MONASH University

MONASH
BUSINESS
SCHOOL

notes

Shuofan Zhang

4/16/2018

Hello everyone, I am Shuofan, my research project is about a comparison between human and computer. Now let’s briefly remind ourselves of what the goal and the motivations are in this study.

Reminder of the first presentation

Firstly, we want to train the computer to read residual plots and make relevant decisions. Secondly we want to test its performance against human and conventional tests.

Residual plots are the main diagnosic tool for many issues. People may say there are always distribution tests that can be used to solve those issues. Why do we still need plots?

Why plots?

As we have talked about last time, same statistics does not necessariliy mean same relationship in the data.

All plots in this page have same mean, standard deviation and correlation up to two decimal places.

Visual inference

To make valid inference from residual plots, a protocol called lineup was developed and proved valid by Wickham, Di, Hofmann in 2013.

“The protocol consists of generating 19 null plots (could be other number), inserting the plot of the real data in a random location among the null plots and asking the human viewer to single out one of the 20 plots as most different from the others”. I believe many of you have received my survey questions, so you know how it works very well.

This is an example of lineup. The null hypothesis in this question is there is no elationship between the residuals and the fitted values. The alternative is there is some. The real plot is the third one, the null will be rejected if the third is picked. Since we have 20 plots in one lineup, if all plots are from the same population, the probability for it to be picked is 1/20 which is 0.05. This is probability of getting Type I error in this lineup question.

Different type of plots can be used for different hypothesis purpose.

Convolutional NN

CNN has advanced substantially recent years especially in computer vision problems. (slide)

Aside: Chihuahuas

In normal vision problems, computers performs not as well as human.
They cannot even tell the difference between the dogs and cuki.

Our Experiments

To test the computer’s ability of reading residual plots, we rendered two experiments in this paper.

The first one is linear vs. null and the second one is heter vs. homo.

First Experiment: Linear vs. null

Amazon turk study

Related to our first experiment here, this person conducted a large study in 2013. Three experiments were conducted to evaluate the effectiveness of the lineup protocol relative to the equivalent test statistic used in the regression setting.
Their second experiment examined $H_o : \beta_k = 0$ vs $H_a : \beta_k \neq 0$

- 70 lineups of size 20 plots were used in this experiment
- 351 evaluations were made by human subjects
-
- Trained computer model was used to classify plots from this study. Accuracy was compared with results by human subjects.

Example lineup from experiment 2

answer is 20

First experiment: linear vs. null

Human experiment procedures (diagram 1)

Here is a diagram illustrating one example of the procedures included in the human experiment. As we can see, one real plot is simulated from the linear model specification, and 19 null plots are generated under the null situation. Together they formed one lineup question. And this lineup question is then evaluated by different people. The associated p.values for the real plot in this lineup can be calculated given the total number of evaluations for this lineup, and the total number of correct answers. we will explain this a little bit more later.

We should note here the real plot can also be generated from the null, depends on the research interest.

First experiment: linear vs. null

Computer model procedures (diagram 2)

Here is the diagram of computer model procedures invovled in our first experiment.

1. A total number of 480,000 data sets have been generated in this experiment, half of them are from linear model, half of them are from the null. They were randomly selected into three sets, train, validation and test. The test set will be hidden from the model until the model is well trained. In both of our two experiments, 200,000 data sets were set apart as test set. We made the test set so large that we can make reliable comparison with them.
2. Next the scatter plots of X and Y were generated for all the simulated data, and saved as 150 $\times$ 150 pixel images.
3. And then CNN was trained on all the image data to recognize patterns from linear and the null.
4. 10 iterations over all training data were done, each iteration gives us a trained model.
5. Accuracy of the prediction on the test set was obtained for all the 10 models

First experiment: linear vs. null

Computer model procedures

repeat

First experiment: linear vs. null

Data simulation

The model is designed as this which is the same with the one used in Turk experiment 2. And all the parameters in our model were designed to cover the range used in their second experiment.

Distributions of X has impact on the shape of the scatters. For instance, if X is generated from a uniform distribution, then the plots will look like a square especially when the sample size is large; while more like a circle if X follows normal distribution. Here we used normal distribution.

- $\beta_0 = 0$
  Intercept is set to be zero, because it will not change the patterns in the data plots.
- $\beta_1 \sim U[-10, -0.1] \bigcup [0.1, 10]$
  β_1 is designed to be uniformly generated from -10 to 10 (excluding -0.1 to 0.1).
- $\varepsilon \sim N(0, \sigma^2)\ where\ \sigma \sim U[1, 12]$
  ε is designed to be uniformly generated from 1 to 12.
- $n = U[50, 500]$
  The sample sizes of each data set vary from 50 to 500.

