GHENT
UNIVERSITY

FACULTY OF ENGINEERING
AND ARCHITECTURE

DEPARTMENT ELECTRONICS AND INFORMATION SYSTEMS
RESEARCH GROUP IDLAB – IMEC

DATA ANALYTICS: MATHEMATICAL BACKGROUND AND ALGORITHMS

Tijl De Bie

GHENT
UNIVERSITY

TAKE-AWAY MESSAGE

Mathematics is key to data analytics

Search:
algorithms and optimization

Assessment:
statistical learning theory

GHENT UNIVERSITY

MACHINE LEARNING

Data		Label
Credit card transaction properties		Fraudulent or not
Patient symptoms		Diagnosis
Movie properties	➔	Rating
Customer properties		Churn probability
…		…

$\boldsymbol{x}_i$ y_i

Given n pairs $(\boldsymbol{x}_i, y_i)$, infer a general rule

GHENT UNIVERSITY

MACHINE LEARNING

- **Hypothesis space** $\mathcal{H}$
- Find $f \in \mathcal{H}$ such that $f(\boldsymbol{x}) \approx y$ most of the time

- Ideal approach:
  - Define **loss function**: $L(f(\boldsymbol{x}), y)$
  - **Risk** = expected loss given f:

$$R(f) = \mathbb{E}\{L(f(\boldsymbol{x}), y)\}$$

  - Find $f \in \mathcal{H}$ minimizing the **Risk**:

$$\min_{f \in \mathcal{H}} R(f)$$

GHENT UNIVERSITY

EMPIRICAL RISK MINIMIZATION

- $R(f) = \mathbb{E}\{L(f(\boldsymbol{x}), y)\}$ is unknown
- Can be estimated!

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} L(f(\boldsymbol{x}_i), y_i)$$

- **Empirical Risk Minimization:**

$$\min_{f \in \mathcal{H}} \hat{R}(f)$$

GHENT UNIVERSITY

$$\min_{f \in \mathcal{H}} \hat{R}(f)$$

Search

- How to find f efficiently?
- How does this depend on $\mathcal{H}$?

Assessment

- Given $\hat{R}(f)$, what about $R(f)$?
- How does this depend on $\mathcal{H}$?

