CE888 Assignment
Project 1: Learning under Covariate Shift in the Data (Supervised Learning)

NIRUPMA PANDEY
Registration number: 1906463

Abstract—Internet availability in portable devices has boomed the steaming fashion and made it easy to collect data, encouraging the need for data mining to extract information to predict consumers behavior/like results in developing a smarter market campaign. In this project, we are aiming to understand and deal with one of the types of non-environmental data shift, covariate shift, which results in dramatic variation in prediction. We are going to demonstrate the procedure to find drift in provided data-set and work on a proposed methodology to tackle this problem, followed by calculation of classification accuracy.

Index Terms—Data Shift, Covariate Shift, AUC score, Classification accuracy.

1 Introduction

With a rapid increase in usage of cell phones, technical advancement like the internet of things and a wide network of sensors, have promoted web browsing, resulting in pulling off an immersive number of data. The increase of internet streaming and easy availability of data has encouraged machine learning and data mining to extract useful information upon finding traits/patterns to predict an outcome. But sometimes, due to the presence of several environment/non-environmental factors causes data shift, which results in biased prediction. The aim of this project to trace the covariate shift in data and methods to fix it.

1.1 Data shift and it's types

In predictive modeling, data shift is a common problem caused due to uneven joint distribution of input and output at training and testing stages. When a large data-set is collected over a long period of time, multiple variables/environment changes might have happened, potentially produce a non-uniform distribution of train and test data, resulting in the lower testing score[1].

Broadly data shifts are divided into three types:

(a) Covariate Shift - shift in independent variables, also known as covariates.

(b) Prior Probability Shift - shift in target variable in train and test data sets.

(c) Concept Shift - shift in relationship between independent variables and target variables.

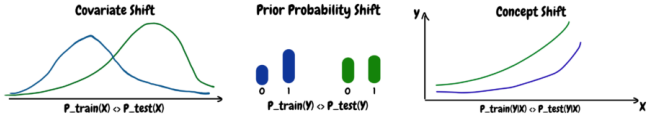

Fig. 1. Three types of dataset shift. (1) Different X distributions in covariate shift. (2) Different Y distributions in prior probability shift. (3) Different relationship between Y and X in concept shift

Working on data-set with any one of the above-mentioned shifts could results in uneven distribution of features, resulting in model failure and generating false prediction. Hence, ideally, to get the most accurate prediction, we need to check for shifts, followed by the application of techniques to minimise the effect, upon applying a model to get the final result.

2 Background

In supervised learning, the aim of learning is to infer an input-out dependency from training data, using which output of an unknown test data can be predicted. We generally assume that input points of both train and test are having the same probability distribution[3]. Covariate shift is a specific class of selection bias that arises when the marginal distributions of the input features X are different in the source and the target domains while the conditional distributions of the target Y given X are the same, mathematically

$P_{train}(X) \neq P_{test}(X) and P_{train}(Y|X) = P_{test}(Y|X)$,

where X is a feature [1]

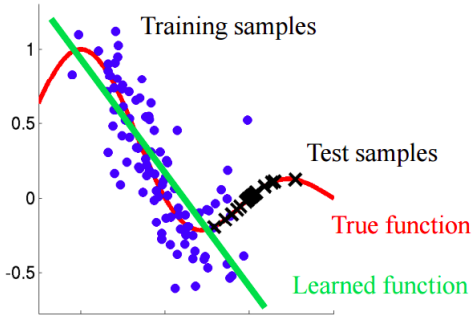

Fig. 2. Covariate shift in data

The problem of covariate shift can be found in real-world instants very easily, for example, data of working-age group

for an insurance policy, data collected over a long period of time on working efficiency of machinery, filtering spam emails, brain-computer interfacing[3]. In the first example, if data of people aged between 25-60 has been collected and distributed such that training contains fairly young candidates, whereas elderly people are classified as a test, it results in a significant difference in mean distribution.

Similarly, if records collected on the working efficiency of machines in a production plan over a period of time, may result in a significant difference in distribution, due to wear and tear of some/most of the parts.

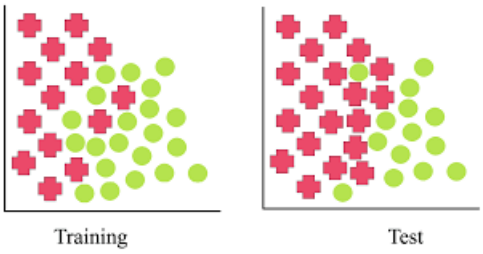

Fig. 3. Example showing distribution difference in test and training data

3 Methodology

In the section we are going to propose and discuss methods to find, then minimise covariate shift in a given data set.

3.1 Methods to find Covariate shift in a data-set

Methods to find Covariate shift in a data-set One of the easiest and popular method of tracing a drift in data distribution is plotting histogram between corresponding independent variables of train and test, we can visualise if their distribution is uniform or with a drift.

