) \in D : y_i = -1\}, \quad D^+ = \{(\boldsymbol{x}_i, y_i) \in D : y_i = +1\}$$

Let L be a line in $\mathbb{R}^2$. We say that $\underline{L \text{ separates } D}$ if D^- and D^+ are contained in opposite sides of L.

We say that D is <u>linearly separable</u> if there is a line L that separates D.

Not all datasets are linearly separable: The dataset

$$D := \Big\{ \big((0,0),0\big), \big((1,0),1\big), \big((1,1),0\big), \big((0,1),1\big) \Big\}$$

is not linearly separable.

Linear separators need not be unique: The linear separators of

$$D := \Big\{ \big((0,-1),0\big), \big((0,1),1\big) \Big\}$$

are precisely the lines

$$y = mx + b, \quad b \in (-1,1).$$

Finding linear separators

Problem: Given a dataset, D, find a vector $\boldsymbol{u}$ and a nonzero vector $\boldsymbol{w}$ such that

$$D \cap H^+_{\boldsymbol{u},\boldsymbol{w}} = D^+ \quad \text{and} \quad D \cap H^-_{\boldsymbol{u},\boldsymbol{w}} = D^-$$

or show that no such $\boldsymbol{u}$ and $\boldsymbol{w}$ exist.

We'll begin by analyzing a special case:

Special case: Given a dataset D, find a nonzero vector $\boldsymbol{w}$ such that

$$D \cap H^+_{\mathbf{0},\boldsymbol{w}} = D^+ \quad \text{and} \quad D \cap H^-_{\mathbf{0},\boldsymbol{w}} = D^-$$

$L_{\mathbf{0},\boldsymbol{w}}$ does **not** separate D if and only if

$$D^- \cap H^+_{\mathbf{0},\boldsymbol{w}} \neq \varnothing \quad \text{or} \quad D^+ \cap H^-_{\mathbf{0},\boldsymbol{w}} \neq \varnothing,$$

...

...or, equivalently, if and only if

$$\Big(y_i = -1 \quad \text{and} \quad \boldsymbol{w} \cdot \boldsymbol{x}_i > 0\Big) \quad \text{or} \quad \Big(y_i = +1 \quad \text{and} \quad \boldsymbol{w} \cdot \boldsymbol{x}_i < 0\Big)$$

for some i, or, equivalently, if and only if

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i) < 0$$

for some i, or, equivalently, if and only if

$$\min(y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i), 0) < 0$$

for some i, or, equivalently,

$$\sum_i \min(y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i), 0) < 0,$$

or, equivalently, if and only if

$$\sum_i \max(-y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i), 0) > 0.$$

View the term

$$L(\boldsymbol{w}, \boldsymbol{x}_i, y_i) := \max(-y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i), 0)$$

as a **penalty** or **loss** for $\boldsymbol{x}_i$ being misclassified by $\boldsymbol{w}$, i.e., lying on the wrong side of the line through $\mathbf{0}$ normal to $\boldsymbol{w}$.

Assume $\boldsymbol{x}_i$ is correctly classified, then

$$\Big(y_i = -1 \quad \text{and} \quad \boldsymbol{w} \cdot \boldsymbol{x}_i < 0\Big) \quad \text{or} \quad \Big(y_i = +1 \quad \text{and} \quad \boldsymbol{w} \cdot \boldsymbol{x}_i > 0\Big),$$

in which case $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i) > 0$ and

$$-y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i) < 0.$$

Thus, the penalty assessed for $\boldsymbol{x}_i$ being misclassified by $\boldsymbol{w}$ is

$$L(\boldsymbol{w}, \boldsymbol{x}_i, y_i) = \max(-y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i), 0) = 0,$$

appropriate since, by hypothesis, $\boldsymbol{x}_i$ is classified correctly!

Conversely, Assume $\boldsymbol{x}_i$ is correctly classified, then

$$\Big(y_i = -1 \quad \text{and} \quad \boldsymbol{w} \cdot \boldsymbol{x}_i > 0\Big) \quad \text{or} \quad \Big(y_i = +1 \quad \text{and} \quad \boldsymbol{w} \cdot \boldsymbol{x}_i < 0\Big),$$

in which case $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i) < 0$ and

$$-y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i) > 0.$$

Thus, the penalty assessed for $\boldsymbol{x}_i$ being misclassified by $\boldsymbol{w}$ is strictly positive:

$$L(\boldsymbol{w}, \boldsymbol{x}_i, y_i) = \max(-y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i), 0) > -y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i).$$

Define the **cost** associated with $\boldsymbol{w}$ by

$$C(D, \boldsymbol{w}) = \sum_i L(\boldsymbol{w}, \boldsymbol{x}_i, y_i) = \sum_i \max(-y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i), 0).$$

Then $L(\boldsymbol{w})$ separates D if and only if

$$C(D, \boldsymbol{w}) = 0.$$

General case

$$L(\boldsymbol{u}, \boldsymbol{w}, \boldsymbol{x}_i, y_i) = \sum_i \max(-y_i(\boldsymbol{w} \cdot (\boldsymbol{x}_i - \boldsymbol{u})), 0)$$

$$\begin{aligned} C(D, \boldsymbol{u}, \boldsymbol{w}) &= \sum_i L(\boldsymbol{u}, \boldsymbol{w}, \boldsymbol{x}_i, y_i) \\ &= \sum_i \max(-y_i(\boldsymbol{w} \cdot (\boldsymbol{x}_i - \boldsymbol{u})), 0) \end{aligned}$$

Find

$$\operatorname*{argmin}_{\boldsymbol{u}, \boldsymbol{w}} C(D, \boldsymbol{u}, \boldsymbol{w}).$$

The Perceptron Algorithm

$$D = \left\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\right\}$$

Suppose that D is linearly separable, i.e., that there is unit vector a $\boldsymbol{w} \in \mathbb{R}^n$ such that

$$\operatorname{sign}(\boldsymbol{w} \cdot \boldsymbol{x}_i) = y_i$$

for all i, or, equivalently, that

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i) > 0$$

for all i.