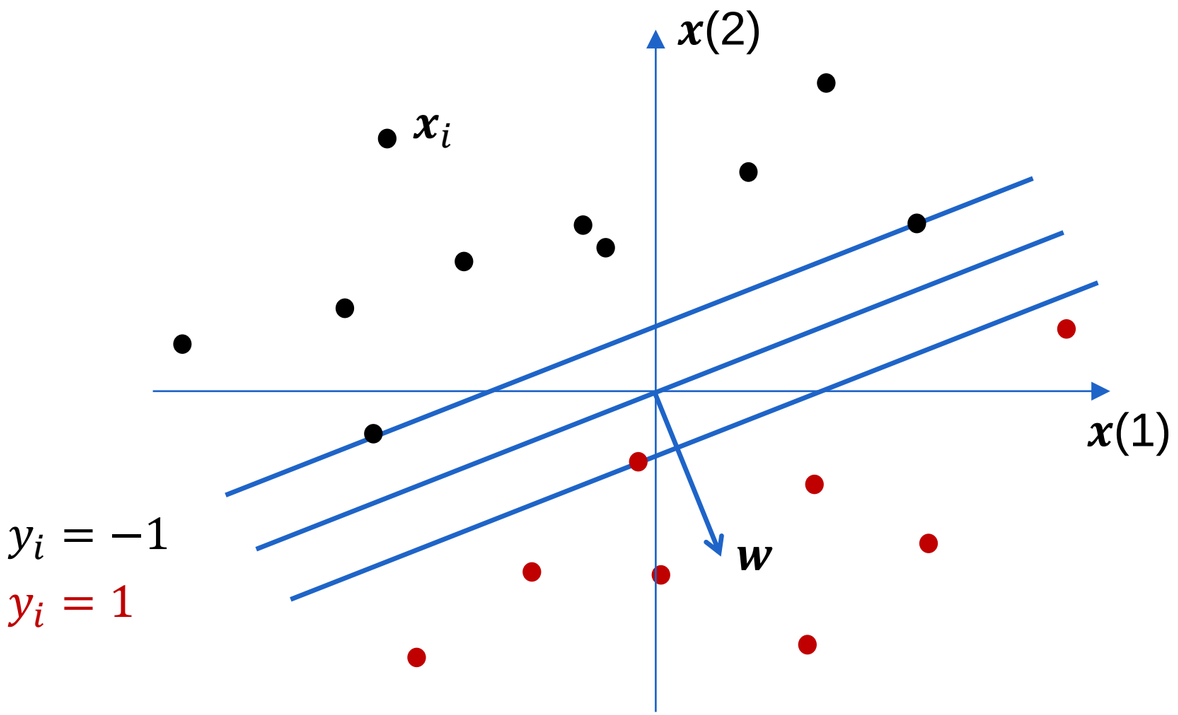
GHENT UNIVERSITY

SEARCH

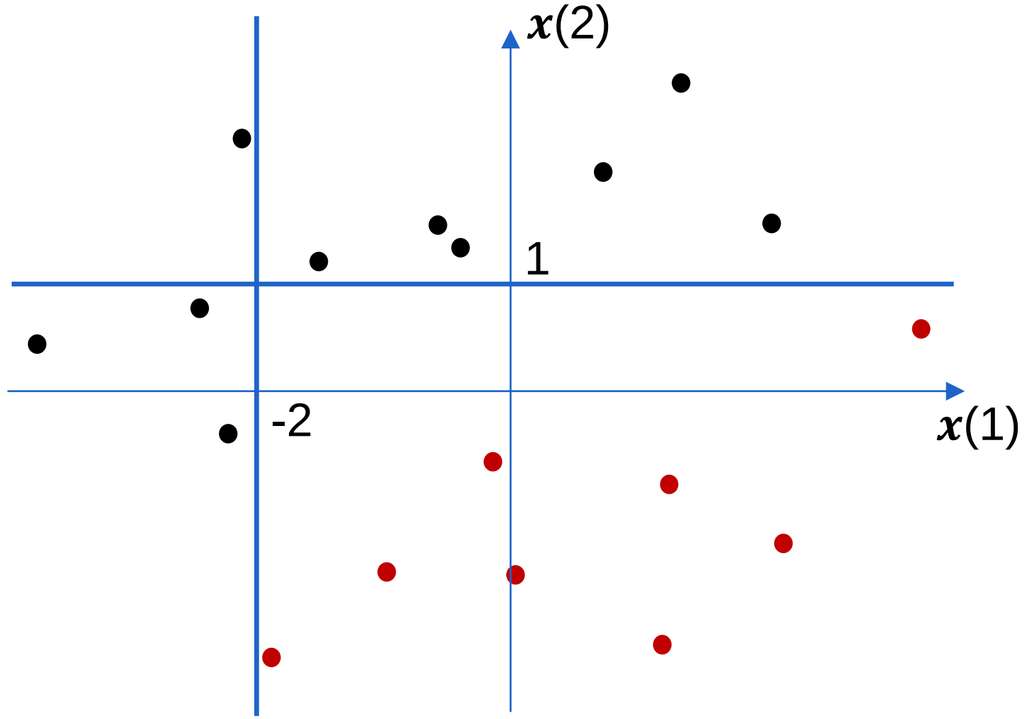
GHENT
UNIVERSITY

ALGORITHMS & OPTIMIZATION

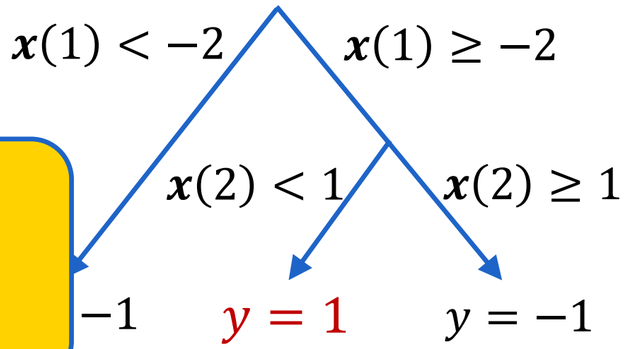
Support Vector Machines



Decision Trees



$$\max_{w} \frac{2}{w^T w} \quad \text{subject to:} \quad y_i(x_i^T w) \geq 1$$



Optimization theory
Algorithms
Computational complexity theory

GHENT UNIVERSITY

ASSESSMENT

GHENT
UNIVERSITY

STATISTICAL LEARNING THEORY

- Given f, statements of the following type:

$$P\left(R(f) > \hat{R}(f) + \varepsilon\right) < \delta(\varepsilon)$$

- Typically, $\delta(\varepsilon)$ decreases rapidly with increasing ε
- For a given f, easy to do!

