The Machine Learning Validation Challenge

Dr. Thomas Spieker, CR/AIR1, 22.11.2023

BOSCH

Introduction
Dr. Thomas Spieker

- **Department:** CR/AIR1
- **Position Title:** Research Engineer
- **Start Date:** 01.04.2020
- **Prior Experience:**
  - Joint Masters Degree in High Energy Physics
    (ETH Zürich & École Polytechnique Paris-Saclay)
  - PhD in Particle Physics with ATLAS
    (Kirchhoff Institute for Physics Heidelberg)
  - Topic: *“Detector Corrected Search for Dark Matter in Monojet and Vector Boson Fusion Topologies with the ATLAS Detector”*
- **Current Topic Assignment:** Machine Learning Validation & Verification

Who we are

Our company in figures

In 2022

88.2

billion euros sales revenue

3.8

billion euros EBIT from operations

421,300

Bosch associates worldwide at year-end (approx.)

468

subsidiaries and regional companies in 60 countries

Who we are
Our business sectors

Mobility Solutions

Industrial Technology

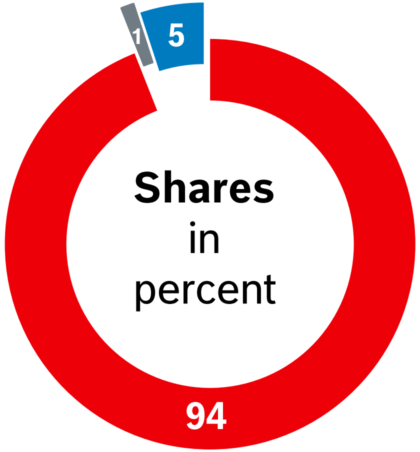
Energy and Building Technology

Consumer Goods

Who we are

Our ownership structure

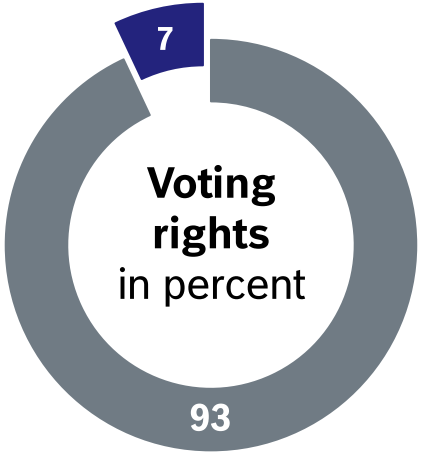
The vast majority of company shares are held by Robert Bosch Stiftung GmbH, a charitable foundation, while the vast majority of voting rights are held by Robert Bosch Industrietreuhand KG, an industrial trust. In other words, the foundation is the largest shareholder in Robert Bosch GmbH, but has no voting rights.



