Introduction

Aims of project

Models:

- Logistic Regression
- Random Forest
- Gradient Boosting
- XGBoost

Roadmap:

- Creation of synthetic data and data affected by shift
- Evaluation of model performance on the data
- Identification of improvement strategies (R.W.A.)

Dataset shift

- **Dataset shift** is a common problem in machine learning.
- It occurs when the distribution of the training data differs from the distribution of the test data.
- This can lead to a decrease in the performance of the model.

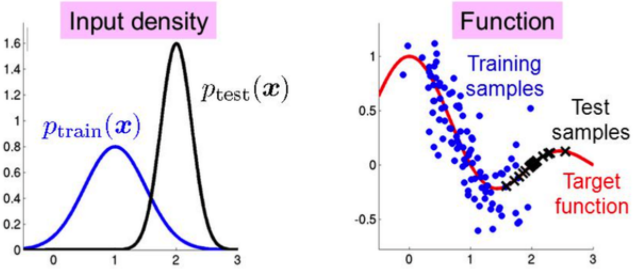
The two most common and well-studied causes of Dataset shift are:

- **Sample selection bias** (e.g. Economic studies)
- **Non stationary environments**

Covariate shift

Consider a target variable X and a response variable Y. Let P_{tra} denote the probability distribution of the training data and P_{tst} denote the probability distribution of the test data. A **covariate shift** occurs when:

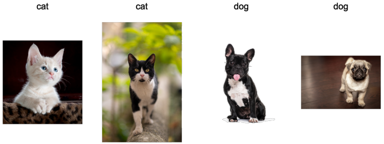
$$P_{\text{tra}}(Y \mid X) = P_{\text{tst}}(Y \mid X) \quad \text{but} \quad P_{\text{tra}}(X) \neq P_{\text{tst}}(X)$$



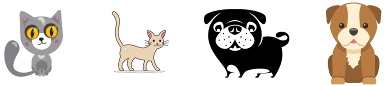
Example

Consider a model designed to distinguish between images of cats and dogs:

Training set:

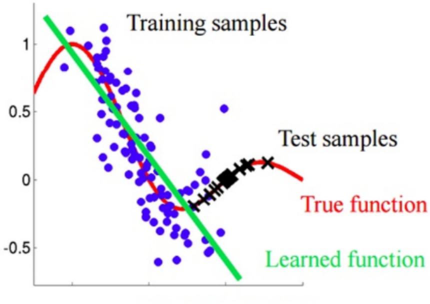


Test set:



Model will not accurately distinguish between cats and dogs because the feature distribution will differ.

Inaccurate Model



Changes in the features distribution can significantly impact the model's accuracy.