Model Testing using Datasets with artificial Dataset Shift

Anonymous Author(s)
Affiliation
Address
email

Abstract

The generalization capability of machine learning models depends significantly on their ability to handle dataset shift. Unfortunately, testing a models for this ability is not straightforward. In this paper, we propose several dataset splits for existing real-world regression and classification datasets. Our splits are designed to create different distributions between training and test set, which resembles dataset shift occurring in real-world problems. Therefore, using our proposed dataset splits, models can be directly tested for their capability of handling several types of dataset shift. We provide the datasets as well as the proposed dataset splits as a python package, which can be accessed conveniently. By that, we hope that the proposed splits become a standard benchmark for testing generalization capabilities under dataset shift.

1 Introduction

The aim of a machine learning model is to generalize to new situations. Unfortunately, testing a model for its ability to handle new situations is not straightforward. It is often unclear how robust models behave in new environments or whether they can be transferred to other applications [11]. This is due to the fact that the distribution of the training data might differ from the distribution of the data with which a model is tested. These differences between training and testing data are referred to as dataset shift [9]. Dataset shift describes the phenomenon, that conditions under which the model is developed might differ compared with the conditions of the final system. There are numerous reasons why dataset shift can happen and for real-world tasks dataset shift is rather the rule than the exception.

In this paper, we propose artificially created dataset splits for several real-world regression and classification tasks, which take into account the most common forms of dataset shift in real-world applications. Using the proposed splits, machine learning models can be tested directly for their capability of handling data set shift. By that, it can be identified to which shifts a certain model is vulnerable and its applicability to a certain real-world task can be estimated. We developed the splits in order to provide multiple examples of different types of dataset shift occurring frequently in real-world problems. Therefore, if the target application domain is known, a model can be evaluated using the sets designed for the respective dataset shift.

Our data splits are provided as a Python-Package together with the corresponding datasets. This ensures that the dataset splits can be integrated into a machine learning pipeline with hardly more than 2 lines of code. Moreover, we provide scoring functions, which can be used to calculate the corresponding score on each dataset (Hand [6] argued that providing standard scoring functions can increase the comparability of results achieved on the respective dataset). We hope that by making the splits conveniently accessible, they become a new standardized benchmark for supervised learning tasks. This makes reported results more robust and better comparable.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

The remainder of this paper is organized as follows. We first recapitulate the definitions of dataset shift and its major types from literature. In the following section, we introduce our use-cases, which represent various tasks prone to dataset shift. Subsequently, we evaluate the performance of three baseline models (namely neural networks, extreme learning machines and linear models using ridge regression) for our use-cases and show that measuring the performance on randomly split datasets is over-optimistic for the considered tasks. Finally, we summarize our findings and give a conclusion.

2 Background

In real-world applications, the conditions under which a system is tested usually differ from the conditions under which it was developed [9]. Translated to the machine learning domain, this should lead to the setting that the test set differs in some way from the dataset used for training. Therefore, the aim is to make proper predictions in one environment, whereas we only have data about another, second environment. Usually, the two environments are closely related (obviously, if they differ too much, we can not infer how to make predictions in the first environment given only the data from the second environment).

In [8], Moreno-Torres et al. propose three main types of dataset shift occurring in typical real-world problems. They focus on classification problems, which are defined by a set of features x, a target variable y and a joint distribution $P(y, x)$. There are two different kinds of problems. First, there are problems in which the target variable is causally determined by the values of the features $x \rightarrow y$. This is for example the case in credit card fraud detection, where the behavior of the user determines its class label. In the second kind of problems, the target variables causally determine the values of the features $y \rightarrow x$. A typical example for this kind of problems is medical diagnosis, where the diagnosed disease determines the symptoms we observe. The three major types of dataset shift for these kinds of problems are

- **Covariate Shift**: Appears only in $x \rightarrow y$ problems (in an ideal setting). Covariate Shift refers to the case that the distribution of the input variables x change, i.e. $P_{train}(y|x) = P_{test}(y|x)$, but $P_{train}(x) \neq P_{test}(x)$. Note that machine learning model models, in theory, should not be affected by covariate shift [9]. However, in practice, covariate shift often leads to deteriorated results.
- **Prior Probability Shift** appears only in $y \rightarrow x$ problems (in an ideal setting). It refers to a shift in the distribution of the target variable y, i.e. $P_{train}(y|x) = P_{test}(y|x)$, but $P_{train}(y) \neq P_{test}(y)$.
- **Concept Shift** arises in both kinds of problem settings. Concept shift describes scenarios, where the relationship between features and target variables is different in the test set compared to the training set. For $x \rightarrow y$, concept shift is defined as $P_{train}(y|x) \neq P_{test}(y|x)$ and $P_{train}(x) = P_{test}(x)$, and for $y \rightarrow x$ problems $P_{train}(y|x) \neq P_{test}(y|x)$ and $P_{train}(y) = P_{test}(y)$ respectively. This is the type of dataset shift which is most challenging.