Robert Bosch GmbH

ERBO II GmbH/Bosch family

Robert Bosch Stiftung GmbH



The Bosch family

Robert Bosch Industrietreuhand KG

Where we want to go
Our research and development

In 2022

7.2

billion euros
R&D expenditure

8.2%

R&D intensity

85,500

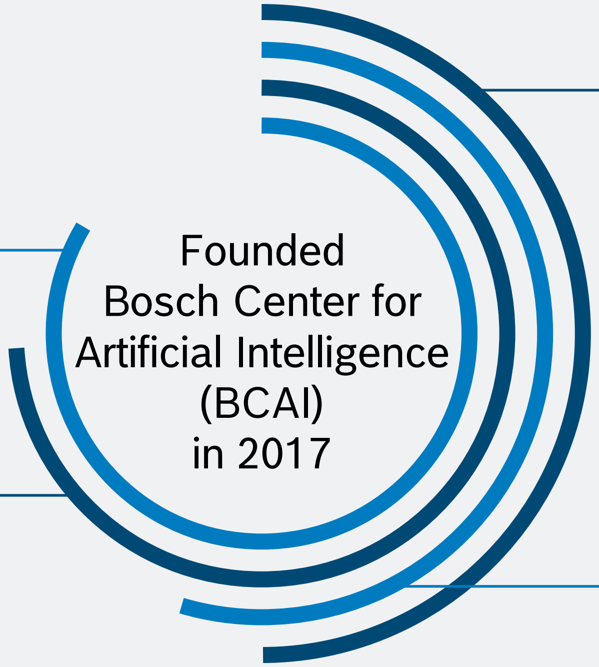
associates work
in R&D, including

44,000
software developers

136

R&D locations
worldwide

Where we want to go
Our AI commitment



Founded
Bosch Center for
Artificial Intelligence
(BCAI)
in 2017

One of the biggest european industry research labs for Machine Learning research

- Impressive publication record in top ML conferences
- Excellent PhD program

Close collaboration with business units to implement their ML solutions jointly

Enabling Bosch: training 1000s of employees in ML and data science

Incorporated and transformed into an active community within the CR structures in 2022

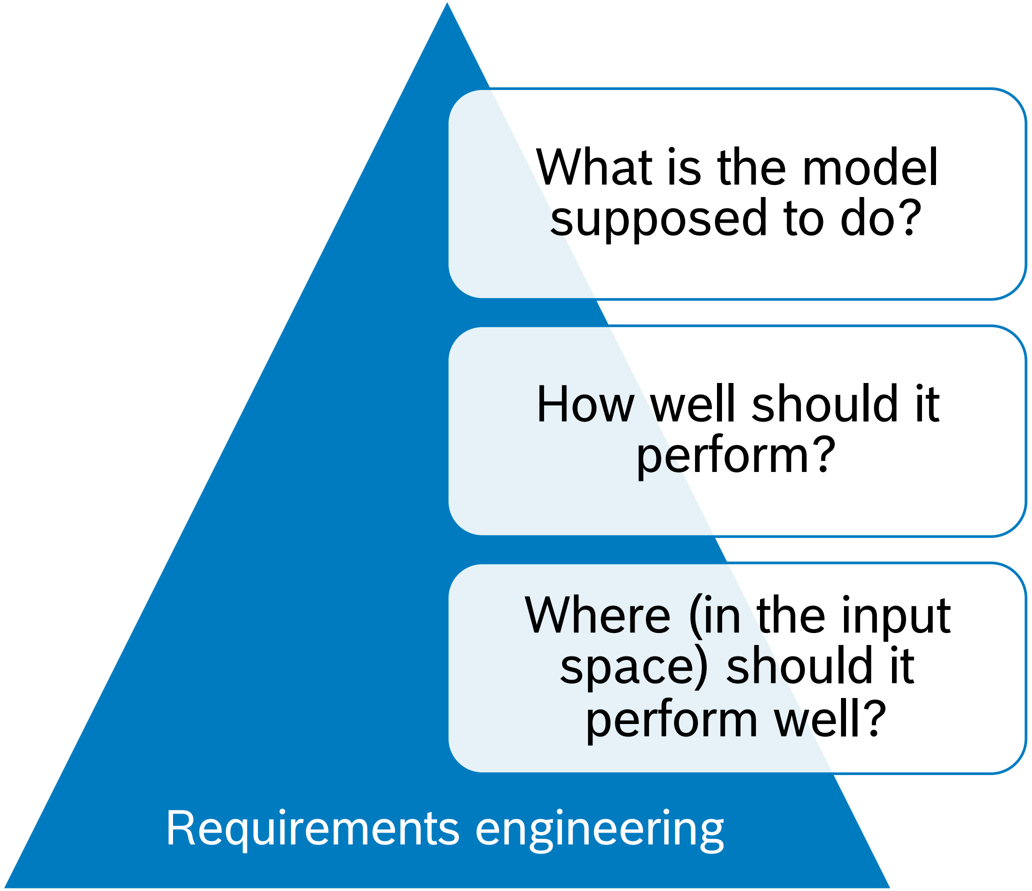
Where we want to go
Validation research

Work in close collaboration with our business units

Typical example: Powertrain Solutions

- Virtual sensor: Replace a physical sensor of a physical property by a function inferring the propery from periferal measurements
- Physical modelling underperforming
- High quality expectation
- High legal requirements
- Potentially safety critical

The ML Validation Challenge
Question 0



What is the model supposed to do?

