Learning under Covariate Shift in the Data CE888 Assignment 1

Belia Maria Baez Molina, *1900952*

***Abstract*—For this first part of the project about the covariate shift, it was necessary to analyze and compare two datasets in order to see if there is a change between the input and the output, training to the testing phase.**

The differences in topics and relevance between the data contribute to having a more complete analysis of how the covariate shift affects or not the learning under covariate shift. Showing the difference between learning with covariate shift and after minimizing the covariate shift.

***Index Terms*—covariate shift, deep neural networks , dataset, LaTeX, paper, template.**

I. Introduction

ACCORDING to The MIT Press [1], "Dataset shift is a challenging situation where the joint distribution of inputs and outputs differs between the training and test stages", an example in Fig1. It is important to determine if the shift on the data set can become a problem in the algorithm. The effect of the shift in the data depends on the size of the dataset and the ability of the data scientist to resolve it [2].

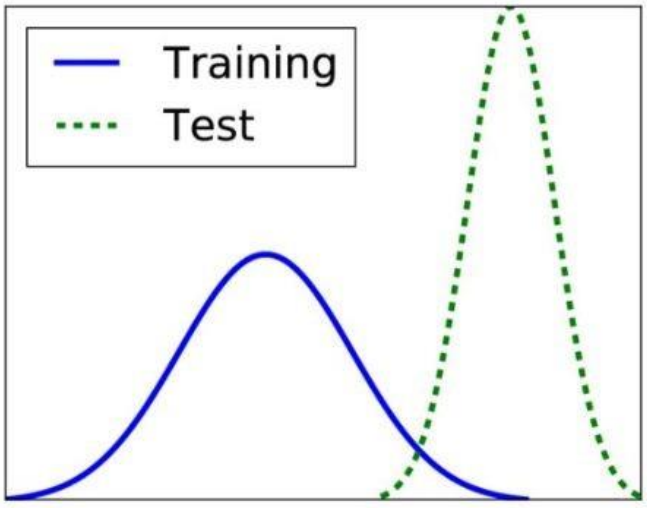


Fig. 1. Example of covariate shift: training and test data having different distributions [2.1].

For this project is needed to learn the data under covariate shift, [1] defines it as "a simpler particular case of dataset shift where only the input distribution changes (covariate denotes input), while the conditional distribution of the outputs given the inputs p(y—x) remains unchanged". This topic is important because it has an impact in the deep neural network, [3] says that a covariate shift is "the change in the distribution of network activations due to the change in network parameters during training." In order to get better training for the network, the internal covariate shift needs to be reduced, because it can impede the training of deep neural networks [2].

II. Background

Sugiyaaa, Krauledat and Müller [4] with the support of other authors said that ". . . from many applications such as off-policy reinforcement learning (Shelton, 2001), spam filtering (Bickel and Scheffer, 2007), bioinformatics (Baldi et al., 1998; Borgwardt et al., 2006) or brain-computer interfacing (Wolpaw et al., 2002), the covariate shift phenomenon is conceivable.

Understanding the importance of the applications that can be affected, [4] has developed a new learning technique that "can alleviate misestimation due to covariate shift".

Another work that is relevant into the area is the adaptive learning with covariate shift-detection for motor imagery-based brain-computer from [5] where they present "an on-line brain-computer interfaces (BCI) system using previously developed covariate shift detection (CSD)-based adaptive classifier to discriminate between mental tasks and generate neurofeedback in the form of visual and exoskeleton motion".

Also, [5] start's the project with a covariate shift-detection which detects the possible shift in "features extracted from motor imagery-based brain responses". After that detection test, the investigation of [5], starts updating the classifier in those testing phases.

This proposed method [5], uses real data and "show a statistically significant improvement in the classification accuracy of the BCI system over traditional learning and semi-supervised learning methods."

III. Methodology

In order to start working with a dataset, is important to inspect them mainly to see if there is a covariate shift in there. That is why this is the first task of the project. One of the most useful methods for detecting a change over time in the data is with plotting histograms. It helps "to detect whether your model predictions change over time, but also check if your most important features change over time.[2]"

Being said that, in order to achieve the purpose of the project, there were selected two datasets from the community Kaggle.com, those were Santander Customer Transaction Prediction and Titanic: Machine Learning from Disaster. Both datasets were previously used in different challenges.