Moreno-Torres et al. mention that there are also other types of dataset shift, but these types hardly occur in real-world problems. Moreover, they are so hard to resolve that we currently consider them impossible to solve [8]. Therefore, we focus on the listed three major types.

3 Use-Cases

In this paper, we propose several datasets which can be used for testing the capability of a model to handle dataset shift. The proposed datasets are based on existing real-world regression and classification datasets collected from the UCI Machine Learning Repository [4]. Our dataset splits are designed to resemble real-world scenarios which are prone to dataset shift. An overview over all use-cases together with the types of dataset shift we assume in the respective case is given in Table 1. Moreover, Table 2 shows the statistics of the datasets for which our use-cases are designed.

Character Font Images *A model for character recognition should generalize to new fonts.* The dataset provides numerous examples of several characters written in different fonts and scanned from various devices such as hand scanners, desktop scanners or cameras. Achieving a high score on

Table 1: Overview over the proposed use-cases. For each case we indicate if we assume covariate shift (Cov.-S.), prior probability shift (PPS) or concept shift (Con.-S.) in the data.

Use Case	Cov.-S?	PPS?	Con.-S?
Character Font Images	Yes	No	No
Pen-Based Digits Recognition	Yes	No	No
Simulated Electrical Grid Stability	Yes	No	No
Parkinson Speech	Yes	No	No
Hand Postures	Yes	No	No
Wine Quality	No	Yes	No
Polish Companies Bankruptcy	No	Yes	Yes

Table 2: Overview over the statistics of the datasets used.

Use Case	Dim.	# Train	# Val	# Test
Character Font Images	400	239,766	9,460	115,864
Pen-Based Digits Recognition	16	5,995	1,499	3,498
Simulated Electrical Grid Stability	11	6,400	1,600	2,000
Parkinson Speech	26	758	206	244
Hand Postures	36	39,118	13,545	25,432
Wine Quality	11	2,638	1,085	1,075
Polish Companies Bankruptcy	64	27,703	5,910	5,910

the test set means that your model generalized well and learned the underlying task of recognizing characters - independent of their particular appearance.

Pen-Based Digit Recognition *A model for recognizing digits should generalize to new individuals drawing the digits.* The aim is to learn which digits have been drawn based on resampled coordinates of the drawing process [1]. The test set consists of digits drawn by individuals who did not contribute samples to the training set. Hence, in this use-case, it is tested if a model which recognizes digits generalizes to new individuals drawing digits.

Simulated Electrical Grid Stability *A model predicting electrical grid stability should generalize to new regions with other consumer behavior.* Based on simulated data of electrical grid stability [2], the aim of this use-case is to virtually test if models trained in one region generalize to another region with different consumer behavior. The given data is split such that in the validation and testing set more energy is consumed than in the training set (e.g. as is the case if you trained your model on data from a residential area and would like to test it on data of an industrial area).

Parkinson Speech *A model predicting if an individual suffers from the parkinson disease should generalize to new patients.* The dataset used for this use-case [10] consists of voice features of several sound recordings. The aim is building a model, which predicts if the voice features of an individual indicate that he suffers from the parkinson disease. The test set consists of other individuals than the ones recorded for the training set. Therefore, in this use-case, it is tested if the model correctly predicts the status of new patients. Note that even if this is a $y \to x$ problem, we refer to the shift in this dataset as covariate shift due to the finite number of individuals in the dataset.

Motion Capture Hand Postures *A model for hand posture recognition should generalize to new individuals performing postures.* Based on the Motion Capture Hand Postures dataset [5], the aim of this use case is to predict the correct hand posture given the coordinates of 11 markers. Note that in the dataset, there are many missing values and the marker positions have been permuted between different recordings. The test set consists of hand postures performed by other individuals than the ones in the training set. Hence, this use-case tests the capability to recognize postures of new individuals.