How well should it perform?

Where (in the input space) should it perform well?

Requirements engineering

The hardest possible requirement:

The ML model has to predict
exactly the right value
everywhere,
every time,
no exceptions

The ML Validation Challenge
The problem with accuracy

95% accuracy is fine.

What does that even mean?

5 out of 100 examples are misclassified.

Yes, but which 5 examples were wrong?

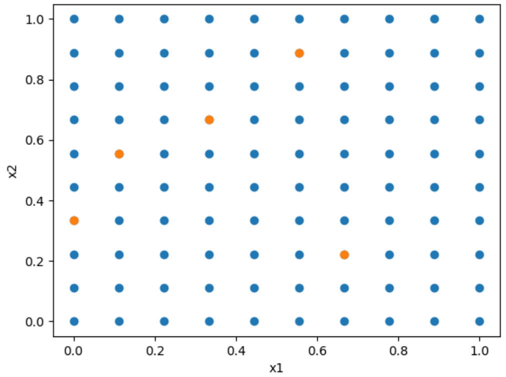
Were they clustered, or scattered randomly?

Are your test inputs distributed according to the distribution you expect in the field?

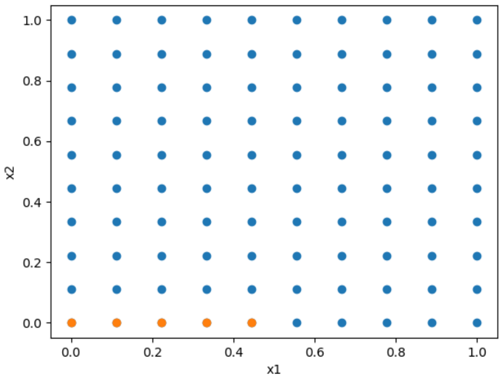
The ML Validation Challenge
How to interpret accuracy

- Test data uniform

errors random



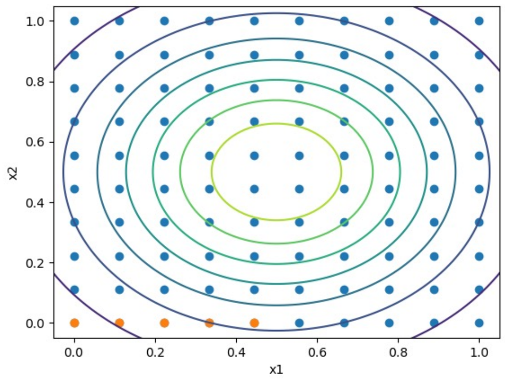
errors clustered



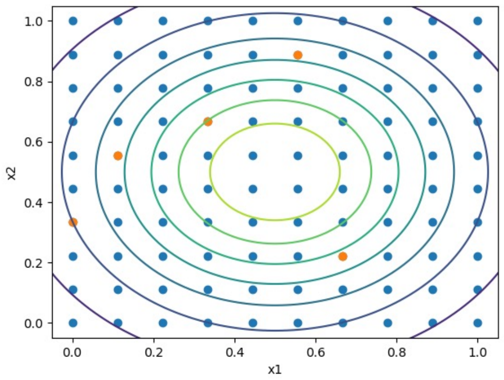
  - in 5% of the input space the model is erroneous
  - random errors → generalization issues?

