Will the sun shine? – An accessible dataset for teaching machine learning and deep learning

Florian Huber [1] **Dafne van Kuppevelt** [2] **Peter Steinbach** [3] **Colin Sauze** [4] **Yang Liu** [2] **Berend Weel** [2]

Abstract

Hands-on teaching of modern machine learning and deep learning techniques heavily relies on the use of well-suited datasets. We present a novel tabular dataset that was specifically created for teaching machine learning and deep learning to an academic audience. The dataset contains intuitively accessible weather observations from 18 locations in Europe. It was designed to be suitable for a large variety of different training goals and to avoid reaching unrealistically high prediction accuracy. Teachers or instructors thus can choose the difficulty of the training goals and thereby match it with the respective learner audience or lesson objective. The compact size of the dataset makes it possible to quickly train common machine learning and deep learning models on a standard laptop so that it can be used in live hands-on sessions.

1. Introduction

We will start with the obvious: Modern machine learning (ML) and deep learning (DL) techniques have rapidly gained popularity in many different fields of academic research, culture, and society as well as in many industries. ML & DL allow stakeholders to handle larger datasets, or make entirely new applications possible. While some expectations towards machine learning might prove more difficult than expected (Mitchell, 2021), it is becoming clear that ML and DL techniques are here to stay. Currently, they represent a heavily sought-after skill set in many areas of society (Lee et al., 2019). This is also reflected by a growing number of online courses and tutorials, as well as many new university degrees related to ML & DL.

As ML & DL are data-driven by definition, the use of suitable datasets is key to teaching it. At the time of writing, there is an incomprehensible amount of public datasets available that can be used to train machine learning models. But making a good choice in this overwhelming abundance of data is far from trivial. When designing hands-on teaching material for deep learning, we noticed that a large part of the available datasets has strong limitations, such as not being suited for demonstrating typical difficulties that commonly occur when applying deep learning techniques. One reason for this is likely that many of the most commonly used datasets are geared towards basic, rather simple, machine learning tasks (e.g. Iris dataset (Bezdek et al., 1999)). While DL can be applied to such data, many more advanced concepts are difficult to show (e.g. overfitting, unbalanced data). This can easily raise unrealistic expectations and make it hard to move from tutorial material to actual real-world problems.

Another issue in teaching situations is that many datasets require domain knowledge to understand what the included features mean. This can for instance be botanical terms (e.g. Iris dataset (Bezdek et al., 1999) or mushroom dataset (Society, 1981; Dua & Graff, 2017)) or medical terms (e.g. Parkinson dataset (Little et al., 2009)). Other datasets contain very abstract features which are either hard to interpret for non-experts or simply have no intuitive explanation at all such as the popular credit card fraud detection dataset (Carcillo et al., 2021) (https://www.kaggle.com/mlg-ulb/creditcardfraud). Again other datasets – some of them commonly used for tutorials – come with hidden ethical issues, such as the Boston housing dataset (Chen, 2021; Carlisle, 2020).

Notable exceptions are image-based datasets, which tend to be more intuitively accessible and also exist in all kinds of flavors and difficulty levels (Voulodimos et al., 2018).

Our goal here is to create a tabular dataset that is well suited for hands-on courses and tutorials on machine learning as well as deep learning. Ideally, this means the dataset com-

[1] Centre for Digitalization and Digitality, University of Applied Sciences Düsseldorf, Düsseldorf, Germany [2] Netherlands eScience Center, 1098XG Amsterdam, the Netherlands [3] Helmholtz AI, Helmholtz-Zentrum Dresden-Rossendorf, Core Facility for Information Services and Computing, Bautzner Landstr. 400, 01328 Dresden, Germany [4] Department of Computer Science, Aberystwyth University, Penglais, Aberystwyth, SY23 3DB, United Kingdom. Correspondence to: Florian Huber <florian.huber@hs-duesseldorf.de>.

