Machine Learning Course - CS-433

Linear Regression

Sep 22, 2016

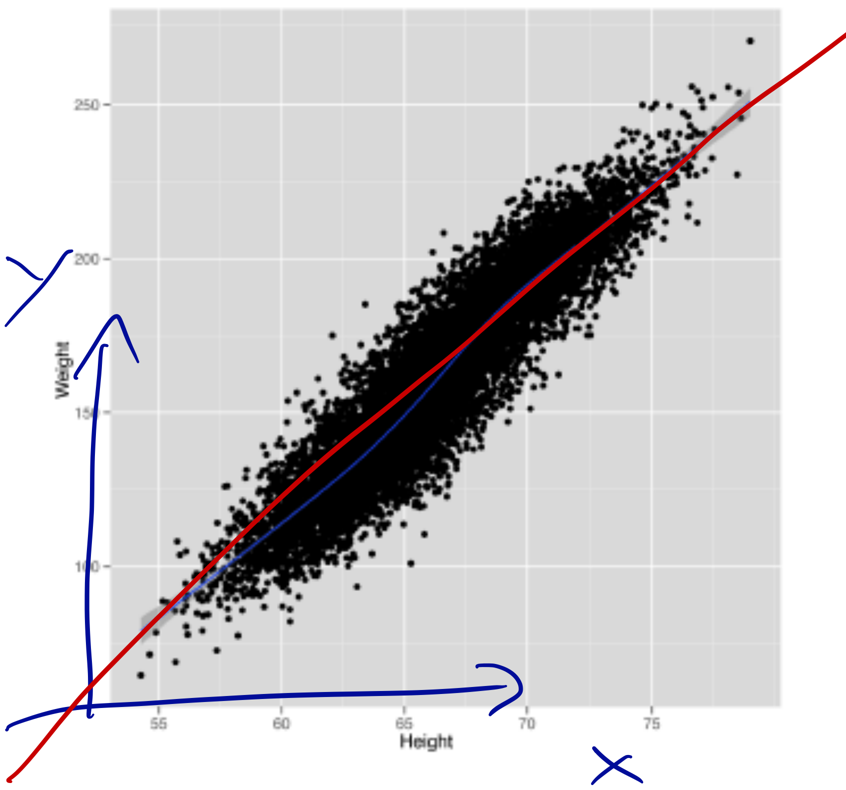
minor changes by Martin Jaggi 2016

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

1 Model: Linear Regression

What is it?

Linear regression is a model that assumes a linear relationship between inputs and the output.



$y_n \approx f(x_n)$

$\|$

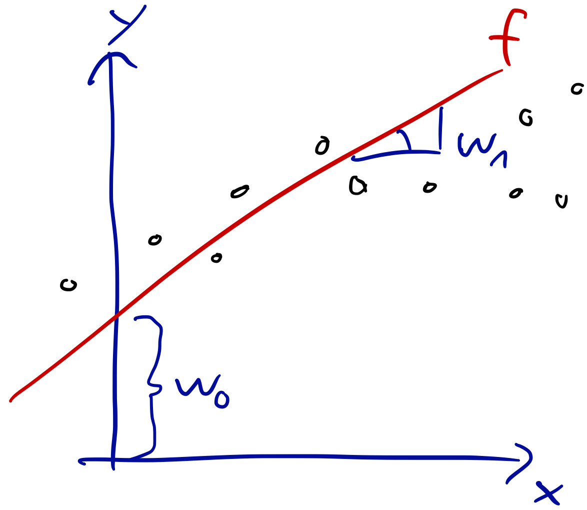
$x_n^T w$

Why learn about *linear* regression?

Plenty of reasons: simple, easy to understand, most widely used, easily generalized to non-linear models. Most importantly, you can learn almost all fundamental concepts of ML with regression alone.

scalable

$x \in \mathbb{R}$



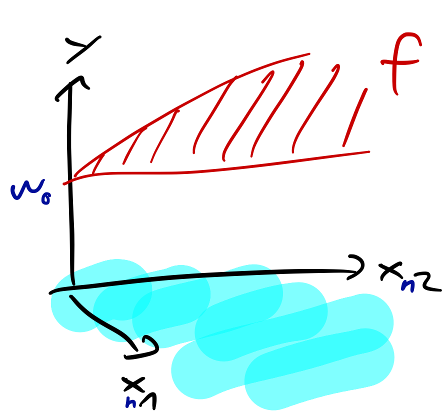
Simple linear regression

With only one input dimension, we get simple linear regression.

$$y_n \approx f(\mathbf{x}_n) := \underline{w_0 + w_1 x_{n1}}$$

Here, $\mathbf{w} = (w_0, w_1)$ are the two parameters of the model.

$x \in \mathbb{R}^D$



Multi(ple) linear regression

If our data has multiple input dimensions, we obtain multivariate linear regression.

$$
\begin{aligned}
y_n &\approx f(\mathbf{x}_n) \\
&:= w_0 + w_1 \overline{x_{n1}} + \ldots + w_D \overline{x_{nD}} \\
&= \mathbf{x}_n^T \mathbf{w} + w_0 \\
&= \tilde{x}_n^T \tilde{w}
\end{aligned}
$$

$$\tilde{x}_n = \begin{pmatrix} 1 \\ x_n \end{pmatrix} = \begin{pmatrix} 1 \\ x_{n1} \\ \vdots \\ x_{nD} \end{pmatrix}$$

Learning / Estimation / Fitting

Given data, we would like to find $\mathbf{w} = [w_0, w_1, \ldots, w_D]$. This is called learning or estimating the parameters or fitting the model.

To do so, we need an optimization algorithm, which we will discuss in the chapter after the next.

find f
(good w)

Additional Notes

Alternative when not using an 'offset' term

Above we have used $D + 1$ model parameters, to fit data of dimension D. An alternative also often used in practice, in particular for high-dimensional data, is to ignore the offset term w_0.

$$
\begin{aligned}
y_n \approx f(\mathbf{x}_n) &:= w_1 x_{n1} + \ldots + w_D x_{nD} \\
&= \underline{\mathbf{x}_n^T \mathbf{w}}
\end{aligned}
$$

in this case, we have just D parameters to learn, instead of $D + 1$.

As a warning, you should be aware that the number of weight parameters of a machine learning model in general can be very different from D (being the dimension of the input data). Think for example of a neural network (more details later).

Matrix multiplication

To go any further, one must revise matrix multiplication. Remember that multiplication of $M \times N$ matrix with a $N \times D$ matrix results in a $M \times D$ matrix. Also, two matrices of size $M \times N_1$ and $N_2 \times M$ can only be multiplied when $N_1 = N_2$.

The $D > N$ Problem

Consider the following simple situation: You have $N = 1$ and you want to fit $y_1 \approx w_0 + w_1 x_{11}$, i.e. you want to find $\mathbf{w} = (w_0, w_1)$ given one pair (y_1, x_{11}). Is it possible to find such a line?

This problem is related to something called $D > N$ problem (in statistics typically named $p > n$). It means that the number of parameters exceeds number of data examples. In other words, we have more variables than we have data information. For many models, such as linear regression, this makes the task *under-determined.*

These problems are all solved by using regularization, which we will learn later.