- Field data Gaussian-distributed

errors random



errors clustered



  - experts might tell where errors are more/less harmful
  - need to “translate” this to the actual (field data) distribution

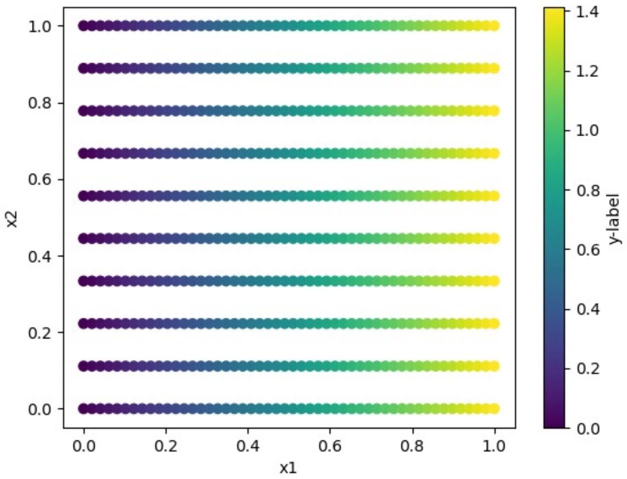
If test data is not distributed as the field data, then the 95% accuracy is meaningless.

The ML Validation Challenge

The reality

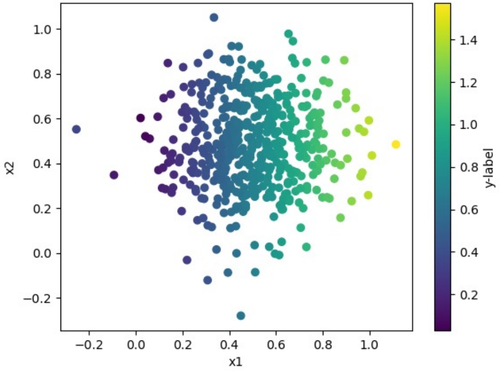
Test bench

- Data collection with fixed design of experiments (DoE)

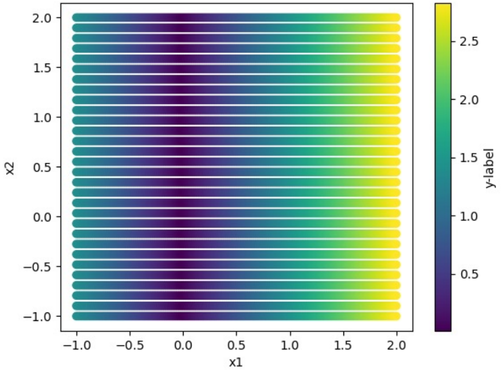


Field data

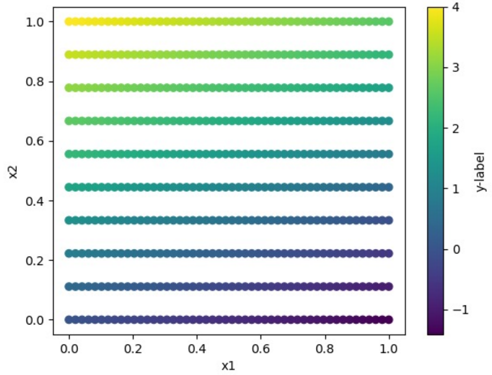
1. *Might* have different distribution



2. *Might* have different support



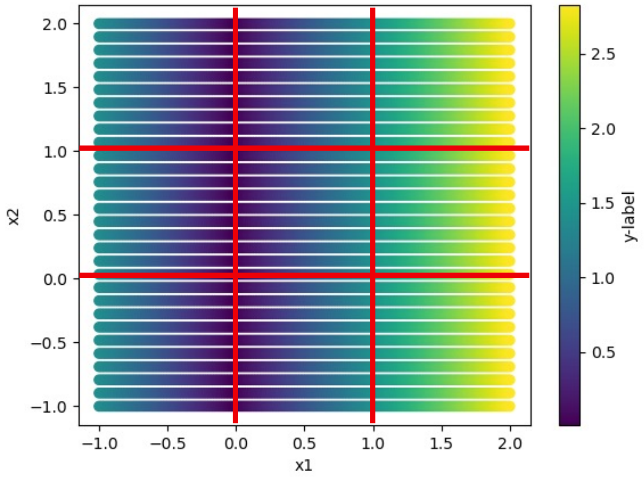
3. *Might* have different input-output relation