Continuously, to obtain the data from Kaggle, it was downloaded to the virtual machine of google colab via the API of the dataset. kaggle competitions download - Santander-customer-transaction-prediction in Fig2, Kaggle competitions download -c titanic on Fig3. Once is already download, it is

necessary to unzip them to start working. The process consists of after unzipping them, save the train.csv and test.csv file into variables for the easy manipulation. After that, with the method dataset.head() the first five rows are shown to verify if the data was stored correctly.

Upload files, train and test

```
[ ] #train of santander
    train_sntdr = pd.read_csv(path_sntdr + 'train.csv')
    train_sntdr.head()
```

	ID_code	target	var_0	var_1	var_2	var_3	var_4	var_5	var_6	var_7	var_8	var_9	var_10
0	train_0	0	8.9255	-6.7863	11.9081	5.0930	11.4607	-9.2834	5.1187	18.6266	-4.9200	5.7470	2.9252
1	train_1	0	11.5006	-4.1473	13.8588	5.3890	12.3622	7.0433	5.6208	16.5338	3.1468	8.0851	-0.4032
2	train_2	0	8.6093	-2.7457	12.0805	7.8928	10.5825	-9.0837	6.9427	14.6155	-4.9193	5.9525	-0.3249
3	train_3	0	11.0604	-2.1518	8.9522	7.1957	12.5846	-1.8361	5.8428	14.9250	-5.8609	8.2450	2.3061
4	train_4	0	9.8369	-1.4834	12.8746	6.6375	12.2772	2.4486	5.9405	19.2514	6.2654	7.6784	-9.4458

5 rows × 202 columns

Fig. 2. Method dataset.head(), visualization of the store data.

Upload of files train and test

```
[ ] #Train for titanic
    train_titanic = pd.read_csv(path_titanicData + 'train.csv')
    train_titanic.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age
0	1	0	3	Braund, Mr. Owen Harris	male	22.0
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0
2	3	1	3	Heikkinen, Miss. Laina	female	26.0
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0
4	5	0	3	Allen, Mr. William Henry	male	35.0

Fig. 3. Method dataset.head(), visualization of the store data.

By practicality, the process is first made with one dataset, in this case the Santander one. Afterward, both .csv files are stored is necessary to regulate them so both files are with the same parameters and format in order to compare them correctly.

To make sure they are in the same format is necessary to erase the columns that have strings instead of integers as the data value. A histogram is made only out of numbers. That is using the method dataset.drop()with the parameters the columns to be erased.

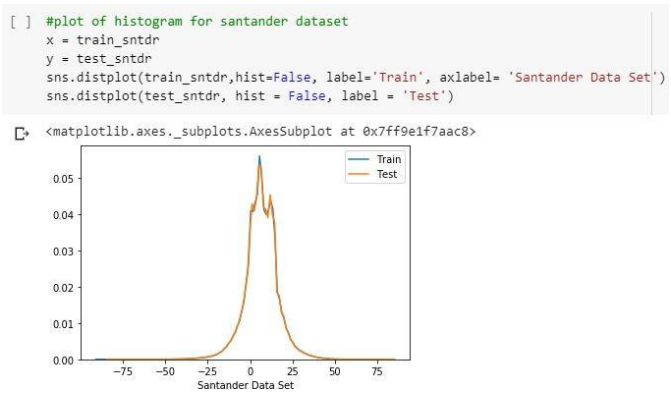
Subsequently, is time to plot the histogram to verify if there is any shift in the data. Those are made by the method sns.distplot(). This type of graphic helpful to observe easily the change in data because is a continuous line. Once the Santander dataset is plotted, the same process is needed to be made in the Titanic dataset.

IV. Results

Now that the histograms are plotted, they can be inspected to decide if there is a shift in the data or not.

As a result of the plotting, it is easily notated that in the Santander Fig4, plot there is not any change, they are practically the same graph for the data in the training file as in the test file.

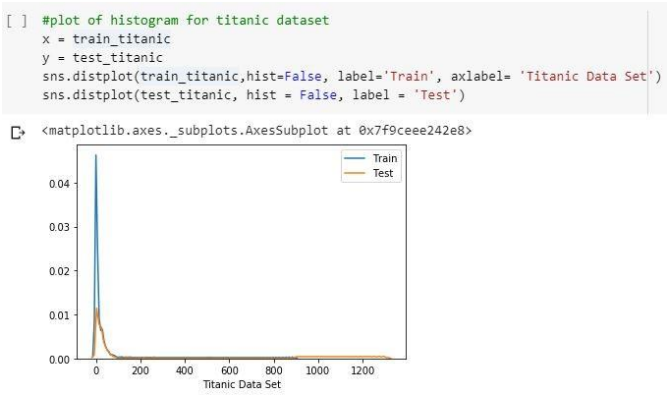
On the other hand, the Titanic, Fig5 plot has a notorious difference in the first values of the graph and that is the determinant to classify that the Titanic dataset has a covariate shift.



