Ethics of Data Science – Part III

Week 2: Reproducibility and Robustness

Champion/challenger deployment pipeline

Dr. Chris Anagnostopoulos, Hon. Senior Lecturer

Stability of Stochastic Gradient Descent

Input:

- Training data $\mathbf{x_1}, \ldots, \mathbf{x_n}$ (often augmented non-deterministically, e.g., randomly flip an image)
- Initialisation $\boldsymbol{\theta_0}$ and hyperparameters (e.g., learning rate) – sometimes initialized randomly
- Order of presentation of training data to algorithm $\mathbf{i_1}, \ldots, \mathbf{i_n}$ – sometimes shuffled randomly

Internal sources of randomness:

- Potential random steps during training (e.g., simulated annealing, or drop out)
- Lack of determinism in underlying primitives (e.g., GPU / multi-threading)
- Numerical stability (e.g., in vanishing or exploding gradients)

Ways to measure output variability:

- Variability of final weights $\boldsymbol{\theta_N}$; or *argmax* on predictions; correlation of predictions; and more.

- Hardt, Moritz, Ben Recht, and Yoram Singer. "Train faster, generalize better: Stability of stochastic gradient descent." *International conference on machine learning*. PMLR, 2016.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. JMLR, 15(1):1929–1958, 2014.
- Summers, Cecilia, and Michael J. Dinneen. "Nondeterminism and instability in neural network optimization." *International Conference on Machine Learning*. PMLR, 2021.
- Haber, Eldad, and Lars Ruthotto. "Stable architectures for deep neural networks." *Inverse problems* 34.1 (2017): 014004.

How to improve the stability of your algorithm

Defending against overfitting also produces stability:

- "Convexify" the problem (e.g., "L1 magic" [Candes, 2005], Lasso [Tibshirani, 2011])
- Specifically introduce stability as a criterion (e.g., stability selection [Meinshausen et al, 2010])

Ensemble methods are more stable than single-run methods:

- Execute multiple runs (via bootstrap or otherwise) and construct ensemble (e.g., random forests)

Honestly communicating any residual instability requires additional experimentation:

- Execute multiple runs (via bootstrap or otherwise) and report run-to-run variability

- Candes, Emmanuel, and Justin Romberg. "l1-magic: Recovery of sparse signals via convex programming." *URL: www. acm. caltech. edu/l1magic/downloads/l1magic. pdf* 4.14 (2005): 16.
- Tibshirani, Robert. "Regression shrinkage and selection via the lasso: a retrospective." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.3 (2011): 273-282.
- Meinshausen, Nicolai, and Peter Bühlmann. "Stability selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4 (2010): 417-473.

How to improve the stability of your algorithm

Variable selection algorithms are well known for having wildly variable stability profiles:

1. Forward selection
2. Lasso
3. Stability selection

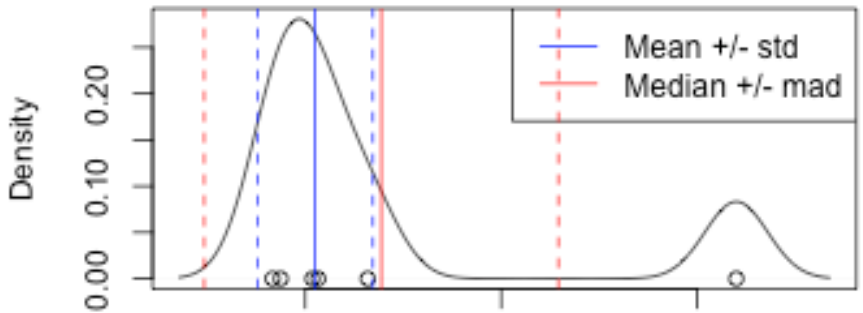
How could you measure the stability of these algorithms in an experiment?

- Candes, Emmanuel, and Justin Romberg. "l1-magic: Recovery of sparse signals via convex programming." *URL: www. acm. caltech. edu/l1magic/downloads/l1magic. pdf* 4.14 (2005): 16.
- Tibshirani, Robert. "Regression shrinkage and selection via the lasso: a retrospective." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.3 (2011): 273-282.
- Meinshausen, Nicolai, and Peter Bühlmann. "Stability selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4 (2010): 417-473.

A brief history of robust statistics

Many classical estimators make an assumption of normality, or similar (unimodal, symmetric distributions). If the distribution has fat tails and/or outliers, then these estimators suffer massively.

Robust estimators were proposed that are slightly less efficient in the “ideal” (i.e., Normal) case, but remain accurate estimators of the mean or dispersion of non-Normal distributions.



Robustness in Machine Learning

Robustness to outliers

- Outliers are usually datapoints that take extremely large (positive or negative) values. Sometimes we refer to “inliers” – i.e., any low density points, as outliers too.
- Pure tree-based methods will generally be robust to outliers in the test data, as the prediction of a tree-based method cares about which side of a split a datapoint is, not its magnitude.
- However, most methods will suffer from the presence of outliers in the training data given that parameter fitting usually involves minimizing some distance metric between predictions and data. Stability-favoring ensemble-based methods may be less sensitive to outliers.
- An alternative is to remove outliers from the test set, and flag them separately as such. There are many outlier detection methods that can be used for this purpose.

Robustness in Machine Learning

Robustness to dataset shift

- Introducing some ability to forget past data or detect concept drift is essential in modern pipelines (for example, treating data prior to or during the Covid pandemic differently).
- There are multiple types of dataset shift with covariate shift – e.g., P(X) moving – and concept drift – e.g., P(y|X) is moving – being the two most important ones for classification problems.
- Prototypical solutions include a continuous learning training regime (e.g., such as a rolling window), or a drift-detection based re-training strategy (e.g., a Champion/Challenger setup)
  - *Retrain every 3 months on the most recent 2 years' worth of data.*
  - *Deploy a model trained on 2 years' worth of historical data. Every month, compare the current model in production with a model retrained on most recent 2 years' worth of data.*
- Here too there are many methods available that offer additional robust to dataset shift.

Conclusion

- Determinism is about getting the same output when you have exactly the same input.
- Stability is about getting similar output when you have similar input, but in a DS context this should be interpreted as getting similar results (=XAI + predictions) when training on similar data.
- Robustness is about staying performant when your input or environment are being “weird”. Important examples of statistical “weirdness” are outliers, and dataset shift.

You can improve the determinism, stability and robustness of your pipelines significantly by a combination of an appropriate choice of model, data quality mitigations, and model fitting pipeline.