The ML Validation Challenge
Different support

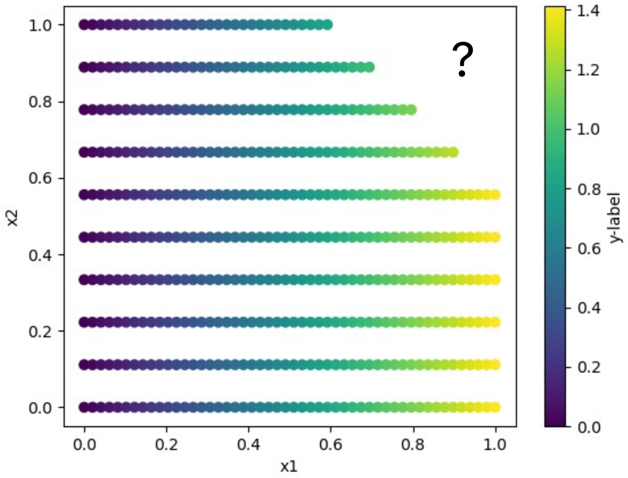
In reality there might occur inputs, which you have not seen in training/validation/testing data
→ Requires monitoring in production

- Simple: trim inputs



- Outlier classification:

→ requires negative examples

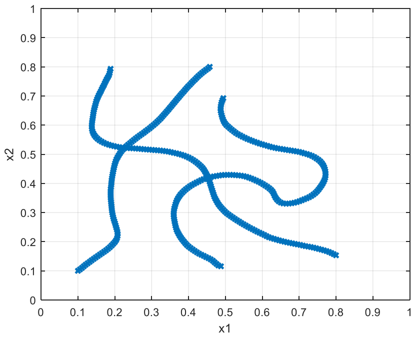


The ML Validation Challenge

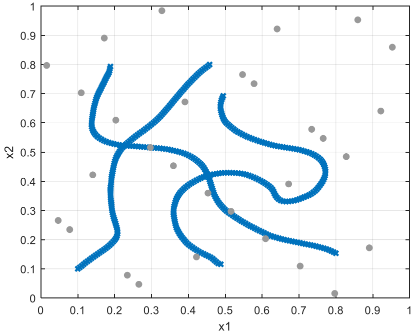
ML Model Input Monitoring

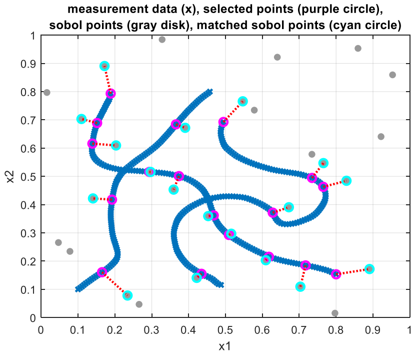
Objective

- Input space 10 .. 15 dim
- Valid input region is represented by training data
  - trajectories (= non-uniform)
  - vast amount of data points
- Invalid (outside) region not specified explicitly
- Monitoring function must run on ECU with AMU, in 100ms raster

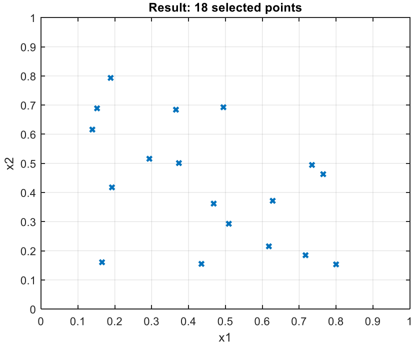


- **Step 1:** reduce data to space filling subset
- **Step 2:** generate out of distribution data
- **Step 3:** provide resource efficient classifier
  - RBF network with per-axes kernel width
  - Sigmoid output layer for classification
  - Training objective: log loss