First experiment: linear vs. null

Histogram linear

This is an overview of parameter values used in the linear class simulation, for computer model training. Histogram of each parameters are almost rectangles, the points in scatter plot of n against beta and sigma are evenly spread. According to these plots, we can say good coverage is obtained across the parameter space.

First experiment: linear vs. null

Histogram null

As for the null model, beta 1 is always 0, so we omitted two plots associated with beta1.

CNN R code

This is the R code used to build CNN. It is a stack of layer of this conv_2d and max_pooling_2d layers. We will talk a little bit about what convolution and max pooling do in a sec.

How convnets works: Model structure

The model structure associated with the code is this. After each layer of convolution and max pooling operation, the shape of outputs changes. The original 150*150 pixel image was transformed into a 1D tensor in the end. And a prediction will be made using this 1D tensor with the sigmoid function. The sigmoid function is indeed the logistic function.

How convnets works: Diagram of convolution and max pooling

Here is how convnet and max pooling works. Turns out a convolution is nothing but a element wise multiplication. i.e. dot product of the image matrix and the filter. As we can see from this picture, different filter produces different output matrix which extracts different features from the data. Max pooling is used to down-sample the input to enable the model to make assumptions about the features so as to reduce over-fitting. In this example, we use 22 *max pooling filter, so for every non-overlapping area in this matrix, we extract the max number. For example, the first* 22 elements in the first matrix is 3, 3, 2, 3, so the max is 3. which gives us the first element in the output.

First experiment: linear vs. null

training and validation metrics of convnets

The top picture is the accuracy achieved in training set and validation set by the CNN after each iteration. while the bottom is the loss. The red one is obtained in training set, while the green one is obtained in validation set. Normally the accuracy for the training set is always increasing because of overfitting. So we refer to the green line, and we can see the acc stops increasing after the fourth iteration, and pretty stable. Hence, we select the fourth, sixth, eighth and the tenth model to have them tested on the unseen test set.

First experiment: linear vs. null

convnets model selection

We separately tested the accuracy of the convnets on the linear group and the null group in test set. So we can have an idea of the estimated test power and type i error of the convnets.

The performance for models from different epoch are very close to each other. So we selected the 8th model which has the highest overall accuracy. Just by a little bit.

We should note that since the majority of the data plots in Turk's experiment have been generated with linear relationship (when the alternative hypothesis is true), it is a disadvantage for the computer. Because alpha of the model we chose is approximately 2% which is smaller than 5%.

For a fair competition, the Type I error (α) is supposed to be held the same for all test methods. However, we do not have direct control over the α of the computer model. Therefore, a 2% significant t-test and 2% significant human conclusion is also included to give a complete picture of the comparison.

First experiment: linear vs. null

Computer calculation

This slide describes how we calculate the performance of computer on Turk study data.

1, regerated 70 “real plots”

2, create a seperate test directory for 70 exclusively

3, prediction accuracy of CNN

First experiment: linear vs. null

Human calculation

The conclusion of human evaluation is obtained differently from the computer's. Because human evaluated "lineup", not only the "real plots". The performance is tested in five steps:

- Count total number of evaluations made by human for one lineup (N) and the number of correct answers for that lineup (k);
- Obtain N and k for all 70 lineup;
- Calculate p-value associated with each real plot using the formula introduced in section 2 of Majumder et al (2013);
- Draw conclusion: reject the null when the calculated p-value is smaller than α.
- The accuracy of the conclusions the 70 real plots is presenting for the human performance.

6666666666666666666666666666666666666666666 Now i am about to tell you the final results of the first comparison between human and computer. Do you want to have a guess first? Who do you think wins this competition? 666666666666666666666666666666

First experiment: linear vs. null

Comparing results

For those who think the computer wins, I have to say sorry, I really tried to train the computer well, but human wins. This is the rank of this game. Interestingly, human did really well, and its result is robust to a smaller p.value. 2% t-test and the CNN give the same accuracy on this data set.

First experiment: linear vs. null

Aside

An interesting aside discussion related to this comparison is t-test and cnn always performs similarly to each other.
I am thinking It is possible that the convnets is in fact doing the same thing as t-test in this case. Or the strategy it learned in this case turns out to be the t-test.
To check this idea, we calculated the accuracy of t-test with different alpha (from 0.005 to 0.1 with 0.005 increments) on all 200,000 test sets. The estimated power and overall accuracy were recorded. When $\alpha = 0.015$, the overall accuracy reaches its maximum. This value approximately coincide with the α chosen by the convnets.
And since the α of convnets is from 0.0142 to 0.0347 in this case, we truncated the t-test data to make this figure. From this graph, we can see the two tests indeed perform very similarly, but t-test has overall better performance. i am not saying that it is confirmed cnn is doing t-tes secretly, for now it is still my guess.

second exp

Human experiment setup

The experiment is to evaluate the human ability of reading heteroskedasticity from residual plots. It is rendered at Monash University, Melbourne Australia. The participants are all students or lecturers in this university.

