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

Registration Number: 1908015

Abstract—An aim of this report is learning under covariate shift in the datasets and to more understand a scenario where your machine learning model can be failed by covariate shift. Two different datasets will be chosen and then investigated to see whether there is a covariate shift in those datasets or not. The chosen datasets are, Sberbank Russian Housing Market and For Medical Cost Personal Datasets , from Kaggle website. the ROC-AUC metric and histogram plot will be used for an estimate and show how much covariate shift the features of datasets have. In this study, feature selection and feature importance will be used when there is a covarite shift in features of datasets and Root Mean Square Error (RMSE) will be used to estimate test error and then compare the results between employing feature selection, feature importance, and no any techniques applied with drifting features.

1 Introduction

Supervised learning is the most popular used among data scientists compared to another two types of machine learning, which are unsupervised and reinforcement learning. Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output $Y = f(X)$. The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data. Supervised learning problems can be further grouped into regression and classification problems.

- Classification: A classification problem is when the output variable is a category, such as "red" or "blue" or "disease" and "no disease".
- Regression: A regression problem is when the output variable is a real value, such as "dollars" or "weight".

However, the problem of supervised learning which usually occur is that there can be a shift in the data [1]. Dataset shift is occurred when there is a difference in the joint distribution of training and testing [2]. There are several types of dataset shift such as Covarite shift, Prior probability shift, Concept shift, etc.

- Covariate shift appears only in $X \rightarrow Y$ problems, and is defined as the case where $P_{tra}(y|x) = P_{tst}(y|x)$ and $P_{tra}(x) \neq P_{tst}(x)$
- Prior probability shift appears only in $Y \rightarrow X$ problems, and is defined as the case where $P_{tra}(x|y) = P_{tst}(x|y)$ and $P_{tra}(y) \neq P_{tst}(y)$
- Concept shift defined as $P_{tra}(y|x) \neq P_{tst}(y|x)$ and $P_{tra}(x) = P_{tst}(x)$ in X$\rightarrow$ Y problems. $P_{tra}(x|y) \neq P_{tst}(x|y)$ and $P_{tra}(y) = P_{tst}(y)$ in Y$\rightarrow$ X problems

Covariate shift, one of the most appearance, is a simpler particular case where only the input distribution changes (covariate denotes input), while the conditional distribution of the outputs given the inputs $p(y|x)$ remains unchanged [1] and there is another name of Covariate shift which is called "population drift" [2]. Typically, it is determined as "Covariate shift appears only in X $\rightarrow$ Y problems, and the case where the conditional probability in trainning and testing remains the same ($P_{train}(y|x) = P_{test}(y|x)$), but the input distribution *P(x)* changes between training and testing, i.e., ($P_{train}(x) \neq P_{test}(x)$). Figure 1 shows an example of difference in distribution between training and testing test which cause Covariate shift. There will be an example case where covariate shift occur below; Suppose we wish to generate a model to diagnose breast cancer. Suppose, moreover, that most women who participate in the breast screening test are middle-aged and likely to have attended the screening in the preceding 3 years. Consequently our sample includes mostly older women and those who have low risk of breast cancer because they have been tested before. This problem is referred to as sample selection bias. The examples do not reflect the general population with respect to age (which amounts to a bias in $P_{tr}(x)$) and they only contain very few diseased cases (i.e. a biasin $P_{tr}(y|x)$).

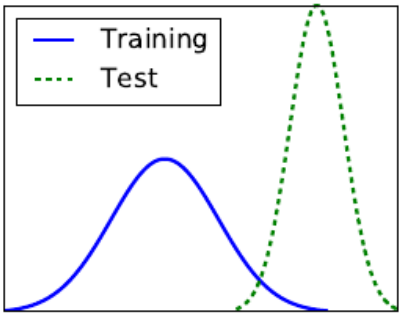

Fig. 1: an example of difference in distribution between training and testing test which cause Covariate shift

The purpose of this study is to understand how to detect covariate shift and see if the performance of using feature selection e.g. dropping features, keeping important features will make the model get better accuracy compare to not using feature selection.

2 BACKGROUND

2.1 Previous Works

In the previous study, [8] Authors stated that standard learning methods such as maximum likelihood estimation are no longer consistent but weighted variants according to the ratio of test and training input densities are consistent. Hence, estimating the density ratio is one of the key issues in covariate shift adaptation. For high dimensional cases, naive method could not perform well. In this paper, Authors proposed a direct importance estimation approach which does not involve density estimation. Since the studies showed that *Kullback-Leibler Importance Estimation Procedure* (KLIEP) probably outperform existing approaches in importance estimation including the logistic regression based method, and it provides to improving the prediction performance in covariate shift scenarios.