- However, there is a problem…

GHENT UNIVERSITY

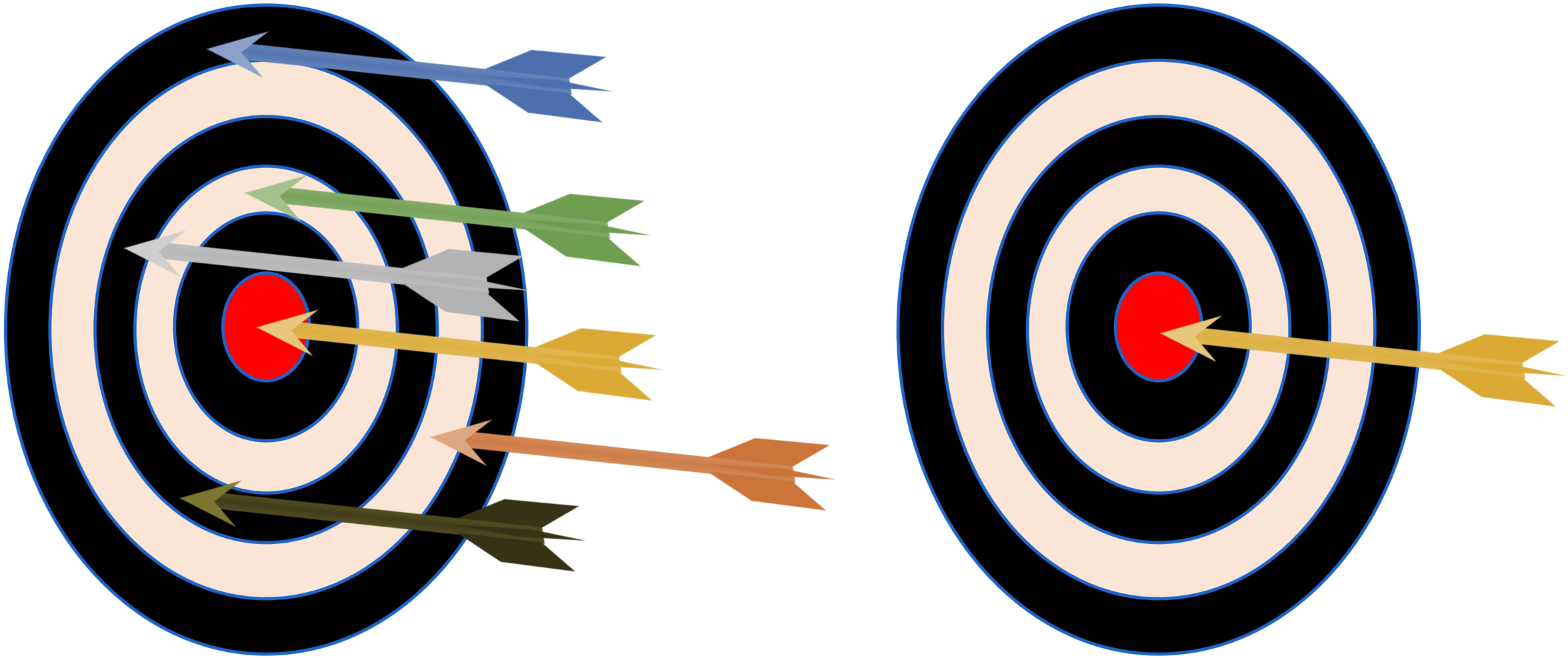
THE PROBLEM

- **Friend A:** “Last night I spotted a European roller (scharrelaar) in my garden!”
- **Friend B:** “I’m not impressed, I’ll check tonight and I’m sure I’ll spot on in my garden too.”
- Lo and behold, friend B does spot a European roller…

- **Friend C** would like to see a European roller too. Which garden should she try?

GHENT UNIVERSITY

THE PROBLEM



GHENT
UNIVERSITY

FIXING THE PROBLEM

- Quick fix:

$$P\left(R(f) > \hat{R}(f) + \delta\right) < |\boldsymbol{\mathcal{H}}|\varepsilon(\delta)$$

- Not very tight bound
- Fails when $|\mathcal{H}| = \infty$
- **Statistical Learning Theory:**

$$P\left(R(f) > \hat{R}(f) + \delta\right) < \boldsymbol{C_{\mathcal{H}}}\varepsilon(\delta)$$

GHENT UNIVERSITY

TAKE-AWAY MESSAGE

Mathematics is key to data analytics

Search:
algorithms and optimization

Assessment:
statistical learning theory

Both theories inform how to choose $\mathcal{H}$

GHENT UNIVERSITY

FACULTY OF ENGINEERING
AND ARCHITECTURE

Tijl De Bie

Professor

IDLAB – IMEC
ELECTRONICS AND INFORMATION SYSTEMS

E tijl.debie@ugent.be
T +32 9 264 33 66
M +32 468 29 69 74

www.ugent.be

Ghent University
@ugent
Ghent University

GHENT
UNIVERSITY