Week 2: Gradients

http://mlvu.github.io

February 19, 2020

In this session we will practice searching for a good or optimal model using the *gradient*. We will define the *loss function* of a model, with respect to some target data, and we will will minimize the loss function with respect to the parameters.

In the following explanation, some parts are replaced by green dots. Fill these in.

1 Partial derivatives

We'll begin by reviewing the idea of *partial derivatives*: derivatives of functions with multiple inputs. If you haven't worked with derivatives for a while (or ever), start by having a look at the following resources:

- https://www.youtube.com/watch?v=ANyVpMS3HL4
- https://www.youtube.com/watch?v=hCLfogkqzEk
- http://betterexplained.com/articles/derivatives-product-power-chain/

A partial derivative is the derivative with respect to one of the parameters, treating all the others as constants. For a simple example, consider the function

$$f(a, b) = 3a^2 + b^2 - ab + a.$$

We first take the derivative with respect to a:

$$\frac{\partial f(a, b)}{\partial a} = \frac{\partial (3a^2 + b^2 - ab + a)}{\partial a} = 6a - b + 1\ .$$

Note that the second term of the function (b^2) does not contain a, i.e. it is constant with respect to a, so it disappears from the derivative.

Then, we take the derivative wrt. to b:

$$\frac{\partial f(a,b)}{\partial b} = \frac{\partial(3a^2 + b^2 - ab + a)}{\partial b} = 2b - a \ .$$

The *gradient* $\nabla f(a,b)$ is simply all the partial derivatives of a function arranged in a (row) vector:

$$\nabla f(a,b) = \left(\frac{\partial f(a,b)}{\partial a}, \quad \frac{\partial f(a,b)}{\partial b}\right) = (6a - b + 1, \quad 2b - a) \ .$$

If we imagine the function f as a "landscape" over the plane defined by the a and b-axes, the gradient at each point in that plane, is an arrow pointing in the direction in which the ascent is the steepest.[1] A marble placed at some point, and released, would roll in the opposite direction to the gradient.

To find the point at which the function has reached its minimum, we must find out where all partial derivatives are equal to zero.[2]

question 1: To find the a and b for which f(a, b) is minimal, we set the partial derivatives to zero. Fill in the blanks (indicated by "...") in this derivation:

$$\begin{aligned} 6a - b + 1 &= 0 \\ a &= (b-1)/6 \end{aligned} \qquad \begin{aligned} 2b - a &= 0 \\ b &= a/2 \\ b &= \frac{b-1}{6}\frac{1}{2} \\ b - \frac{b}{12} &= -\frac{1}{12} \\ b &= -\frac{1}{11} \end{aligned}$$

$$a = \frac{-\frac{1}{11} - 1}{6} = -\frac{2}{11}$$

A quick check on Wolfram Alpha shows that this solution is correct.

[1] We can interpret a vector $\mathbf{x} \in \mathbb{R}^n$ as a point in $\mathbb{R}^n$, but also as a *direction*. The direction is that of the arrow between the origin and the point $\mathbf{x}$. The gradient only makes sense as a direction.

[2] To be precise, if all partial derivatives are zero, the function may be at a minimum, a maximum, or a plateau. It may even be a saddle-point, a place where the function is a minimum in one dimension and a maximum in another. For the purposes of this exercise, you may assume that if the gradient is zero, you have found a minimum.

1.1 Rules

While it's important to understand intuitively what differentation means (see the *better explained* link above), actually finding a specific derivative is usually a very mechanical process of matching existing rules and rewriting a function to a useful form. The following rules are usually sufficient: Let c be a constant, independent of x, and let $f(x)$ and $g(x)$ be arbitrary functions of x. Then:

$$\frac{\partial c}{\partial x} = 0 \qquad \text{the } \textit{constant} \text{ rule}$$

$$\frac{\partial x^n}{\partial x} = n x^{n-1} \qquad \text{the } \textit{exponent} \text{ rule}$$

$$\frac{\partial x}{\partial x} = 1 \qquad \text{(follows from the exponent rule)}$$

$$\frac{\partial c f(x)}{\partial x} = c \frac{\partial f(x)}{\partial x} \qquad \text{the } \textit{constant factor} \text{ rule}$$

$$\frac{\partial (f(x) + g(x))}{\partial x} = \frac{\partial f(x)}{\partial x} + \frac{\partial g(x)}{\partial x} \qquad \text{the } \textit{sum} \text{ rule}$$

$$\frac{\partial f(g(x))}{\partial x} = \frac{\partial f(g(x))}{\partial g(x)} \frac{\partial g(x)}{\partial x} \qquad \text{the } \textit{chain} \text{ rule}$$

To find the derivative of $f(x, y, z)$ with respect to y, you write down

$$\frac{\partial f(x, y, z)}{\partial y},$$

fill in the definition of f, and match the result (whole, or part of it) with the left-hand side of one of these rules. You replace it by the right-hand side and keep going until all the ∂'s are gone.