Proceedings of the 3^{rd} Teaching in Machine Learning Workshop, PMLR, 2022. Copyright 2022 by the author(s).

bines

- Cognitively easily accessible. Without specialized domain knowledge, students can very quickly understand what the data is about and what the features and labels mean.
- Small enough to allow for fast import, processing, and model training even on a standard laptop.
- Complex enough to show frequently faced difficulties in applying deep learning, such as overfitting or unbalanced data.
- Realistic and related to an actual (scientific) problem setting
- Flexible use-cases so that the same dataset can be used for multiple tasks which helps to lower the cognitive load of handling several different datasets in a workshop setting. In our case: Regression and classification task, single label and multi-label classifications, possible use for time series classification, regression, and forecasting tasks.

We target activities for learners at the undergraduate as well as post-graduate level in research and academia.

1.1. Why weather data?

Weather data is something that everyone is familiar with, for instance from daily weather forecasts. Observational variables like daily minimum or maximum temperature or sunshine hours do not need any further introduction and impose less cognitive load on the learner. At the same time, weather data represent actual scientific measurements and can immediately be linked to interesting scientific use-cases or stories. In addition, weather data contains a rich selection of typical phenomena that researchers or data scientists will frequently face when handling data of various origins. Some variables such as the daily mean temperature expose well-behaving float entries with a characteristic nearly symmetrical distribution due to the diurnal cycles. Other variables, such as sunshine hours or precipitation, display highly skewed distributions. Precipitation values, for instance, can even be seen as a mixture of categorical and continuous values since many days show no precipitation at all.

2. Construction of the weather prediction dataset

The initial meteorological data was retrieved from ECA&D (Klein Tank et al., 2002) - a project that publicly provides daily observations at meteorological stations throughout Europe and the Mediterranean. 18 different European cities or locations were selected from this dataset of which multiple daily observations were available through the years 2000 to 2010. These locations were: Basel (Switzerland), Budapest (Hungary), Dresden, Düsseldorf, Kassel, München (all Germany), De Bilt and Maastricht (the Netherlands), Heathrow (UK), Ljubljana (Slovenia), Malmo and Stockholm (Sweden), Montélimar, Perpignan and Tours (France), Oslo (Norway), Roma (Italy), and Sonnblick (Austria). We provide the raw downloaded data of all these locations from ECA&D on zenodo: https://doi.org/10.5281/zenodo.4964287.

Recordings of daily meteorological observations for these 18 locations in the original dataset span different time ranges. Some contain collections that go back to the 19th century. For this ML dataset, we only selected entries from the years 2000 to 2010 resulting in 3654 daily observations per location. The dataset is eventually obtained by merging data from all 18 locations. The data in addition consists of different observations. All selected locations provide data for the variables ‘mean temperature’, ‘max temperature’, and ‘min temperature’. Wherever available, we also included data for the variables ’cloud_cover’, ’wind_speed’, ’wind_gust’, ’humidity’, ’pressure’, ’global_radiation’, ’precipitation’, ’sunshine’. After collecting and merging the data, further cleaning was performed: Columns containing features with $> 5\,\%$ invalid entries (“-9999”) were removed, and columns with $\leq 5\,\%$ invalid entries where kept - invalid entries therein were replaced by mean values. For all 18 locations, this resulted in a total of 165 variables (or features) over the course of 3654 days.

Finally, we transformed several data units to achieve numerical ranges of the present values which are more similar to each other. This makes the data more suitable to be used for ML or DL even without additional (pre-)processing. We deliberately did not choose to fully standardize the data as we aspire to keep the presented units and values as intuitively accessible as possible. As such, temperature is given in $^{\circ}\mathrm{C}$, wind speed and gust in $\mathrm{m/s}$, humidity in fraction of $100\,\%$, sea level pressure in units of $1000\,\mathrm{hPa}$, global radiation in $100\,\mathrm{W/m^2}$, precipitation amounts in cm, sunshine in hours.

The dataset is available under:
https://doi.org/10.5281/zenodo.7053722

3. Example use-cases for teaching ML and DL

The data is collected and processed such that multiple key aspects in ML and DL can be addressed. It also leaves room to choose different levels of complexity or difficulty. In the following, we will describe three examples to illustrate how we believe the dataset could be used in teaching ML or DL.

