170008773

16th March 2018

Abstract

Keywords:

1 Introduction

In this report we were asked to desing and impolement a regression model to predict energy consumption in resident houses. In this report we will mirror the work of (Candanedo, Feldheim & Deramaix, 2017) in several ways, but depart from it in others which we will discuss now. Like Candanedo et al. we assume that it is desirable to better understand the impact different features have on energy consumption for prediction of other important phenomena, such as but not limited to "determine[ing] adequate sizing of photovoltaics and energy storage to diminsih powre flow into the grid [and] to detect abnormal energy use patterns" (Candanedo et al., 2017). Therefore we will develop several regression models that can tell us more about the impact certain features have on the overal energy usage. Here we will use three algorithms: `AdaBoost`, Random forests and polynomial regression with regularisation. See section 2.4 for a more indepth explanation.

2 The learnign process

2.1 Cleaning the data

2.2 Analysing the data

2.3 Feature selection

This is relatively straight forward. Since part of our objective is to discover underlying relations betweeen the features and the regression target, we will use all the features available in our analysis. It would however be possible that certain extracted features would yield better performance but this is unfortunately ouside the scope of this report.

2.4 Selecting and training the model

Models Since it is our goal to gain a better understanding of how the various features impact the overal energy usage, it behoves us to pick models that can easily provide this information. A lot of regression models, like Neural Networks do not provide this information natively. This would defeat a large part of the

purpose of our research. We will discuss the models we chose to implement below and why they are suited to this task.

Metric We must furthermore select a metric to judge our model by. Again we wanted to use a metric that was supported by our libraries, so this cut down the number of possibbilities considerably. We also wanted a metric that would let us interpret the results easily, since we are not just interested in raw performance, but in interpretability. This narrowed the possibilities down to the Root-Mean-Squared-Error (RMSE) or the Mean-Absolute-Error (MAE). We eventually elected MAE since this provides a more neutral view of the accuracy (Willmott, Matsuura & Robeson, 2009).

2.4.1 Polynomial regression with regularisation

Polynomial regression does not natively provide very good information about the relative importance of features. This can be done however, using a combination of normalisation and regularisation. Regularisation is originally ised to combat overfitting in polynomial regression. It does this by reducing the scale of certain features and enlarging other ones. If we make sure to properly normalise all of the features before training, then the regularisation doesn't have to compensate for scale. This means that after training we can identify important features by looking at the regularisation coefficients. Unreliable features will have small regularisation coefficients while important ones will have large ones. We could have done this with linear regression, but after examinening the data we found that the correlation between each individual feature was relatively low, leading us to conclude that linear regression would not suffice. We therefore decided to use quadratic regression.

2.4.2 AdaBoost

`AdaBoost` is an ensemble method developed in (Freund & Schapire, 1997). Perhaps counter-inuitively `AdaBoost` is a sample selection algorithm instead of a feature selection algotihm. It works to identify "hard examples" and works to improve accracy on those, under the assumption that that will also boost performance on the "easy examples". The way that this can still provide insight into the feature space, is that we can look at the similarities between the "easy examples". If for example all the easy examples have a large magnitude in one feature, we can deduce that that featre has a large impact. While this deduction is not a native it can yield much more insight than other black-box regression methods.

Choice of Boosting algorithm `AdaBoost` is part of a class of ensemble methods called boosting algorithms. While `AdaBoost` is an older algorithm, and has worse performance than newer algorithms in this catagory (e.g. NH-Boost.DT or Squint-boost (Luo & Schapire, 2014; Otten, 2016; Vente, 2016)), it will serve for our purpose. We decided to use this one because it is supported in scikit-learn (Pedregosa et al., 2012) and implementing the other regression algorithms from scratch is outside the scope of this report. The astute reader will remark that random forests (discussed below) were developed in response to `AdaBoost` but we included it anyway to determine the viability of using a boosting algorithm in the way we described.

2.4.3 Random Forests

Random forests is another ensemble method developed by (Breiman, 2001). He writes "Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. T[Random Forests] use a random selection of features to split each node [...]

2.5 Evaluating the model

2.6 Discussing the results

3 Conclusion

word count:

References

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324. arXiv: /dx.doi.org/10.1023{\%}2FA{\%}3A1010933404324 [http:]

Candanedo, L. M., Feldheim, V. & Deramaix, D. (2017). Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, *140*, 81–97. doi:10.1016/j.enbuild.2017.01.083

Freund, Y. & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1, part 2), 119–139. doi:https://doi.org/10.1006/jcss.1997.1504

Luo, H. & Schapire, R. E. R. (2014). A Drifting-Games Analysis for Online Learning and Applications to Boosting. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 1368–1376). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/5549-a-drifting-games-analysis-for-online-learning-and-applications-to-boosting

Otten, H. (2016). A theoretical examination of boosting algorithms. Retrieved from https://www.universiteitleiden.nl/binaries/content/assets/science/mi/scripties/otten.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. *... of Machine Learning ...* *12*, 2825–2830. Retrieved from http://dl.acm.org/citation.cfm?id=2078195%7B%5C%%7D5Cnhttp://arxiv.org/abs/1201.0490

Vente, D. (2016). *A practical comparison of boosting algorithms* (Bachelor Thesis, Univeristy of Leiden). Retrieved from https://www.universiteitleiden.nl/binaries/content/assets/science/mi/scripties/vente.pdf

Willmott, C. J., Matsuura, K. & Robeson, S. M. (2009). Ambiguities inherent in sums-of-squares-based error statistics. *Atmospheric Environment*, *43*(3), 749–752. doi:10.1016/j.atmosenv.2008.10.005