Wine Quality *A model for predicting wine quality should correctly predict test wines which have been selected biased towards high-quality wines.* The challenge of this use-case is to train a model to predict wine quality [3] using examples of the full quality spectrum. After training, the testing is carried out using the wines provided by an upper-class wine merchant. Therefore, this use-case tests the capability of handling prior probability shift, since we assume that the merchant tends to have higher-quality wines. Note that this is a virtual scenario and the wines from the testing set are not actually provided from a wine merchant. Instead, we split the training, validation and test set in order to achieve certain characteristics.

Polish Companies Bankruptcy *A model predicting if a company goes bankrupt trained on historical data should generalize to data acquired more recently.* The aim of this use case is to build a model, which predicts if a company goes bankrupt within one year (based on the Polish Companies Bankruptcy dataset [12]). As usual for real-world tasks, you can train your model solely on historical data. The test set, however, consists of data acquired more recently. We assume that whether a company goes bankrupt depends on the economic environment, which changes over time. Therefore, this use-case tests if your model can handle concept shift between training and testing data.

4 Experiments

In our experiments, we compare test scores obtained using our dataset splits with scores obtained on randomly split data. For that, we train three baseline models on each of the datasets. These models are

1. Extreme Learning Machine: An Extreme Learning Machine proposed by Huang et al. [7]. We are using 200 neurons, the sigmoid function as nonlinear activation and ridge regression to determine the output weights (with regularization hyperparameter $\lambda = 0.001$).
2. Neural Network: A simple neural network with 2 hidden layers, 200 neurons in each of the hidden layers and the ReLU function as nonlinear activation. The model is trained until convergence on a randomly split validation set using the Adam optimizer.
3. Ridge Regression: Ridge Regression as linear baseline compared to the other, nonlinear methods. We use $\lambda = 0.001$ as hyper parameter to weight the regularization term.

All hyperparameters and the code for the baseline experiments are available online. In order to account for stochastic effects, we repeated every experiment 30 times (the individual outcome of each experiment can be accessed in the python package).

The neural network achieves best performance for most of the use-cases, followed by the Extreme Learning Machine. Even if the model obtained by ridge regression is simple compared to the other approaches, it achieves remarkable results on most of the datasets. For all comparisons, the baseline models achieved better test-performances when the data was split randomly in contrast to using our dataset splits. This shows that for our data splits, the distribution between training and testing sets differs, which represents the induced artificial dataset shift. Therefore, our splits can be used to test a model for its capability to cope with dataset shift. The outcomes of the experiments are depicted in the supplementary materials.

5 Conclusion

In this paper, we proposed dataset splits for various real-world regression and classification tasks. Our use-cases are based on several existing real-world datasets with different characteristics. Moreover, we developed the splits such that the resulting use-cases feature different types of dataset shift that occur in real-world scenarios. The proposed splits can be used to test the generalization capabilities of a model under dataset shift.

We provided a convenient way of accessing the proposed datasets and splits in order to maximize the accessibility. Furthermore, we showed for all our proposed use-cases that reporting the performance on a randomly split dataset leads to over-optimistic results for the respective target application. The results and code of our experiments are publicly available and can be used as baselines for future work using our dataset splits.

References

[1] F. Alimoglu and E. Alpaydin. Methods of combining multiple classifiers based on different representations for pen-based handwritten digit recognition. In *Proceedings of the Fifth Turkish Artificial Intelligence and Artificial Neural Networks Symposium (TAINN 96)*. Citeseer, 1996.

[2] V. Arzamasov, K. Böhm, and P. Jochem. Towards concise models of grid stability. In *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 1–6. IEEE, 2018.

[3] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.

[4] D. Dua and C. Graff. UCI machine learning repository, 2017.

[5] A. Gardner, J. Kanno, C. A. Duncan, and R. Selmic. Measuring distance between unordered sets of different sizes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 137–143, 2014.

[6] D. J. Hand. Classifier technology and the illusion of progress. *Statistical science*, pages 1–14, 2006.

[7] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[8] J. G. Moreno-Torres, T. Raeder, R. Alaiz-RodríGuez, N. V. Chawla, and F. Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

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

[10] B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun. Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics*, 17(4):828–834, 2013.

[11] L. Torrey and J. Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global, 2010.

[12] M. Zikeba, S. K. Tomczak, and J. M. Tomczak. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*, 2016.