Notes on Machine Learning Approach to Modeling Human Editing Process of IRS Tax Returns

Richard W. Evans and Daniel Silva-Inclan

September 17, 2018

1 Regression Model

When U.S. households file their taxes every year, they submit data to the IRS. I think we can divide these data fields for each household i into the subset of fields that ever get edited $\boldsymbol{x}_i$ and the rest of the fields that never get edited $\boldsymbol{z}_i$. This will reduce the dimensionality of the prediction problem. Define each element of the vectors $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ as a variable. Our goal is to accurately predict the change in $\boldsymbol{x}_i$ vector of variables as a function of $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$. Equation (1) is a representation of the true data generating process that we are trying to estimate.

$$\boldsymbol{y}_i \equiv \Delta\boldsymbol{x}_i = f\left(\boldsymbol{x}_i, \boldsymbol{z}_i\right) \tag{1}$$

We can then define the edited data as $\tilde{\boldsymbol{x}}_i$,

$$\begin{aligned} \tilde{\boldsymbol{x}}_i &\equiv \boldsymbol{x}_i + \hat{f}\left(\boldsymbol{x}_i, \boldsymbol{z}_i\right) + \boldsymbol{\varepsilon}_i \\ \tilde{\boldsymbol{x}}_i - \boldsymbol{x}_i &= \widehat{\Delta\boldsymbol{x}_i} + \boldsymbol{\varepsilon}_i \\ \Delta\boldsymbol{x}_i &= \hat{\boldsymbol{y}}_i + \boldsymbol{\varepsilon}_i \\ \tilde{\boldsymbol{x}}_i &\equiv \hat{\tilde{\boldsymbol{x}}}_i + \boldsymbol{\varepsilon}_i \end{aligned} \tag{2}$$

The vector with the hat "$\hat{}$" symbol is the estimated model of the predicted change in $\boldsymbol{x}_i$ from Equation (1), and $\hat{\tilde{\boldsymbol{x}}}_i \equiv \boldsymbol{x}_i + \widehat{\Delta\boldsymbol{x}_i}$ is the predicted edited version of the variables (the original values plus the predicted change in those values).

The model in Equation (1) is necessarily a regression model and not a classifier because the elements in $\Delta\boldsymbol{x}_i$ can each take on a continuum of values. We will estimate a machine learning model $\hat{f}\left(\boldsymbol{x}_i, \boldsymbol{z}_i\right)$ using a series of training sets that minimizes some criterion on the errors $\boldsymbol{\varepsilon}_i$ in a series of test sets.

2 Data

We have data on the original data $(\boldsymbol{x}_i, \boldsymbol{z}_i)$ and human-edited data $\tilde{\boldsymbol{x}}_i$ for some household tax returns i. We can calculate the variable of interest, the change in the editable variables $\boldsymbol{y}_i \equiv \Delta\boldsymbol{x}_i$ as the following.

$$\boldsymbol{y}_i \equiv \Delta\boldsymbol{x}_i \equiv \tilde{\boldsymbol{x}}_i - \boldsymbol{x}_i \tag{3}$$

- The data in $\boldsymbol{y}_i$ should have a lot of zeros in it.
- The variables in the vector $\boldsymbol{y}_i$ will likely have drastically different scale. Therefore, we will need to make some normalization of $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, the inputs to the model $f(\cdot, \cdot)$.

3 Model Estimation

We can test a number of tuned statistical learning models (e.g., multi-layer perceptron, SVM, random forest) to see which ones most accurately predict the vector of changes $\hat{\boldsymbol{y}}_i$ in a series of test sets. Géron (2017) is a great book for training statistical learning models using Python's `Scikit-Learn` machine learning library and training those models using the `TensorFlow` interface.

References

Géron, Aurélien, *Hands-On Machine Learning with Scikit-Learn & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, O'Reilly Media, Inc., 2017.