Define a sequence $\boldsymbol{w}_0, \boldsymbol{w}_1, \ldots$ of vectors in $\mathbb{R}^n$ by:

Algorithm The Perceptron Algorithm

$k \leftarrow 0$
$\boldsymbol{w}_0 \leftarrow \mathbf{0}$
while $y_i(\boldsymbol{w}_k \cdot \boldsymbol{x}_i) \leq 0$ for some i **do**
    $k \leftarrow k + 1$
    $i_k \leftarrow \min\{i : y_i(\boldsymbol{w}_{k-1} \cdot \boldsymbol{x}_i) \leq 0\}$
    $\boldsymbol{w}_k \leftarrow \boldsymbol{w}_{k-1} + y_{i_k} \boldsymbol{x}_{i_k}$
end while

Note that $i_1 = 1$.

Theorem: (Rosenblatt, 1957) The perceptron algorithm terminates after $k < R^2/r^2$ steps.

Proof: Let k be such that $y_{i_k}(\boldsymbol{w}_{k-1} \cdot \boldsymbol{x}_{i_k}) \leq 0$.

Lower bound on $\|\boldsymbol{w}_k\|^2$:

Set

$$R = \max_i \|\boldsymbol{x}_i\| > 0.$$

$$\begin{aligned}
\|\boldsymbol{w}_k\|^2 &= \|\boldsymbol{w}_{k-1} + y_{i_k}\boldsymbol{x}_{i_k}\|^2 \\
&= \|\boldsymbol{w}_{k-1}\|^2 + \|\boldsymbol{x}_{i_k}\|^2 + 2y_{i_k}(\boldsymbol{w}_{k-1} \cdot \boldsymbol{x}_{i_k}) \\
&\geq \|\boldsymbol{w}_{k-1}\|^2 + \|\boldsymbol{x}_{i_k}\|^2 \\
&\geq \|\boldsymbol{w}_{k-1}\|^2 + R^2 \\
&\geq \|\boldsymbol{w}_{k-2}\|^2 + 2R^2 \\
&\vdots \\
&\geq \|\boldsymbol{w}_0\|^2 + kR^2 \\
&= kR^2
\end{aligned}$$

Upper bound on $\|\boldsymbol{w}_k\|^2$:

Set

$$r = \min_i |\boldsymbol{w} \cdot \boldsymbol{x}_i| > 0.$$

$$\begin{aligned}
\boldsymbol{u} \cdot \boldsymbol{w}_k &= \boldsymbol{u} \cdot (\boldsymbol{w}_{k-1} + y_{i_k} \boldsymbol{x}_{i_k}) \\
&= \boldsymbol{u} \cdot \boldsymbol{w}_{k-1} + y_{i_k} (\boldsymbol{u} \cdot \boldsymbol{x}_{i_k}) \\
&\geq \boldsymbol{u} \cdot \boldsymbol{w}_{k-1} + r \\
&\geq \boldsymbol{u} \cdot \boldsymbol{w}_{k-2} + 2r \\
&\vdots \\
&\geq \boldsymbol{u} \cdot \boldsymbol{w}_0 + kr \\
&= kr
\end{aligned}$$

$$kr \leq \boldsymbol{u} \cdot \boldsymbol{w}_k \leq \|\boldsymbol{u}\| \|\boldsymbol{w}_k\| = \|\boldsymbol{w}_k\|$$

$$k^2 r^2 \leq \|\boldsymbol{w}_k\|^2$$

$$k^2 r^2 \leq \|\boldsymbol{w}_k\|^2 \leq kR^2$$

$$kr^2 \leq \|\boldsymbol{w}_k\|^2 \leq R^2$$

$$k \leq \frac{R^2}{r^2}$$

Feature maps

What do we do if our data isn't linearly separable?

Embed your data in a higher dimensional space in which it is linearly separable.

The higher dimensional space is called **feature space** and the function mapping **data space** — the ambient space of our data — into this feature space is called a **feature map**.

Consider the linearly inseparable dataset

$$D := \Big\{ \big((0,0),0\big), \big((1,0),1\big), \big((1,1),0\big), \big((0,1),1\big) \Big\}$$

Define a **feature map** $\phi : \mathbb{R}^2 \to \mathbb{R}^3$ by

$$\phi(x_1, x_2) = \big(x_1, x_2, (x_1 - x_2)^2\big)$$

Then

$$\phi(D) = \Big\{ \big((0,0,0),0\big), \big((1,0,1),1\big), \big((1,1,0),0\big), \big((0,1,1),1\big) \Big\}$$

The points in with class label 0 and 1 have third coordinates 0 and 1, respectively.

The plane

$$x_3 = \frac{1}{2}$$

separates the classes.

The points in data space mapped by ϕ into this separating plane are those that satisfy

$$(x_1 - x_2)^2 = \frac{1}{2}$$

$$x_1 - x_2 = \pm\frac{1}{2}$$

This pair of lines separates the original dataset D.

More examples in the Jupyter notebook.

Where do features come from?

Many classification techniques require features to be hand-crafted for a given application.

A more robust approach is to, as much as possible, *learn* the features. Neural networks use this approach.