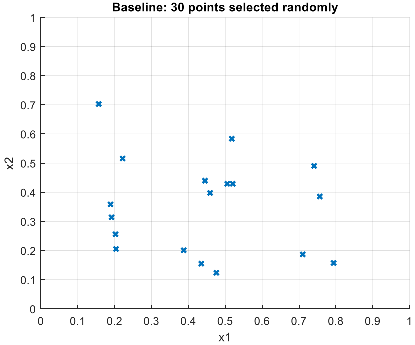




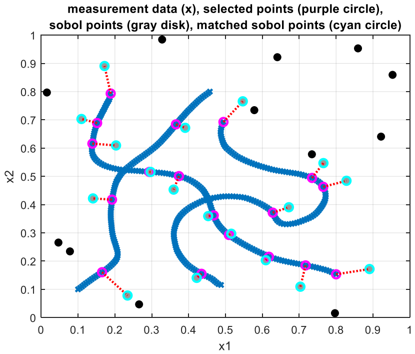
measurement data (x), selected points (purple circle), sobol points (gray disk), matched sobol points (cyan circle)



Result: 18 selected points



Baseline: 30 points selected randomly

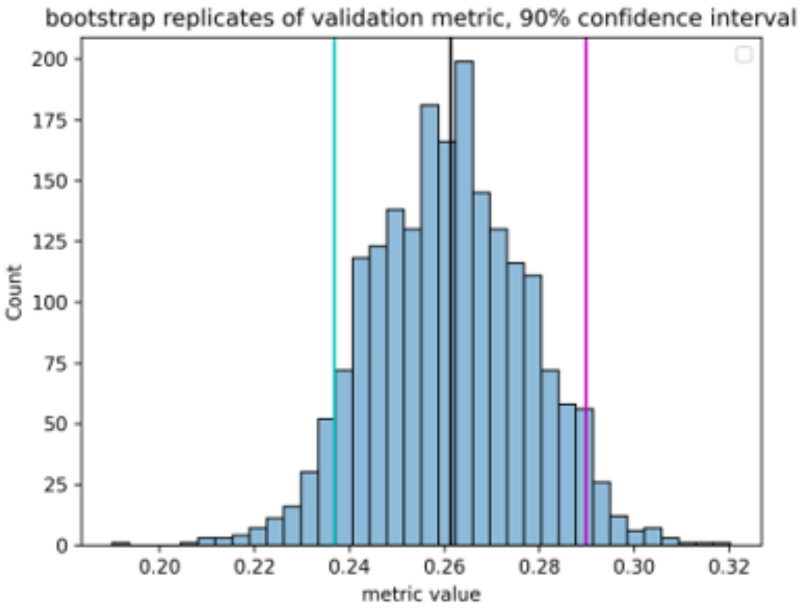


measurement data (x), selected points (purple circle), sobol points (gray disk), matched sobol points (cyan circle)

The ML Validation Challenge
Bootstrap Confidence Intervals

- Assume test data is distributed as in the field, but limited statistics
- Add bootstrap confidence intervals to indicate uncertainty of evaluation metric