New Importance Estimation Method can be formulated as follows:

$$w(x) = \frac{p_{te}(x)}{p_{tr}(x)}$$

the key restriction is to avoid estimating densities $p_{te}(x)$ and $p_{tr}(x)$ when estimating the importance $w(x)$

Kullback-Leibler Importance Estimation Procedure (KLIEP) can be modeled the importance $w(x)$ by the following linear model:

$$\hat{w}(x) = \sum_{l=1}^{b} \alpha_l \rho_l(x)$$

where $\alpha_l^b{}_{l=1}$ are parameters to be learned from data samples and $\rho_l(x)_{l=1}^b$ are basis functions. Therefore, this paper showed method, called KLIEP, does not involve density estimation so it is more advantageous than a naive KDE-based approach particularly in high-dimensional problems. Comparing to KMM which also directly gives importance estimates, KLIEP is still more useful since it is equipped with a model selection procedure. Hence, KLEIP can demonstrate a promising approach for covariate shift adaptation.

In this study, [6] has experimented non-stationary environments (NSEs) where the input data distribution could be shifted over some certain period. Authors stated that the industrial process might be affected by the effect of NSEs although there is no shift has already occurred in the system. Somehow, the process that they have taken into the developing could still contribute the additional useful information to the system, and it could enhance the performance of the system in terms of system accuracy. For instance, there is the supervisor who can monitors and analyses the input information to detect the outliner (false positive) to the system, this types of information can conclude to the enhancing the system accuracy if we take the adequate correction. Then, when the expert label the data set before using for training or testing it could require much more human labour in the way of time and confusion that can occur during the tasks especially for the big data and real-time systems. There are several key points of adaptive mechanism for non-stationary systems as follows:

- the data samples must be intelligently collected for classifier parameter tuning and future use, if this is applicable
- the data from the current environment is treated as a representation of new knowledge, therefore it can be useful for adaptation
- the shift-detection or process monitoring mechanism is required to check the stationarity of the process
- pruning of irrelevant data is required to be finished and make sure there is no relevant information loss

Finally, they delevoped the classifiers with the common assumption which the data distribution is in the stationary state throughout the training and testing sets. The performance is however affected the in non-stationary conditions.

3 METHODOLOGY

3.1 Method

In this method of study, we will study and learn how to detect datasets (Sberbank Russian Housing Market) which are under covariate shift step-by-step as follows:

- Choose two different datasets which are under covariate shift or make the covariate shift environment condition. Training and testing sets must have a different distribution in the features.
- Train the algorithm without employing any techniques
- Predicting the variable of interest from training sets and compare the result of prediction to test set then, record the result for comparison
- Selecting the features which distribution are different between training and testing sets by using AUC-ROC method
- Comparing the features which have different distribution with important features and then keep the features which are important
- Train the algorithm with featured selection
- Test the algorithm with the same test data again and record the performance
- Comparing the result between without employing techniques and featured selection technique whether featured selection can do better performance or not

For Medical Cost Personal Datasets, we can see that there is no features under covariate shift. Hence, we make an experiment step as follows:

- making feature age by reordering young to old
- making different distribution of age feature in training and testing sets
- plotting histogram to see that there is a different distribution in age feature which can be seen in Figure 2.
- fitting randomforest algorithm with all features and record the result of RMSE
- dropping drifting feature (*age*) and again fitting randomforest algorithm
- record the result of RMSE again and compared the result each other

3.2 Dataset

Two selected datasets were downloaded from Kaggle webesite. Two datasets are as follows:

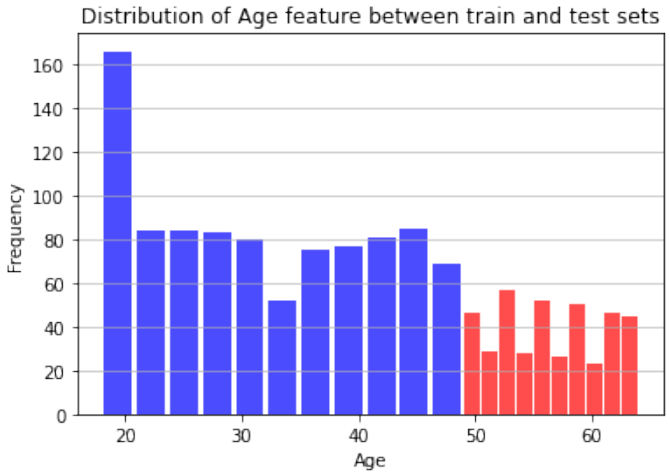

Fig. 2: The illustration shows different distribution of age feature between training and testing sets

