Assignment: Statistical Investigation

Sebastian Danckwardt, Nils Ekström, Reza Rezvan, Ivan Wely

October 28, 2022

1. Finding and processing the data.

   (a) Find the release date and the number of parameters of the n = 6 following models: ELMo, GPT-2, Megatron LM, Turing NLG, GPT- 3, and (very recent) Megatron-Turing NLG. Record the dates and number of parameters in a table for each model $i \in \{1, \ldots, n\}$. Discuss which sources you used and how you understood release date and number of parameters. Discuss data quality.

Name	Release Date	# of parameters
ELMo	2018	93.6 Million
GPT-2	February 14, 2019	1.5 billion
Megatron LM	September 17, 2019	8.3 billion
Turing NLG	February 13, 2020	17 Billion
GPT-3	June 11, 2020	175 billion
Megatron-Turing NLG	October 11, 2021	530 Billion

Table 1: Release Date and number of parameters

https://allenai.org/allennlp/software/elmo
https://arxiv.org/abs/1805.06556
https://www.researchgate.net/figure/Parameter-configuration-of-ELMO_tbl2_349764218

https://en.wikipedia.org/wiki/GPT-2

https://arxiv.org/abs/1909.08053

https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

https://en.wikipedia.org/wiki/GPT-3

https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/

   (b) Prepare a scatter plot with time as horizontal axis and number of parameters as vertical axis.
   Check scatterplot.png

   (c) How can you represent the release dates for each model as continuous variable x_i? Decide and record the values of x_i and the logarithm to the basis 10 of the number of parameters (let's call those values y_i) for each model $i \in \{1, \ldots, n\}$ in a second table.

   (d) Prepare a scatter plot with time x_i as horizontal axis and the logarithm to the basis 10 of the number of parameters y_i as vertical axis.
   Check scatterplot_log.png

X_i	$Y_i, Y_i = \log_{10}(X_i)$
$X_1 = 2018$	$Y_1 = -1.028\ldots$
$X_2 = 2019$	$Y_2 = 0.176\ldots$
$X_3 = 2019$	$Y_3 = 0.919\ldots$
$X_4 = 2020$	$Y_4 = 1.230\ldots$
$X_5 = 2020$	$Y_5 = 2.243\ldots$
$X_6 = 2021$	$Y_6 = 2.72\ldots$

Table 2: Release Date and log of number of parameters

(e) Are there any visual outliers?

Yes, GPT-3 is quite a outlier, which isn't weird, it was a huge deal in the world of AI models.

2. Fitting a linear model..

(a) A simple regression model postulates relation between explained variable (logarithmic model size yi) and explanatory variable (time xi)

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{1}$$

Explain why the exponential growth hypothesis suggest such a model.

If the relationship is presumed exponential and you take the log of the explanatory variable , then the curve goes from exponential to straight, because log(exp(x)) == x. That way you can fit a linear model instead of an exponential.

(b) Perform a regression to find estimates for α and β

Performing linear regression yields:
$\hat{\alpha} = -2110.42$ and $\hat{\beta} = 1.04535$

(c) Add the regression line to the scatter plot of 1d.

Check scatter_regression.png

(d) Prepare a scatter plot of the residuals. Does the model give a good fit? Discuss model validation.

Check scatter_residuals for a picture, as we can see there's only one outlier which makes the model less believeable. But under these circumstances it's still an okay model. The expected value of the resiudals is ≈ 0.2 which is quite close to 0. So we can consider this regression model to be okay.

(e) Add the curve corresponding to the regression line in 1d to the scatter plot of 1b.

Check scatterplot_test.png

(f) Give an interpretation of the slope: how many years does it take for parameter number to increase by a factor of 10 in terms of β?
We have $y_i = -2110.42 + 1.04535x_i$ so to increase by a factor of 10, meaning $y_i \longrightarrow 10y_i$ takes ≈ 9 years

3. Making predictions.

(a) Is there any significant linear relation between the logarithmic parameter number and time? Formulate a null hypothesis and test at significance level 0.05. Assume normal errors with unknown variance σ^2.

$H_0 : \beta = 0$
$H_1 : \beta \neq 0$

$CI_\beta = \hat{\beta} \pm t_{\alpha/2}(n-2)\frac{s}{\sqrt{S_{xx}}} = \cdots = [0.70912796, 1.38157204]$

Since $H_0 \notin Ci_\beta$ we can reject the null-hypothesis and say that H_1 is true. Therefore our conclusion is that, yes, there's a significant linear relation between the (logarithmic) parameter number and time (A positive one at that as well).

(b) Find a 95% confidence interval for the slope. Use it to determine a 95% confidence interval for the length of time in which the model parameter number tends to increase by a factor of 10.

$CI_\beta = \hat{\beta} \pm t_{\alpha/2}(n-2)\frac{s}{\sqrt{S_{xx}}} = \cdots = [0.70912796, 1.38157204]$

$CI_{x10} = [7.23813142599, 14.101827264]$

(c) What model size can we expect for models in next spring (April 2023)? Give a 95% confidence interval. Let's see if that comes true...

(d) What do think: Will the model make a reasonable prediction of model size in 2030?