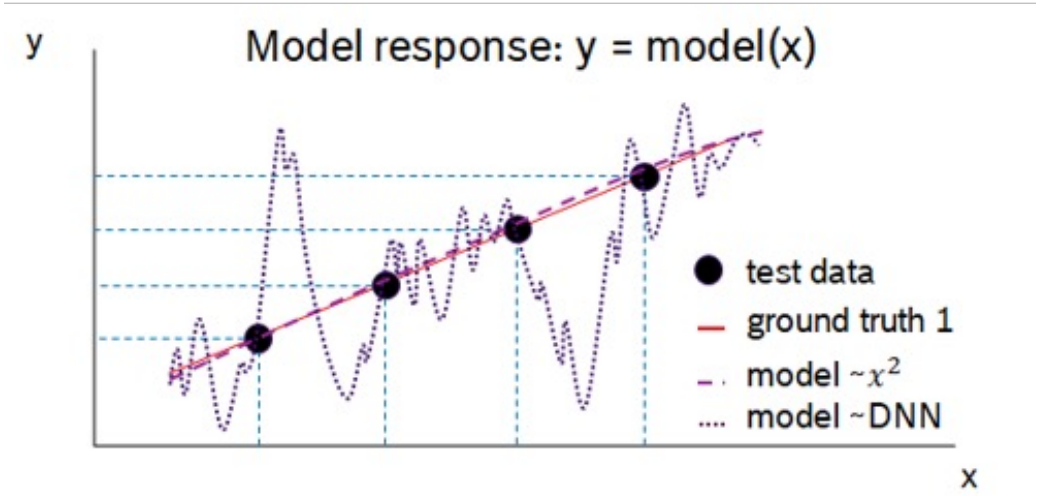
- Method details:
  - Repeat a few thousand times:
    - Form a new validation sample by sampling from the validation data (size N) N times with replacement
    - Compute the metric
  - In the histogram of the results, find the 5% and 95% percentiles
  - Apply some corrections for non-normality of the distribution (BCa algorithm)



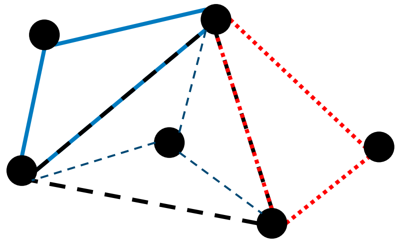
The ML Validation Challenge

Model generalization

- Assume that you want your model to predict “almost” exactly the right value everywhere
- ML model behavior only known at test points
- Model complexity may lead to unintended behavior in between
- Must make assumptions about ground truth

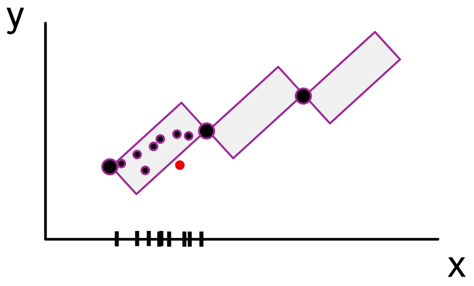


The ML Validation Challenge
Statistical testing of model smoothness



Objective: Quantify model behavior in entire input space (i.e., interpolation space)

- Compute convex hull and span with simplices
- Define 'corridor of truth'
  - Maximal and minimal bound to the y-label which the model is allowed to predict
  - Linear interpolation, truth function – each with allowed offset; Lipschitz continuity,…



- Sample and check within each simplex
- Hypothesis testing
  1. $H_0\colon \forall x \in X' \mid \|f(x)-y\| < c$ *aka* $p_0 = 0$
  2. $\mathrm{H}_1\colon \exists S \subset X'$ where $\emptyset \neq S = \{x \in X' \mid \|f(x)-y\| > c\}$ *aka* $p_0 > 0$
- Given no outliers are found

$$P(k=0|H_1) = (1-p_0)^n < \alpha \quad \Leftrightarrow \quad p_0 > 1-\alpha^{\frac{1}{n}}$$

- Can exclude the probability p_0 to find outliers at a confidence level $(1-\alpha)$

„If the model produces outliers, we would have seen it"

The ML Validation Challenge
Statistical Validation

- Given enough data (from the real world distribution) one can just test…
  - need to check distribution is the right one
  - need to check “really enough data” – since this is relative to the size of the input space
  - need to check that the train-test split is reasonable (don’t want to lie to yourself)
- Performance metric must measure system relevant properties

- What about particle physics?
  - plenty of data which can be considered IID
  - certainly according to the “real world distribution”
- But beware:
  - how much is enough data?
  - trigger systems?
  - extrapolation?

Bosch Research

Thanks for your attention. Still curious? Check us out online and visit our website and LinkedIn account.

Website

LinkedIn