3.1. Binary classification – picnic weather tomorrow?

We provide an extra set of labels in a separate file that allows learners to work on the following binary classification task: is tomorrow going to be the right weather for a picnic or not? We defined this criterion as a day with $< 0.1\,\mathrm{mm}$ precipitation and a maximum temperature $\geq 18\,°\mathrm{C}$. Training a random forest (scikit-learn (Pedregosa et al., 2011)) on data of the first 3 years to predict if the next day will be suited for having a picnic generally performs well. A simple random forest with 50 trees of maximum depth of 6 was correct in predicting picnic weather for the next day in Düsseldorf (Germany) in 3 out of 4 cases.

3.2. Multi-class classification – which month is it?

This might at first not seem like a very plausible task to do. Why predict the month from weather data? But many people will know the sensation of experiencing weather conditions that do not seem appropriate for a particular time of the year, for instance, a particularly warm day in early April, or a frustratingly cold day a week later. Could a classical machine learning model or a neural network learn to predict the correct month based on only the daily weather observations of a single day? Using a random forest classifier (scikit-learn (Pedregosa et al., 2011)) the performance on the reserved test set is shown in fig.1.

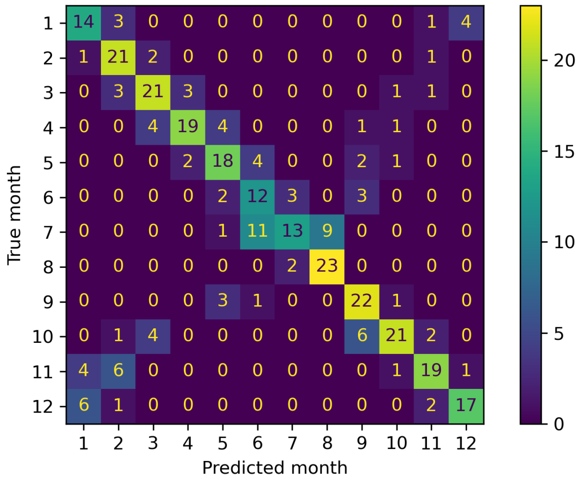

Figure 1. Confusion matrix to evaluate a model's performance. It here compares the month predicted based on daily weather data using a random forest classifier with the true month. We here used a random forest with 10 trees of maximum depth=5. The colorbar and the displayed values are counts within the test set.

3.3. Regression task – How many hours of sunshine will we have at location X tomorrow?

Arguably, the most common type of regression related to weather data is the prediction of weather in the future. A rather approachable way to do this is to predict the sunshine hours in Basel (or any other place in the dataset) for tomorrow based on the weather data of all 18 locations today.

As the data is sorted using the 'DATE' column as key, labels can be created by taking the sunshine hours for the following day, e.g. using the following pandas code (pandas development team, 2020):

```
data.loc[1:(365*3 + 1)]["BASEL_sunshine"]
```

Date and month should be removed from the training data while all other columns can be used without further processing.

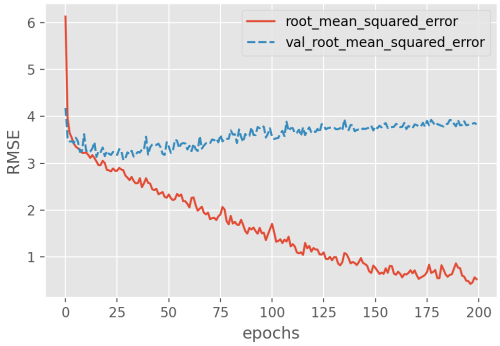

Figure 2. Model performance on predicting tomorrow's sunshine hours as measured by the root mean squared error (RMSE) on the training and validation set during training process showing a clear signature of severe overfitting. Red line: error on training set. Blue line: error on validation set.

As an example, we trained a neural network consisting of three dense layers: input (163 features) → dense layer (100 nodes, ReLU) → dense layer (50 nodes, ReLU) → ouput: dense layer (1 node, no activation function). We further use an Adam optimizer (Kingma & Ba, 2014) and a mean squared error (MSE) loss. This model converges well during training, but also shows drastic overfitting (fig.2).