3.2.1 *Sberbank Russian Housing Market*

The dataset was collected to help the investor by making predictions about realty prices so renters, developers, and lenders are more confident when they sign a lease or purchase a building. No one wants uncertainty about one of their biggest expenses. Training data is from August 2011 to June 2015, and test set is from July 2015 to May 2016. The dataset's features are 292 columns as follows: ID, timestamp, full_sq, life_sq, floor, max_floor, material,build_year, num_room, kitch_sq, cafe_count_5000_price_2500, cafe_count_5000_price_4000,cafe_count_5000_price_high, big_church_count_5000,church_count_5000, mosque_count_5000, leisure_count_5000,sport_count_5000, market_count_5000, price_doc, etc.

3.2.2 *Medical Cost Personal Datasets*

The dataset was collected to help the health institutions to predict their cost by predicting individual customers medical costs billed by health insurance. There are seven features as follows:

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg $/m^2$) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance x

where **charges** is variable of interest.

4 Results

The result getting from the experiments will be the number of *Root Mean Square Error* (RMSE) between employing feature selection technique and without applying any techniques. ROC-AUC was used to see which features have different distribution between training set and testing set in each columns.

From dataset of Sberbank Russian Housing Market, Firstly, we used randomforest regressor algorithm to predict the variable of interest (price_doc) without using any techniques. The result of RMSE is 0.40116 as shown in table 2. For using feature selection, we calculated ROC-AUC in each features of datasets to find a different distribution between features of training and testing sets. If the roc-auc score has more than 0.8, it means that those features are drifting. Since we calculated each features between training set and testing set, drifting features are shown in table 1. Next, we employed feature importance of randomforest classifier to show features which are important to price_doc. Then, we compared drifting features to important features. we can see that the features 'life_sq' and 'kitch_sq' are mutual. Hence, we kept these two features to predict **price_doc**. The result of predicting by dropping drifting features and keeping important features is 0.39759 which can be seen in table 2.

From Medical Cost Personal Datasets, the dataset was inspected which there is no any features under covariate shift. Hence, there will be two different scenarios of this dataset. First, the dataset was made to be under covariate shift in age feature which training set has only age between 18 to 49 and testing set has only age between 49 to 64 and other features remain the same. Then, Randomforest algorithm was fitted and predicted. The result of RMSE is 5748.18. For dropping drifting age feature, Randomforest algorithm was fitted and predicted again. So, it can be seen that the result of RMSE is higher than without dropping age feature which is 9074.57 as in table 3. Therefore, it can be said that dropping drifting feature of age cannot improve an accuracy of varible of interest (*charges*). Looking at figure 3, this can explain that is why dropping age feature did not improve an accuracy. The point is that *Age* feature is an important feature to variable of interest (***charges*** feature).