question 2: To practice, let's take the derivative of $(x + y + z)^2$ with respect to y. Fill in the blanks.

$$
\begin{aligned}
\frac{\partial(x+y+z)^2}{\partial y} &= \frac{\partial(x+y+z)^2}{\partial(x+y+z)}\frac{\partial(x+y+z)}{\partial y} && \text{using the chain rule} \\
&= 2(x+y+z)\frac{\partial(x+y+z)}{\partial y} && \text{using the exponent rule} \\
&= 2(x+y+z)\left(\frac{\partial x}{\partial y}+\frac{\partial y}{\partial y}+\frac{\partial z}{\partial y}\right) && \text{using the sum rule} \\
&= 2(x+y+z)(0+1+0) && \text{using the constant and exponent rules} \\
&= 2(x+y+z)
\end{aligned}
$$

2 Linear regression

As discussed in the lecture, the optimal model for standard linear regression can be computed directly. In this assignment we will derive this function. Lets say we are trying to predict the size to which a particular newborn baby will grow in the coming months. We have measured the child in its first few months and found the following data:

age(a, months)	height (h, cm)
$a_1 = 0$	$h_1 = 30$
$a_2 = 2$	$h_2 = 40$
$a_3 = 4$	$h_3 = 50$

Let n be the number of examples we have (3 in this case, but we want to derive a solution for arbitrary n). Our model, f, is a linear function, described by two parameters: the slope s and the intercept b:

$$f_{s,b}(a) = sa + b$$

Where $f(a_i)$ should be as close to h_i as possible. We will use the sum of squared errors as our loss function. That is, we will compute the residual $f(a_i) - h_i$ for each data point, square them, and sum these squared values. In other words, our loss function is

$$\text{loss}(s,b) = \frac{1}{2}\sum_i \left(f_{s,b}(a_i) - h_i\right)^2 = \frac{1}{2}\sum_i \left(sa_i + b - h_i\right)^2$$

The $\frac{1}{2}$ multiplier is there to simplify the derivatives later. It doesn't affect the minimum of the loss function, which is what we're after.[3]

question 3: For $s = 1$ and $b = 0$, and the data given above, what is the loss?

$$\begin{aligned}\text{loss}(1,0) &= \frac{1}{2}\sum_i (1 \times a_i + 0 - h_i)^2 \\ &= \frac{1}{2}\left((0-30)^2 + (2-40)^2 + (4-50)^2\right) \\ &= \frac{1}{2}(900 + 1444 + 2116) = 2230\end{aligned}$$

question 4: The term $f(a_i) - h_i$ represents the difference between our model's prediction and the observed value (the *residual*). Per example, this is a good measure of the error. Why do we not just sum these, and check how far it is from 0? Why square it first, and *then* sum?

If we summed them, one big positive error could cancel out against one big negative error, giving the false impression that the model is highly accurate. We need to square the errors before summing them. The choice for square instead of another approach (like taking the absolute value, or raising to some other power) is more subtle. Squaring emphasizes the loss of large errors compared to the absolute value.

It turns out that if we assume that the data is linear, but with added Gaussian noise, optimizing for the sum-of-squared errors is equivalent to optimizing likelihood (we will discuss this later in the probability lectures.)

question 5: Let's find the derivative of the loss function. One with respect to s and one with respect to b. Fill in the blanks.

[3] In the slides, we used the *mean* sum of squared errors (i.e. we multiplied by $\frac{1}{n}$ as well). Since n is a constant, the two functions have the same minimum and we can use either one.

$$
\begin{aligned}
\frac{\partial \text{loss}(s,b)}{\partial s} &= \frac{\partial \frac{1}{2}\sum_i (sa_i + b - h_i)^2}{\partial s} \\
&= \frac{1}{2}\sum_i \frac{\partial (sa_i + b - h)^2}{\partial (sa_i + b - h_i)} \frac{\partial (sa_i + b - h_i)}{\partial s} \\
&= \frac{1}{2}\sum_i 2(a_i s + b - h_i)a_i \\
&= \sum_i (a_i s + b - h_i)a_i \\
&= \sum_i ({a_i}^2 s + a_i b - a_i h_i) \\
&= s\sum_i {a_i}^2 + b\sum_i a_i - \sum_i a_i h_i
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \text{loss}(s,b)}{\partial b} &= \frac{\partial \frac{1}{2}\sum_i (sa_i + b - h_i)^2}{\partial b} \\
&= \frac{1}{2}\sum_i \frac{\partial (sa_i + b - h_i)^2}{\partial (sa_i + b - h_i)} \frac{\partial (sa_i + b - h_i)}{\partial b} \\
&= \frac{1}{2}\sum_i 2(a_i s + b - h_i) \\
&= \sum_i (a_i s + b - h_i) \\
&= s\sum_i a_i + bn - \sum_i h_i
\end{aligned}
$$

question 6: For many loss functions, setting the gradients equal to zero results in a system of equations that cannot be solved analytically. In that case we will have to *search*. We can still use the gradient though, in an algorithm called *gradient descent*. Starting with the parameter values $s = 1$ and $b = 0$, describe one step of the gradient descent algorithm (with

learning rate 0.01). For these parameters, the gradient is

$$\begin{aligned}
&\left(s\sum_i a_i{}^2 + b\sum_i a_i - \sum_i a_i h_i,\ \ s\sum_i a_i + bn - \sum_i h_i \right) \\
&= \left(\sum_i a_i{}^2 - \sum_i a_i h_i,\ \ \sum_i a_i - \sum_i h_i \right) \\
&= (4 + 16 - (2 \cdot 40 + 4 \cdot 50),\ \ 6 - 120) \\
&= (-260,\ \ -114)
\end{aligned}$$