Overfitting is a very common obstacle in machine learning and deep learning, so this use case can serve as a realistic example of how to identify and avoid this. In a first workshop setting, this scenario was used to illustrate why the training process in deep learning is generally being monitored with a separate validation set. This allows to detect overfitting, and by also using 'Early Stopping' even to prevent severe overfitting. In the here presented case, handling overfitting is far from trivial, which in our view makes it a realistic training example. Different techniques can be applied to improve the results, such as regularization techniques (e.g. dropout (Hinton et al., 2012)), batch normalization, adaptations of the model architecture (e.g. fewer nodes), or applying data augmentation.

For more advanced teaching units, we could conceive to introduce methods to estimate the model uncertainty. One common way to assess the model uncertainty is to activate the dropout layers during prediction to sample predictions from an ensemble of models, termed Monte-Carlo Dropout (Gal & Ghahramani, 2016).

To illustrate this, we trained a model similar to the three-layer model described above. In this case, however, we devised a few attempts to improve performance: an initial batch normalization layer, as well as two dropout layers (dropout rate of 0.2), a third hidden layer (10 nodes), and early stopping. In addition, we trained it on the full dataset. Even with the mentioned improvements, the model prediction of the sunshine hours for the next day is still not overly impressive, see figure 3.

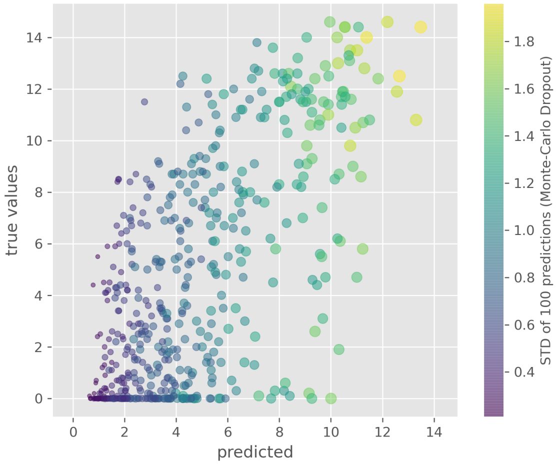

Figure 3. Predicted sunshine hours for the next day in Basel with uncertainty estimation through Monte-Carlo Dropout (circle size + color code give the standard deviation of 100 predictions)

Other properties are easier to predict for the following day, for instance, the global radiation which is related to the sunshine hours but apparently shows much more consistent (hence predictable) behavior. Even more consistent and correlated and thereby better to predict are the temperature data, for instance, the daily maximum temperature (see fig. 4).

4. Conclusions and outlook

We present a novel tabular dataset that is specifically targeted at teaching machine learning and deep learning to an academic audience. The core aim was to combine versatile usability for a wide range of teaching goals with small dataset size ($< 3Mb$). Moreover, the features are intuitively accessible as they are commonly used in daily weather reports, such as minimum and maximum temperature, wind speed, or sunshine hours.

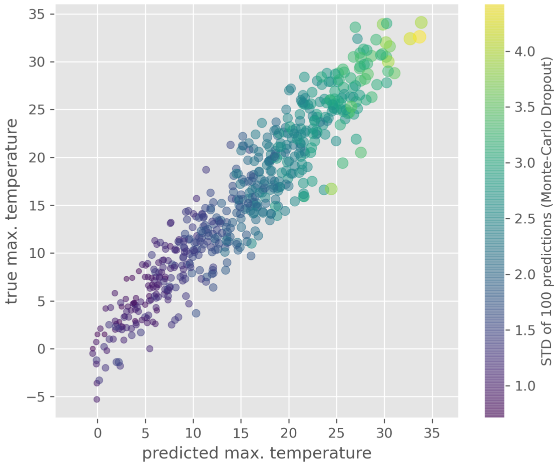

Figure 4. Predicted maximum temperature in Celsius for the next day in Basel with uncertainty estimation through Monte-Carlo Dropout (circle size + color code give the standard deviation of 100 predictions)

We imagine that the dataset can be used for training undergraduate students as well as professional researchers or data scientists. The dataset further allows defining a variety of ML-related training goals, many of which are not easily giving way to unrealistically high prediction accuracy. Teachers or instructors can choose the difficulty of the training goals and thereby match it with the respective learner audience or lesson objective.

