Mathematics for Machine Learning (70015)

Lecture 1.1 - Linear Regression Intro

Linear regression aims to provide a solution to the supervised learning problem; we are given a dataset of N examples of inputs and expected outputs, with a goal of predicting the correct output for a new input. Examples of this include image classification (such as digit classification) and translation. A curve fitting problem in 1-dimension has an input space $\in \mathbb{R}$, and an output space $\in \mathbb{R}$.

To tackle this problem mathematically, we need to first describe the curve fitting problem mathematically. As each input is associated with a single output, this is equivalent to a function in mathematics. We are given a dataset of N pairs, of inputs and outputs, where $\boldsymbol{x_n} \in \mathcal{X}$, which is usually $\mathbb{R}^D$ and $y_n \in \mathcal{Y}$ (usually $\mathbb{R}$ in this case); $\{(\boldsymbol{x_n}, y_n)\}_{n=1}^N$. The goal is to find a function that maps from the input space to the output space **well**; $f : \mathcal{X} \to \mathcal{Y}$.

We need to first find candidates for functions that can perform the predictions. Functions need to be parameterised, such that some numbers $\boldsymbol{\theta}$ map to a function. From here, we need to pick the 'best' function, thus requiring us to define what good and bad functions are. A good function has the property $f(\boldsymbol{x_i}, \boldsymbol{\theta^*}) \approx y_i$; the output of the function closely matches the outputs of the training points. This can be defined with a **loss function**, for example;

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{N} (y_i - f(\boldsymbol{x_i}, \boldsymbol{\theta}))^2$$

Therefore, a good function is chosen by minimising the loss; $\boldsymbol{\theta^*} = \operatorname{argmin}_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$.

Lecture 1.2 - Scalar Differentiation

We can plot the loss against the parameters for a function. This raises two questions; how to change the parameter to make the loss smaller and how we know if we can't get a better loss. The derivative is defined as the limit of the difference quotient (as usual);

$$f'(x) = \frac{\mathrm{d}f}{\mathrm{d}x} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Several examples of this are as follows;

$f(x)$	$f'(x)$
x^n	nx^{n-1}
$\sin(x)$	$\cos(x)$
$\tanh(x)$	$1 - \tanh^2(x)$
$e^x = \exp(x)$	e^x
$\log(x)$	$\frac{1}{x}$

There are also the following rules which combine the basic functions;

- sum rule — describes the derivative of sum of two functions

$$(f(x) + g(x))' = f'(x) + g'(x)$$

- product rule — similarly, for multiplication

$$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$$

- chain rule describes how to differentiate functions that are composed

$$(g \circ f)'(x) = (g(f(x)))' = g'(f(x))f'(x)$$

- quotient rule describes division, special case of the product rule

$$\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$$

This tells us how to change the input in the function; the gradient tells us how much the output changes based on an increase in the input. We first compute the derivative function at a point to find the point's gradient. If the gradient is negative, this tells us the function will decrease if we increase the input (hence increase to minimise), and vice versa; decrease for positive gradients - this is the idea behind gradient descent.

Similarly, we know we are done (at a minimum for the loss function) when there is nothing that can be done to lower the output, hence the gradient must be 0. However, this isn't sufficient to tell us that we have reached a minimum, as a maximum also has a gradient of zero. A minimum has a decreasing function followed by an increasing function; hence the gradient of the gradient (second derivative) is positive.

However, this only gives us a local minima. We should be concerned with getting stuck in a local minima (rather than a global minima) when dealing with non-convex functions. Working through a simple example, with a linear regression problem (aiming to find an optimal a) - the final step takes the second derivative to verify we have a minimum;

$$\begin{aligned}
f(x) &= a \cdot x \\
L(a) &= \sum_{n=1}^{N} (f(x_n) - y_n)^2 \\
\frac{\mathrm{d}L}{\mathrm{d}a} &= \sum_{n=1}^{N} 2(ax_n - y_n)x_n \\
&= \sum_{n=1}^{N} 2ax_n^2 - 2x_ny_n \\
&= 0 \\
2a\sum_n x_n^2 &= \sum_n 2x_ny_n \\
a &= \frac{\sum_n 2x_ny_n}{\sum_n x_n^2} \\
\frac{\mathrm{d}^2L}{\mathrm{d}a^2} &= \sum_{n=1}^{N} 2x_n^2 \\
&\geq 0
\end{aligned}$$