3.2 Methods to deal with drift in a data-set

Feature engineering -All machine learning algorithms need some input data to compress some features to predict output, hence feature engineering is required to meet the specific need of algorithm allow them to work properly. some of the examples are imputation, feature splitting, handling outliers, etc.

Transfer learning- It is a field of machine learning, where we aim at storing knowledge gain from learning source data distribution of a well-performing model and applying it on a different but relevant target data distribution.

3.3 Comparative study of datasets before and after treatment

After minimising the effect of shift, we will perform some comparative study on the data set to measure the accuracy or correctness of the model. One of the best methods of measuring accuracy is by plotting the ROC curve and AUC.

The receiver operating characteristic curve (ROC) is a graphical representation of the performance of the classification model at all classification threshold, by plotting two parameters- True positive rate and False positive rate.

TPR = $\mathrm{TP}\frac{}{TP+FN}$

FPR = $\mathrm{FP}\frac{}{FP+TN}$

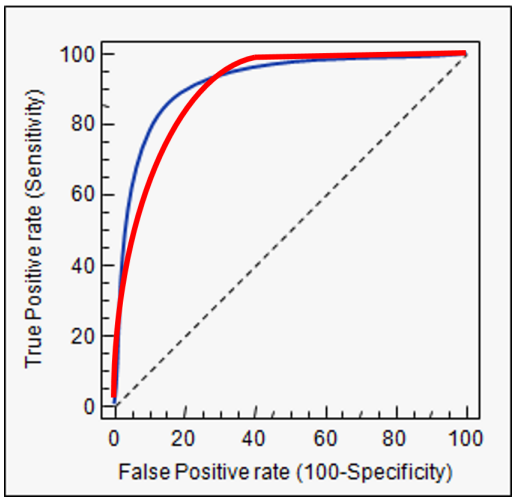

Fig. 4. ROC curve showing False positive rate vs True positive rate

3.4 Algorithm

Here is an algorithm of steps to follow: 1) Add origin feature that tells us whether the observation belongs to training or testing dataset.

2) Merge training and testing datasets.

3) Create a classification model taking examined feature as an explanatory variable and origin feature as a target on a part of the merged dataset (75 percent)

4) Predict on the rest of merged dataset (25 percent)

5) Calculate the value of the Phi coefficient and ROC-AUC score and set a threshold of 0.30 and 0.80 respectively.

If the value of the Phi coefficient and ROC-AUC score is greater than the set threshold we identify the examined variable as drifting and restart the whole algorithm until both values are below the threshold.

4 Result

To find if there is any drift in the provided dataset we will perform the algorithm mentioned in section 3.4. At the end of each iteration we will calculate the Phi coefficient value and ROC-AUC score, after comparing these values with the set threshold we will observe whether data if free from drift or not. Once data is free from any sort of data shift we will be able to apply any of the machine learning techniques to predict an outcome.

5 Discussion

In the study, we will be calculating the ROC-AUC score of treated data and then after comparing the same with a fixed threshold of 0.80, we will consider that the data is not biased and uniformly distributed in train and test data set. Hence, with the help of balanced data, we will be able to predict the accurate result and improve the performance of our model.

6 Conclusion

During this study, we have found that using a biased dataset, could result in loss of important information and not reliable for a recommendation system as it will favor highly weighed variable. Hence it is advisable to check for

the presence of any kind of data shift and minimise/treat if any.

7 Plan

In order to complete the in-depth data analysis, we have divided task in steps and going to follow them as per the mentioned timeline:

1) Data Analysis- Familiarising with the dataset it the first step to start with deep learning, to get familiar with all independent variables and target variables and finding their relationship with each other. The self-set deadline is March 2nd.

2) Finding the best methods to deal with data shifts- this step requires reading some more research papers and books to get an idea in which all approaches can be taken. The self-set deadline is March 16th.

3) Application of methods to find and minimise shift in data- this part requires working on codes and know which are key variables to use and what all parameters need to be considered to get better results.The self-set deadline is March 30th.

4) Report writing and summarising - this is the final step of our project which requires a sound knowledge of IEEE format, overleaf and written content of the whole procedure. The self-set deadline is April 13th.

References

[1] G. Ditzler, M. Roveri, C. Alippi, and R. Polikar, "Learning in nonstationary environments: A survey," IEEE Computational Intelligence Magazine, vol. 10, no. 4, pp. 12–25, 2015.

[2] H. Raza, G. Prasad, and Y. Li, "Adaptive learning with covariate shift-detection for non-stationary environments," in 2014 14th UK Workshop on Computational Intelligence (UKCI). IEEE, 2014, pp. 1–8.

[3] M Sugiyama, M Krauledat, KR MÃžller, "Covariate shift adaptation by importance weighted cross validation" - Journal of Machine Learning . . . , 2007