Four survey are randomly sent to 84 people by email, three of the survey consist of ten lineup questions, and the fourth survey has only four lineup questions. Only one lineup question appears in the survey twice, thus, we have 33 ($10 \times 3 + 4 - 1$) distinct questions in total. A total number of 22 people have participated. Five people evaluated two surveys. One people selected four plots for each lineup, this person's response is removed from the data. In summary, we have 218 effective evaluations from 21 people.

second exp

example

This is an example of lineup question we used in the survey. which one is different? We used different difficulty level in the survey, this is an easy one, every viewer picked plot 2 for this question.

The data used for generating the lineup questions (to make real plots and null plots) were simulated using the same specification as in the next slide.

second exp

heter simlulation

In general, the choice of the parameters is an empirical work. Primarily, we want the residual plots to show more variation; on the other hand, we need to limit the range of these parameters in order to keep the key features in the data.

To better present the heteroskedasticity in the data, a uniform distribution of X is used instead of normal distribution.

Intercept is set to be zero. Because the residual plot but not data plot is used in this experiment. Therefore, the information contained in β_0 will be extracted by the linear regression we fit to the data.

β_1 has little impact in this case as well.

The variance of the error term is a quadratic function of the explanatory variable which controls the magnitude of heteroskedasticity in the model.

The parameter “a” here, following uniform distribution from -5 to 5 (excluding -0.05 to 0.05), is the correlation coefficient between X and the standard deviation. Larger a gives stronger heteroskedasticity. We belive this range is wide enough for our purpose.

This new error term “v” is added to the variance of ε so the relationship

second exp

white test

To provide a reference level of how computer and human perform, a special case of White test is used in this experiment. Every data set simulated from this section has been tested by the White test. The procedure is:

- Estimate OLS model for the data, obtain residuals ($\hat{u}$) and the fitted values ($\hat{y}$). Computer the squared OLS residuals ($\hat{u}^2$) and the squared fitted values ($\hat{y}^2$).
- Run an auxiliary regression as $\hat{u}^2 = \eta_0 + \eta_1 \hat{y} + \eta_2 \hat{y}^2 + \mathit{error}$, obtain the R-squared $R^2_{\hat{u}^2}$
- Calculate the LM statistic which follows χ^2_2 distribution
- Conclude based on p-values given certain α

second exp

traing and validation metrics of convnets

15 iterations were done for this experiment. As we can see the accuracy of the cnn in validation set is very stable. After slighly increase for the first few iterations, it stays at 0.97 something. Loss in validation indicates overfitting after 11th iterations. Same as before, we selected a few models to be tested on the test set.

second exp

CNN model selection

Again, we select the 11th model which has the highest overall accuracy. Interestingly, this time, the cnn performs better than the conventional white test. Its alpha is smaller than 5% white test, and its power is surprisingly also higher than the white test.

second exp

human performance in the exp

This graph is the proportion of people picked the real plots against the simulated “a”. Remember the “a” kinda controls magnitude of the heter. We can see that a and the proportion is strongly related. When a is large, the relationship in the data is stronger, more people are able to pick the real plot.

second exp

comparing results

The comparing results are calculated using the same method given in the first experiment.

This time, the computer finally wins. And the accuracy is far higher than the other approaches.

Discussion

The original design was to include 2 lineup questions which only has homoskedasticity plots in each survey. However instead of using a single constant as the variance of the error terms to generate the “real plot”, I incidentally used a series of random numbers. So for each observation in the sample, it has a distinct variance. Although they are not correlated with X, other undesired patterns like non-linearity and outliers were present in the data. So we had to remove those 6 lineup questions from our data set.

Second Experiment: Heter vs. Homo

Comparing results – p.values

This slide is to illustrate that, there are some similarities in the three tests. x presents for each real plot, y is the p.value associated with that plot given by three tests.

Summary

The convnet can be trained to perform similarly to human perception, when the structure in the residual plots is very specific. Performance on the linear vs no structure is comparable to the human subjects results. Performance on detecting heteroskedasticity is surprisingly good.

Performance of the convnet is a little bit below the results obtained by a *t*-test, but much higher than white-test.

Discussion

If I had more time, I would like to also include other types of structure to the model, like non-linear structure, or time series structure.

I would try other convnets structure or fine tuning the hyper-parameters in the current convnets to test its performance.

I would analyse the results theoretically and try to find the reason of why convnets did so well in heter. . . . .