2.1 Linear Classifiers

Machine Learning 1: Foundations

Marius Kloft *(TUK)*

28 Apr – 5 May 2020

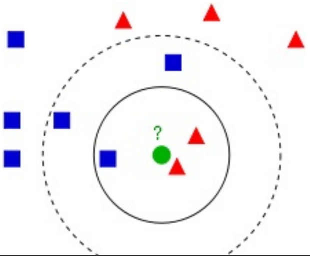
Recap

Machine learning

- computers learning from data
  how to make accurate predictions

Formal problem setting and terminology

- Given **training data** =
  - **inputs** $\mathbf{x}_1, \ldots, \mathbf{x}_n$
  - **labels** $y_1, \ldots, y_n$
- Aim: to compute a function f (called **classifier** or **predictor**) predicting the unknown label y of a new input $\mathbf{x}$)
- Example: k-nearest neighbor algorithm



Contents of this Video

Linear Classifiers

1. Math Notation & Recap
2. Linear Classifiers

Math Notation

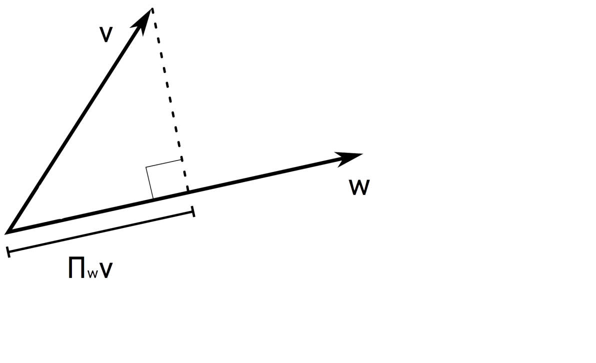
- Vectors $\mathbf{v} \in \mathbb{R}^d$ are thought of as column vectors and denoted with boldface letters.
- Scalars $s \in \mathbb{R}$ are denoted with normal letters.
- Matrices $M \in \mathbb{R}^{m \times n}$ have m rows and n columns and are denoted with normal letters.
- Greek letters can refer to both scalars and vectors, but they are boldfaced if they denote vectors ($\boldsymbol{\lambda} \in \mathbb{R}^d$ vs. $\lambda \in \mathbb{R}$).
- $\mathbf{0}$ and $\mathbf{1}$ are vectors in $\mathbb{R}^d$ with entries all zeros and ones, respectively.
- Transposition of a vector or matrix:
  - if $\mathbf{v}$ is a column vector, then $v^\top$ is a row vector
  - if $M \in \mathbb{R}^{m \times n}$ then $M^\top \in \mathbb{R}^{n \times m}$.
- Scalar product of two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$: $\quad \langle \mathbf{v}, \mathbf{w} \rangle := \mathbf{v}^\top \mathbf{w}$
- Norm of a vector: $\|\mathbf{v}\| := \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} = \sqrt{\mathbf{v}^\top \mathbf{v}}$
- $\mathbf{v} \leq \mathbf{w}$ means $\forall i = 1, \ldots, d : v_i \leq w_i$
- The cardinality of a set S is denoted $|S|$

Math Recap: Projections

Recall from linear algebra:

Definition

The **scalar projection** of a vector $\mathbf{v} \in \mathbb{R}^d$ onto a vector $\mathbf{w} \in \mathbb{R}^d$ is $\Pi_{\mathbf{w}}\mathbf{v} = \mathbf{v}^\top \frac{\mathbf{w}}{\|\mathbf{w}\|}$.



Math Recap: Hyperplanes and Distances

Definitions

- An (affine-)**linear function** is a function $f : \mathbb{R}^d \to \mathbb{R}$ of the form $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$, where $\mathbf{w} \in \mathbb{R}^d (\mathbf{w} \neq \mathbf{0})$ and $b \in \mathbb{R}$.
- A **hyperplane** is a subset $H \subset \mathbb{R}^d$ defined as $H := \left\{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) = 0\right\}$.

Proposition (properties of hyperplanes)

Let H be a hyperplane defined by the affine-linear function $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$.

1. The vector $\mathbf{w}$ is **orthogonal** to H, meaning that: for all $\mathbf{x}_1, \mathbf{x}_2 \in H$ it holds $\mathbf{w}^\top (\mathbf{x}_1 - \mathbf{x}_2) = 0$.
2. The **signed distance** of a point $\mathbf{x}$ to H is given by $d(\mathbf{x}, H) \overset{\text{def.}}{:=} \pm \min_{\tilde{x} \in H} \|x - \tilde{x}\| \overset{!}{=} \frac{1}{\|\mathbf{w}\|} \left(\mathbf{w}^\top \mathbf{x} + b\right)$.

