UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

DISTRIBUTION SHIFT

A Study on Their Effects on Statistical Models and Strategies for Mitigation

Andrea Spinelli, Giacomo Amerio,
Giovanni Lucarelli, Tommaso Piscitelli

University of Trieste

Table of contents

1. Introduction
2. Data Generation
3. Performance Degradation
4. Performance Enhancement

Introduction

Introduction

- Distribution 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.
- In this study, we analyze the effects of distribution shift on statistical models and propose strategies for its mitigation.

Data Generation

Training Dataset: Features

The dataset consists of $n = 10^4$ observations with 3 features and 1 target variable.

Training Dataset: Features

The dataset consists of $n = 10^4$ observations with 3 features and 1 target variable. **Features:**

- $X = (X_1, X_2, X_3) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- $\mu_i \sim \mathcal{U}_{[0,1]}$ for $i = 1, 2, 3$
- $[\boldsymbol{\Sigma}]_{i,j} \sim \mathcal{U}_{[-1,1]}$ for $i, j = 1, 2, 3$

Note: The $\boldsymbol{\Sigma}$ randomly generated has been transformed to a symmetric and positive semidefinite matrix by computing $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T$.

Training Dataset: Target Variable

Dicothomous target variable: $Y \in \{0, 1\}$

$$z = \beta_0 + \sum_{i=1}^{3} \beta_i x_i + \sum_{i=1}^{3} \beta_{ii} x_i^2 + \sum_{i=1}^{2} \sum_{j=i+1}^{3} \beta_{ij} x_i x_j$$

$$Y \sim \mathrm{Be}(p)\,, \quad p = \frac{1}{1 + e^{-z}}$$

Testing Dataset

Performance Degradation

Performance Enhancement

Questions?

References i