2018-01-19

Tom Rochette <tom.rochette@coreteks.org>

July 24, 2025 — daae079c

0.1 Context

0.2 Learned in this study

0.3 Things to explore

1 Problems faced

2 Overview

- How do simple neural networks work?
  - The square of the difference between the output and the target is computed
  - Then this value is reduced to a single value we call the loss
  - Our objective is to minimize this loss as much as possible with the use of an optimizer
- Dense layers
  - Dense layers are tensors of weights $\mathbf{W}$ which are multiplied against the input $\mathbf{x}$ of the layer, to which a bias $\mathbf{B}$ is added

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{B}$$

  - During training, $\mathbf{y}$ is the target we want to predict and $\mathbf{x}$ is the associated observation. Thus, the only parameters (the values we can modify) we have are $\mathbf{W}$ and $\mathbf{B}$
  - In it's simplest form, we have

$$y = Wx + B$$

  - This is a linear equation, where all the values are scalars. If I have an example as part of my training set where $\hat{y} = 3$ ($\hat{y}$ is the expected or target value) and $x = 6$, I can either set $W = 0$ and $B = 3$ or $W = 1/2$ and $B = 0$ or any other valid combination.
  - What happens when we're training a dense layer is that we want as the examples we've extracted from our training set to produce as little error as possible. Thus, if we have a second example where $\hat{y} = 3$ and $x = 7$, $W = 0$ and $B = 3$ produces $y = 3$ while $W = 1/2$ and $B = 0$ produces $y = 3.5$. Thus the first equation produces a better approximation for these two examples compared to the second one.
  - Now, how does one decide which values of W and B are the best for a given training set?
    - Let's initialize $W = 0$ and $B = 0$
    - If for each example in our training set, we compute the resulting y, we can tell how our current values of W and B together are close or far from the value that can best approximate the function represented by the training examples.

$$3 = W \times 6 + B$$
$$0 = W \times 7 + B$$

$$3 =^{?} 0 \times 6 + 0 = 0$$
$$0 =^{?} 0 \times 7 + 0 = 0$$

* From this, we can see that there is an error of 3 in the first training set example, but no error in the second. Given a linear function and two training examples, the best function is one that goes through the two examples, namely $(6, 3)$ and $(7, 3)$. In such a simple case, the equation is

$$W = \frac{y_2 - y_1}{x_2 - x_1}$$
$$B = y_1 - W x_1$$

$$W = \frac{3 - 3}{7 - 6} = \frac{0}{1} = 0$$
$$B = 3 - 0 \times 6 = 3$$

* As such, the optimal values for these two samples are $W = 0$ and $B = 3$.

$$y = 0x + 3$$

* Now we introduce a third example, $(0, 0)$. Given our current values of W and B, we get

$$y = 0 \times 0 + 3 = 3$$

* However, that is not quite right, as we have an error of 3. Since we previously stated that the best linear function that approximates two points is the one that goes through both of them and we have an error with our new third point, it means we'll have to compromise between the three points if we want to reduce our overall error (the error on our 3 training examples).
* What we could do here is to create the 3 linear functions that can be generated by combining 2 points, such that we have the couples $(0, 0)$ and $(6, 3)$, $(0, 0)$ and $(7, 3)$, $(6, 3)$ and $(7, 3)$
  - $y = \frac{3}{6}x$
  - $y = \frac{3}{7}x$
  - $y = 3$
* Given the excluded point as a source of error, we have the follow errors
  - $\frac{3}{6} \times 7 = \frac{21}{6} = 3.5$
  - $\frac{3}{7} \times 6 = \frac{18}{7} \approx 2.57$
  - $3 = 3$
* In every cases there is a certain amount of error, however we can see that $y = \frac{3}{7}x$ has the lowest of the three. Now, we might wonder whether this function is the best, or some other function could be devised... If so, how?
* In the linear function approximation case, we can think of the line we're looking for to be the one that has the smallest distance with all points. We want to optimize $\min \sum_i (\hat{y_i} - y_i)^2$.
* In optimization, what we generally want to do is to find either a minimum or a maximum. In any cases, we can now think that our graph has W as its x coordinate and B as its y coordinate.

3 See also

4 References

- https://en.wikipedia.org/wiki/Linear_regression