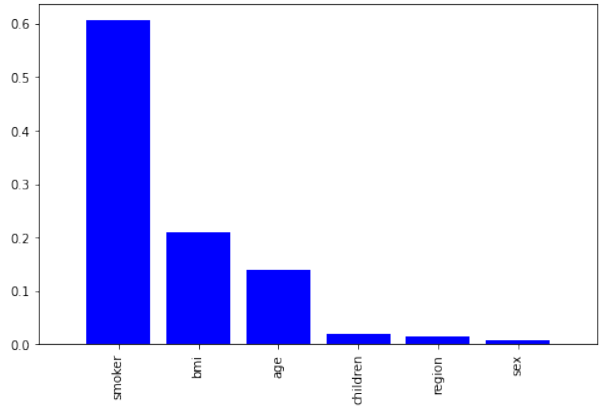

Fig. 3: Important features of Medical Cost Personal Datasets

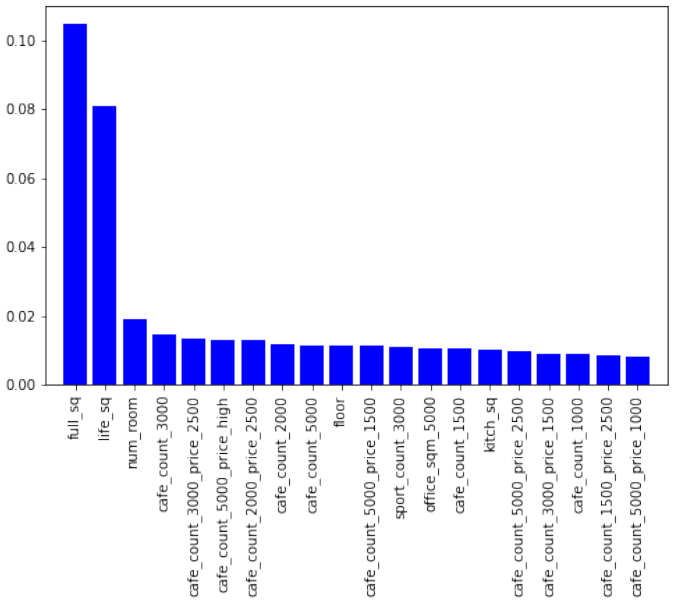

Fig. 4: Important features to price_doc from Sberbank Russian Housing Market dataset

TABLE 1: Features which are under covariate shift between training set and testing set of Sberbank Russian Housing Market dataset

Features	ROC-AUC score
id	1.0
timestamp	0.9144045285714286
life_sq	0.8455408285714286
max_floor	0.8680095571428572
material	0.8547409428571429
build_year	0.9152398857142856
num_room	0.8464369142857142
kitch_sq	0.9581987714285713
state	0.9094977714285715
hospital_beds_raion	0.8905239428571428
cafe_sum_500_min_price_avg	0.8579184142857144
cafe_sum_500_max_price_avg	0.8547958
cafe_avg_price_500	0.8563095857142857

TABLE 2: RMSE score between using feature selection technique and without using techniques of Sberbank Russian Housing Market dataset

Method	Feature selection	Without applying techniques
RMSE	0.39759	0.40116

5 Discussion

From the results we get as we plan in the previous section, we can see that dropping drifting features can improve an accuracy only when unimportant features are dropped. As we can see, From Sberbank Russian Housing Market dataset

TABLE 3: RMSE score between using feature selection technique and without using techniques of Medical Cost Personal Datasets

Method	Feature selection	Without applying techniques
RMSE	9074.57	5748.18

we dropped an unimportant features out and used only important features to fit the model and then predict the housing market price. Hence, the result of dropping feature have lower Root Mean Square Error (RMSE) than used all features so it means that dropping feature method did work. For Medical Cost Personal Datasets, we tried to drop important feature which is *age* and the result was not satisfied because age is the important feature to predict individual medical costs billed by health insurance (charges).

6 Conclusion

In this paper, we only experiment feature selection method to detect the feature which are under covariate shift between training and testing sets. In the future study, we will make an experiment different method of detecting covariate shift for example Importance Reweighting, Adversarial Search, Importance weighted cross-validation.

References

[1] Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N. D. (2009). Dataset shift in machine learning. The MIT Press.
[2] Moreno-Torres, J. G., Raeder, T., Alaiz-RodríGuez, R., Chawla, N. V., Herrera, F. (2012). A unifying view on dataset shift in classification. Pattern recognition, 45(1), 521-530.
[3] Shubham, Jain, 'Covariate Shift Unearthing hidden problems in Real World Data Science', (2017). [Online]. Available: https://www.analyticsvidhya.com/blog/2017/07/covariate-shift-the-hidden-problem-of-real-world-data-science/.
[4] Sugiyama, M., Krauledat, M., MÃžller, K. R. (2007). Covariate shift adaptation by importance weighted cross validation. Journal of Machine Learning Research, 8(May), 985-1005.
[5] Reddi, S. J., Poczos, B., Smola, A. (2015, February). Doubly robust covariate shift correction. In Twenty-Ninth AAAI Conference on Artificial Intelligence.
[6] Raza, H., Prasad, G., Li, Y. (2014, September). Adaptive learning with covariate shift-detection for non-stationary environments. In 2014 14th UK Workshop on Computational Intelligence (UKCI) (pp. 1-8). IEEE.
[7] Bickel, S., Brückner, M., Scheffer, T. (2009). Discriminative learning under covariate shift. Journal of Machine Learning Research, 10(Sep), 2137-2155.
[8] Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. V., Kawanabe, M. (2008). Direct importance estimation with model selection and its application to covariate shift adaptation. In Advances in neural information processing systems (pp. 1433-1440).