For gradient descent we pick the opposite direction (since the gradient points up), multiply by the learning rate, and add the result to the current parameters. Thus, the new parameters are:

$$\begin{aligned}
\begin{bmatrix} s^{\text{new}} \\ b^{\text{new}} \end{bmatrix} &= \begin{bmatrix} s \\ b \end{bmatrix} - \eta \nabla \text{loss}(s, b) \\
&= \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 0.01 \begin{bmatrix} -260 \\ -114 \end{bmatrix} = \begin{bmatrix} 3.6 \\ 1.14 \end{bmatrix} .
\end{aligned}$$

question 7: Set the two derivatives found above equal to zero, and solve to obtain expressions for the optimal model. Make sure that your expression for s does not depend on b, so that the solution can actually be computed. To simplify notation, it can be helpful to use the following conventions for the data means and other statistics:

$$\begin{aligned}
\overline{a} &= \frac{1}{n}\sum_i a_i \\
\overline{h} &= \frac{1}{n}\sum_i h_i \\
\overline{a^2} &= \frac{1}{n}\sum_i a_i{}^2 \\
\overline{h^2} &= \frac{1}{n}\sum_i h_i{}^2 \\
\overline{ah} &= \frac{1}{n}\sum_i a_i h_i
\end{aligned}$$

Fill in the blanks:

$$
\begin{aligned}
s\sum_i {a_i}^2 + b\sum_i a_i - \sum_i a_i h_i &= 0 \\
s\sum_i {a_i}^2 &= -b\sum_i a_i + \sum_i a_i h_i \\
s &= -b\frac{\sum_i a_i}{\sum_i {a_i}^2} + \frac{\sum_i a_i h_i}{\sum_i {a_i}^2} \\
s &= -b\frac{\frac{1}{n}\sum_i a_i}{\frac{1}{n}\sum_i {a_i}^2} + \frac{\frac{1}{n}\sum_i a_i h_i}{\frac{1}{n}\sum_i {a_i}^2} \\
s &= -b\frac{\overline{a}}{\overline{a^2}} + \frac{\overline{ah}}{\overline{a^2}} \qquad (1)
\end{aligned}
$$

$$
\begin{aligned}
s\sum_i a_i + bn - \sum_i h_i &= 0 \\
bn &= \sum_i h_i - s\sum_i a_i \\
b &= \frac{1}{n}\sum_i h_i - s\frac{1}{n}\sum_i a_i \\
b &= \overline{h} - s\overline{a} \qquad (2)
\end{aligned}
$$

Fill equation (2) into equation (1):

$$
\begin{aligned}
s &= -\left(\overline{h} - s\overline{a}\right)\frac{\overline{a}}{\overline{a^2}} + \frac{\overline{ah}}{\overline{a^2}} \\
s &= -\frac{\overline{h}\,\overline{a}}{\overline{a^2}} + s\frac{\overline{a}^2}{\overline{a^2}} + \frac{\overline{ah}}{\overline{a^2}} \\
s\left(1 - \frac{\overline{a}^2}{\overline{a^2}}\right) &= -\frac{\overline{h}\,\overline{a}}{\overline{a^2}} + \frac{\overline{ah}}{\overline{a^2}} \\
s &= \frac{\overline{ah} - \overline{a}\,\overline{h}}{\overline{a^2} - \overline{a}^2}
\end{aligned}
$$

Note that we have now expressed s and b purely in terms of statistics that are easily computed from our data, like the mean height $\overline{h}$ and the mean age $\overline{a}$.

question 8: Fill in the values from the data set, and compute the optimal parameters for this data. Can you explain what the parameters s and b mean? That is, what can they tell us about the baby we've measured? Filling in the data gives us $\overline{ah} = \frac{280}{3}$, $\overline{a} = 2$, $\overline{h} = 40$, $\overline{a^2} = \frac{20}{3}$. This gives us

$$s = \frac{\frac{280}{3} - 2 \cdot 40}{\frac{20}{3} - 4} = \frac{\frac{40}{3}}{\frac{8}{3}} = 5 \tag{3}$$

$$b = 40 - 5 \cdot 2 = 30 \tag{4}$$

This tells us that this baby grows about 5 cm per month, and its size at birth was 30 cm. (Note these values were made up, and are actually a little small for a newborn baby.) See if Wolfram Alpha agrees with your answer.