Scoring out-of-sample performance of IV models

Egor Kraev

June 20, 2022

1 The challenge

Let us consider the following scenario: we have a set of customers X, where each element $x \in X$ is a feature vector for a single customer. We randomly split them into the control set X_0 of customers who don't get access to a particular new feature, and the test set X_1 of customers who get access. Of the latter, the subset $X_{1,f}$ of customers choose to use the feature, and the rest don't. We further observe some outcome y, such as revenue in the 30 following days, for all customers in X.

If we want to measure the impact of the users using the feature, we can begin by assuming that the feature only impacts the outcome for those users who actually used it. We can then treat access to the feature as an instrumental variable, and usage of the feature as the treatment, and fit instrumental variable models. However, how can we compare the performance of different IV models, especially out of sample?

2 The general approach

We propose the following approach: let's call the true impact dy, which by definition will only be nonzero for the customers in $X_{1,f}$ and zero for all others. As we suppose the only impact of making the feature available is through its usage, the conditional distribution of the corrected outcome $\hat{y} \coloneqq y - dy$ given the customer features, $P(\hat{y}|x)$ will be identical between the sets X_0 and X_1.

This allows us to compare IV models out of sample, by holding out a fraction of both X_0 and X_1 as a test set (let's call them X_0^t and X_1^t), training the IV models on the remaining data, and using each of the models to calculate $\hat{y}_i$ on X_1^t (the index i goes over the different models).

We can then introduce some measure of distributional similarity d, and use it to compute the similarity between the distributions of $\hat{y}^i$ over X_0^t and X_1^t, and use that as a score for model i, with higher values denoting worse models, so

$$S_i = d(P(\hat{y}_i|x)\, over X_0^t, P(\hat{y}_i|x)\, over X_1^t)$$

3 Measuring conditional distribution distance from random samples

The challenge is that we are not given the conditional distributions themselves, but rather two sets of pairs (feature vector, corrected outcome), one for the customers who got access to the feature, and another who didn't. How can we use these to estimate a distance between the two conditional distributions?

We propose the following approach: as also in the test dataset the customers' access was assigned randomly, the distribution of customer features will be identical in X_0^t and X_1^t. Therefore, if we combine the feature vector and the corrected outcome into a single extended feature vector, a 'perfect' IV model will result in the extended feature vectors' distributions also being identical.

The problem then becomes to define a distance between the two distributions from which the respective sets of extended feature vectors have been sampled (no longer conditional distributions).

We propose two approaches to doing this. One is to use the energy distance [SR13], as implemented e.g. in the `dcor` package.

Another is to create a new target variable z that equals zero for all points in X_0 and one for all points in X_1, and combine the two holdout extended feature datasets; then train a classifier on the extended vectors, trying to predict z, that is, whether the customer in question had access to the feature or not. This classifier will only have predictive power to the extent that the sampled distributions of the extended feature vectors are different across the two datasets, thus the higher

its predictive power, the worse the score; a perfect model would result in zero predictive power of this classifier.

4 Conclusion

A way to score IV models out of sample, that makes no assumptions about the model structure, allows us to extend the power of AutoML, both for model selection and hyperparameter tuning, to the realm of IV models, in the same way that Qini scores and policy value allow us to do it for CATE models.

References

[SR13] Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.