```
[ ] #plot of histogram for santander dataset
    x = train_sntdr
    y = test_sntdr
    sns.distplot(train_sntdr,hist=False, label='Train', axlabel= 'Santander Data Set')
    sns.distplot(test_sntdr, hist = False, label = 'Test')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff9e1f7aac8>
```


Fig. 4. Plot of histogram for Santander dataset



```
[ ] #plot of histogram for titanic dataset
    x = train_titanic
    y = test_titanic
    sns.distplot(train_titanic,hist=False, label='Train', axlabel= 'Titanic Data Set')
    sns.distplot(test_titanic, hist = False, label = 'Test')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9ceee242e8>
```


Fig. 5. Plot of histogram for Titanic dataset

V. Discussion

One of the tasks for the next assignment is to minimize the shift between the data and see if they see the difference between learning with the covariate shift and after minimizing the covariate shift.

To understand the difference between the two learnings a helpful method to compare the results on both datasets is with AUC scores, which helps to predict which class is the best, or classification accuracy, which helps to defines how many items correctly classified.

VI. Conclusion

As in the method of [5] there was an improvement after adjusting the covariate shift, it is expected to have a similar improvement in this project.

This project permits us to apply different methods seen in class and to understand them better because they combine a small part from different laboratories at the same time.

VII. References

[1]J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, Dataset shift in machine learning. The MIT Press, 2009

[2]M. Stewart, Understanding Dataset Shift,Towardsdatascience, Dec. 2019, Accessed on Feb. 19, 2020. [Online]. Available:

https://towardsdatascience.com/understanding-dataset-shift-f2a5a262a766

[3] Glauner, Patrick State, Radu Valtchev, Petko Duarte, Diogo. (2018). On the Reduction of Biases in Big Data Sets for the Detection of Irregular Power Usage. 10.1142/97898132732380057.

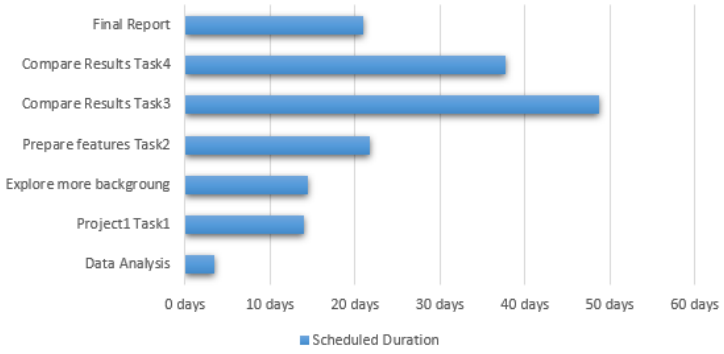
[4] S. Ioffe, C. Szegedy, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv preprint arXiv:1502.03167 (2015).

[5] M. Sugiyama, M.Krauledat, K. R. MÃžller, ”Covariate shift adaptation by importance weighted cross validation”. Journal of Machine Learning Research, May 2007.

[6] A. Chowdhury, H. Raza, Y. K. Meena, A. Dutta and G. Prasad, ”Online Covariate Shift Detection-Based Adaptive Brain–Computer Interface to Trigger Hand Exoskeleton Feedback for Neuro-Rehabilitation,” in IEEE Transactions on Cognitive and Developmental Systems, vol. 10, no. 4, pp. 1070-1080, Dec. 2018.

VIII. PLAN

SOFTWARE DEVELOPMENT PLAN



APPENDIX A
LINK TO A GITHUB PROJECT

This project has been save in the repository CE8882020*inttithubhttps* : *//github.com/beliabaez/CE*8882020*/blob/master/Assignment*1*.ipynb*