[presented at the board]

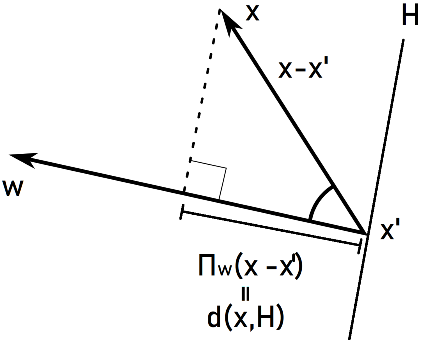
Prop1: $\forall \mathbf{x}_1, \mathbf{x}_2 \in H$ it holds $\mathbf{w}^\top(\mathbf{x}_1 - \mathbf{x}_2) = 0$.

Proof:

Let $x_1, x_2 \in H$. Then $\mathbf{w}^T x_1 + b = 0$ and $\mathbf{w}^T x_2 + b = 0$.
So:
$\mathbf{w}^T x_1 + b = w^T x_2 + b$
$\Rightarrow \mathbf{w}^T(x_1 - x_2) = 0$

Prop2: $d(\mathbf{x}, H) \overset{\text{def.}}{:=} \pm \min_{\tilde{x} \in H} \|x - \tilde{x}\| \overset{!}{=} \frac{1}{\|\mathbf{w}\|} \left(\mathbf{w}^\top \mathbf{x} + b\right)$

Firstly we notice that $\pm \min_{\tilde{x} \in H} \|x - \tilde{x}\| = \Pi_w(x - x')$ for an arbitrary $x' \in H$.



Prop2: $d(\mathbf{x}, H) \overset{\text{def.}}{:=} \pm \min_{\tilde{x} \in H} \|x - \tilde{x}\| \overset{!}{=} \frac{1}{\|\mathbf{w}\|} \left(\mathbf{w}^\top \mathbf{x} + b\right)$

Proof:

$d(\mathbf{x}, H) = \Pi_w(\mathbf{x} - \mathbf{x}') = \frac{\mathbf{w}^T(\mathbf{x} - \mathbf{x}')}{\|\mathbf{w}\|} = \frac{\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mathbf{x}'}{\|\mathbf{w}\|} \overset{*}{=} \frac{1}{\|\mathbf{w}\|}(\mathbf{w}^T\mathbf{x} + b)$

Where $(\star)$:

$\mathbf{x}' \in H \Rightarrow \mathbf{w}^T\mathbf{x}' + b = 0 \iff \mathbf{w}^T\mathbf{x}' = -b$

Linear Classifiers

Definition

A classifier of the form $f(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x} + b)$ is called **linear classifier**.

What are advantages and disadvantages of linear classifiers?

Please **pause** your video here and think about this question for a few minutes...

Linear Classifiers

Definition

A classifier of the form $f(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^\top \mathbf{x} + b)$ is called **linear classifier**.

What are advantages and disadvantages of linear classifiers?

Advantages

+ Easy to interpret
+ In practice: work well surprisingly often
+ Fast

Disadvantages

– Suboptimal performance if true decision boundary is non-linear
  - Occurs for very complex problems such as recognition problems and many others

NCC is a Linear Classifier

Theorem

NCC is a linear classifier $f(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x} + b)$ with $\mathbf{w} := 2(c_+ - c_-)$ and $b = \|c_-\|^2 - \|c_+\|^2$.

Proof

The decision boundary is given by
$H := \{x \in \mathbb{R}^d : \|x - c_-\| = \|x - c_+\|\}$.

We have:

$$\|x - c_-\| = \|x - c_+\|$$
$$\iff \|x - c_-\|^2 = \|x - c_+^2\|$$
$$\iff \sqrt{\sum_{i=1}^{d}(x_i - c_{-i})^2}^2 = \sqrt{\sum_{i=1}^{d}(x_i - c_{+i})^2}^2$$
$$\iff ||x||^2 - 2c_-^T x + ||c_-||^2 = ||x||^2 - 2c_+^T x + ||c_+||^2$$
$$\iff 2(c_+ - c_-)^T x + ||c_-||^2 - ||c_+||^2 = 0$$
$$\iff 2w^T x + b = 0.$$

Thus $H = \{x \in \mathbb{R}^d : 2w^T x + b = 0\}$

Conclusion

Linear Classifier: $f(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x} + b)$

- Fast and easy to understand

Example: NCC

- $f(\mathbf{x}) = \arg\min_{y \in \{-,+\}} \|x - c_y\|$