STA 314 Tutorial 4

David Veitch

University of Toronto

daveveitch.github.io

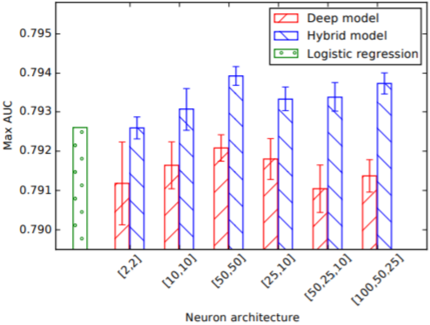
October 4, 2019

Agenda

1. Cool Paper
2. Prediction v Inference
3. Kahoot
4. ISLR 3.7 Exercise 5
5. ISLR 3.7 Exercise 7

Cool Paper

Abstract We describe the Customer LifeTime Value (CLTV) prediction system deployed at ASOS.com, a global online fashion retailer. CLTV prediction is an important problem in e-commerce where an accurate estimate of future value allows retailers to effectively allocate marketing spend, identify and nurture high value customers and mitigate exposure to losses.
https://arxiv.org/pdf/1703.02596.pdf



Performance With NN-Generated + Regular Features for Logistic Regression

Prediction v Inference

Prediction	Inference

Which is more important?

Prediction v Inference

Prediction	Inference
What is the weather tomorrow?	What is the weather in 100 years?

Which question is more important?

Kahoot

Consider the fitted values that result from performing linear regression without an intercept. In this setting the i^th fitted value takes the form

$$\hat{y}_i = x_i \hat{\beta}_i$$

where

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i'=1}^{n} x_{i'}^2}$$

Show that we can write

$$\hat{y}_i = \sum_{i'=1}^{n} a_{i'} y_{i'}$$

And show what does $a_{i'}$ equal

We are given that for a no intercept model, i^{th} fitted value

$$\hat{y}_i = x_i \hat{\beta}$$

where

$$\hat{\beta} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right)}{\left(\sum_{i'=1}^{n} x_{i'}^2\right)}$$

we can re-write this as

$$\hat{\beta} = \frac{\left(\sum_{i'=1}^{n} x_{i'} y_{i'}\right)}{\left(\sum_{j=1}^{n} x_j^2\right)}$$

Then

$$\hat{y}_i = x_i \frac{\left(\sum_{i'=1}^{n} x_{i'} y_{i'}\right)}{\left(\sum_{j=1}^{n} x_j^2\right)} = \sum_{i'=1}^{n} \left(\frac{x_i x_{i'}}{\left(\sum_{j=1}^{n} x_j^2\right)} \right) y_{i'}$$

$$\hat{y}_i = \sum_{i'=1}^{n} a_{i'} y_{i'}$$

$$a_{i'} = \frac{x_i x_{i'}}{\left(\sum_{j=1}^{n} x_j^2\right)}$$

Note: *We interpret this result by saying that the fitted values from linear regression are linear combinations of the response values.*
Question: If for a given $\hat{y}_i$, we have x_i relatively very large what happens?

ISLR 3.7 Exercise 7

It is claimed in the text that in the case of simple linear regression of Y onto X, the R^2 statistic (3.17) is equal to the square of the correlation between X and Y (3.18). Prove that this is the case. For simplicity, you may assume that $\bar{x} = \bar{y} = 0$.

ISLR 3.7 Exercise 7

First we know that since $\bar{y} = 0$ that

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i \left((y_i - \bar{y})^2\right)} = 1 - \frac{\sum_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{\sum_i y_i^2}$$

Also we know that since $\bar{x} = \bar{y} = 0$ then

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 0$$

This implies that

$$R^2 = 1 - \frac{\sum_i \left(y_i - \hat{\beta}_1 x_i\right)^2}{\sum_i y_i^2}$$

ISLR 3.7 Exercise 7

Also,

$$\left(y_i - \hat{\beta}_1 x_i\right)^2 = y_i^2 - 2x_i y_i \hat{\beta}_1 - \hat{\beta}_1^2 x_i^2$$