In this manuscript, we sketched a few possible tasks that could be covered with the dataset. We applied both ”classical” machine learning (random forest) as well as deep learning techniques. First training use-cases have already been explored in practice by the authors in two hands-on training sessions for researchers, one on machine learning and one on deep learning.

The compact size and complexity of the dataset makes it possible to quickly train common ML and DL models on a standard laptop so that it can used in live hands-on sessions.

5. Links to dataset and code

- The ”weather prediction dataset” for teaching ML and DL presented in this article can be found on zenodo: `https://doi.org/10.5281/zenodo.7053722`

- The raw data (used as input) was downloaded from

ECA&D (Klein Tank et al., 2002) in April 2021 can be found on zenodo: https://doi.org/10.5281/zenodo.4964287

- Juptyer notebooks for the creation of the dataset from ECA&D data as well as notebooks for the training of ML and DL models as well as all plots presented in this article can be found on a dedicated GitHub repository: https://github.com/florian-huber/weather_prediction_dataset

Acknowledgement

We acknowledge the data providers in the ECA&D project. Klein Tank, A.M.G. and Coauthors, 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. Int. J. of Climatol.,22, 1441-1453. Data and metadata are available at https://www.ecad.eu. Peter Steinbach was funded by Helmholtz Association's Initiative and Networking Fund through Helmholtz AI.

References

Bezdek, J. C., Keller, J. M., Krishnapuram, R., Kuncheva, L. I., and Pal, N. R. Will the real iris data please stand up? *IEEE Transactions on Fuzzy Systems*, 7(3):368–369, 1999.

Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., and Bontempi, G. Combining unsupervised and supervised learning in credit card fraud detection. *Information Sciences*, 557:317–331, May 2021. ISSN 0020-0255. doi: 10.1016/j.ins.2019.05.042. URL https://www.sciencedirect.com/science/article/pii/S0020025519304451.

Carlisle, M. racist data destruction?, January 2020. URL https://medium.com/@docintangible/racist-data-destruction-113e3eff54a8.

Chen, J. M. An Introduction to Machine Learning for Panel Data. *International Advances in Economic Research*, 27 (1):1–16, February 2021. ISSN 1573-966X. doi: 10.1007/s11294-021-09815-6. URL https://doi.org/10.1007/s11294-021-09815-6.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Klein Tank, A., Wijngaard, J., Können, G., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., et al. Daily dataset of 20th-century surface air temperature and precipitation series for the european climate assessment. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 22(12):1441–1453, 2002. ISSN 1097-0088. doi: https://doi.org/10.1002/joc.773. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.773.

Lee, P., Stewart, D., Loucks, J., and Arkenberg, C. Technology, Media, and Telecommunications Predictions 2019. Technical report, 2019.

Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., and Ramig, L. O. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE transactions on bio-medical engineering*, 56(4): 1015, April 2009. ISSN 0018-9294. doi: 10.1109/TBME.2008.2005954. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3051371/.

Mitchell, M. Why AI is Harder Than We Think. *arXiv:2104.12871 [cs]*, April 2021. URL http://arxiv.org/abs/2104.12871. arXiv: 2104.12871.

pandas development team, T. pandas-dev/pandas: Pandas, February 2020. URL https://doi.org/10.5281/zenodo.3509134.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Society, N. A. *National Audubon Society Field Guide to North American Mushrooms*. Alfred A. Knopf, 1981.

Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience*, 2018:e7068349, February 2018. ISSN 1687-5265. doi: 10.1155/2018/7068349. URL https://www.hindawi.com/journals/cin/2018/7068349/. Publisher: Hindawi.