Taking sum over both sides gets us

$$\sum_i \left(y_i - \hat{\beta}_1 x_i\right)^2 = \sum_i y_i^2 - 2\hat{\beta}_1 \sum_i x_i y_i + \hat{\beta}_1^2 \sum_i x_i^2$$

We know since $\bar{x} = \bar{y} = 0$ that

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - x)^2} = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$

ISLR 3.7 Exercise 7

Then substituting this into the previous equation

$$\begin{aligned}\sum_i \left(y_i - \hat{\beta}_1 x_i\right)^2 &= \sum_i y_i^2 - 2\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2}\right)\sum_i x_i y_i + \left(\frac{\sum_i x_i y_i}{\sum_i x_i^2}\right)^2 \sum_i x_i^2 \\ &= \sum_i y_i^2 - \frac{\left(\sum_i x_i y_i\right)^2}{\sum_i x_i^2}\end{aligned}$$

Bringing this back to the equation for R^2 we get

$$R^2 \ = 1 - \frac{\sum_i \left(y_i - \hat{\beta}_1 x_i\right)^2}{\sum_i y_i^2} = \frac{\sum_i y_i^2 - \sum_i y_i^2 + \frac{\left(\sum_i x_i y_i\right)^2}{\sum_i x_i^2}}{\sum_i y_i^2} = \frac{\left(\sum_i x_i y_i\right)^2}{\sum_i x_i^2 \sum_i y_i^2}$$

ISLR 3.7 Exercise 7

Finally! We compare this last equation to the correlation equation.

From previous page

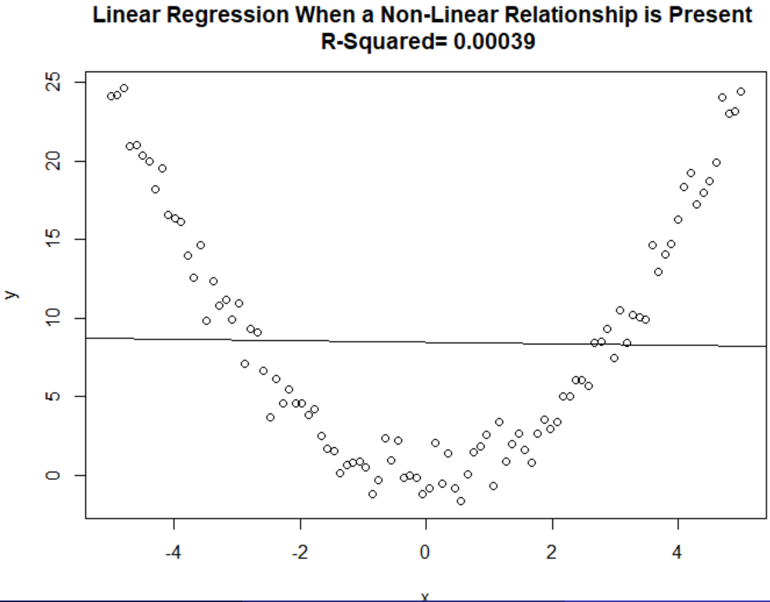
$$R^2 = \frac{\left(\sum_i x_i y_i\right)^2}{\sum_i x_i^2 \sum_i y_i^2}$$

The formula for correlation (and set $\bar{x} = \bar{y} = 0$)

$$Cor\left(X, Y\right) = \frac{\sum_i \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_i \left(x_i - x\right)^2}\sqrt{\sum_i \left(y_i - y\right)^2}} = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}}$$

$$Cor\left(X, Y\right)^2 = \left(\frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}}\right)^2 = \frac{\left(\sum_i x_i y_i\right)^2}{\sum_i x_i^2 \sum_i y_i^2} = R^2$$

Beware R^2


Linear Regression When a Non-Linear Relationship is Present
R-Squared= 0.00039
y
x
25
20
15
10
5
0
-4
-2
0
2
4