Indtroduction to Machine Learning

# Linear models of regression

Nadia Vohwinkel

supervised by
PhD. Akash Ashirbad Panda

November 23, 2020

# Contents

# 1 Linear models of regression

A linear model of regression is a supervised learning method. Given a set of data, the goal is to find a regression function, that best fits the data points and enables us to make predictions for possible new points. This regression function will meet a certain amount of criteria, that will make it a linear function. Considering linear functions is convenient because of their strightforward (mathematical) interpetation and resulting ease to estimate new values.

## 1.1 What is regression analysis?

In general, regression analysis is a mathematical method used in order to predict the value of one or more continuous target variables $t$, given a $N$-dimensional vector $x$ of input variables. Also it helps us to estimate the causal relationship between theese variables, meaning what impact a variable has on an other. This becomes especially important when looking at multivariable linear regression methods. In comparison to some other supervised learning methods, regression has a numerical output, rather than allocating input values to a specific category.

## 1.2 Linear regression

Throughout this paper, we will consider a training data set comprising $N$ observations $\{x_n\}$ and corresponding target values $\{t_n\}$, $n \in \{1, ..., N\}$. We will denote $\mathbf{x} = (x_1, ..., x_N)^T$ and $\mathbf{t} = (t_1, ..., t_N)^T$.
When talking about linear regression, the majority of people imagine a straight line embedded in a plane. This presumption highly restricts the possibilties for linear regression one could take into account. Rather than looking at a function

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \cdots + w_N x_N \tag{1}$$

we wan't to study functions satisfying

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}). \tag{2}$$

Please note, that this second function (2) has $M$ summands, whereas (1) has $N$.
When reagrding both of theese functions, we can see that indeed (1) is a linear function in $x$, whereas (2) possibly is not. But we can also agree on the fact, that both of theese functions are linear in $w$. We call the $\{w_j\}, j \in \{0, ..., M-1\}$ weights and we denote $\mathbf{w} = (w_0, ..., w_{M-1})^T$.
The $\{\phi_i\}$ are called basisfunctions ($i = 0, ..., M-1$).

### 1.2.1 Basisfunctions

Basisfunctions allow a much more broad approach to linear regression. They permit the regression function $y(\mathbf{x}, \mathbf{w})$ not the be linear function in $x$ and thereby extend the range of functions one can use to fit the given data. The choice of basisfunction has a large impact on the complexity and flexibility of the final regression function. There are hardly any restrictions for the selection of a basisfunction. Common examples are polynomial functions, the 'Gaussian basisfunction' or so called wavelets. The regression function will then be a linear combination of the chosen basisfunctions. Later, when duscussing the bias-variance trade-off we will get a closer look und how to choose the right basisfunction.
To facilitate things, we want to adjust the formula from (2) a little bit by adding a "dummy function" $\phi_0(\mathbf{x}) = 1$ in order to obtain

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}). \tag{3}$$

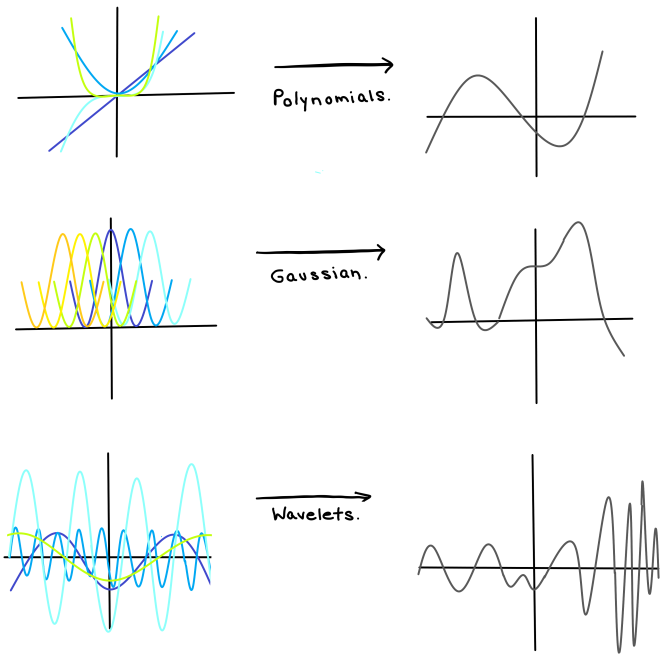
Let's briefly get back to the number of summands in (2). Assuming that $\{x_n\} \in \mathbb{R}$ for $n = 1, ..., N$, we can determine:

$$\phi : \mathbb{R}^N \to \mathbb{R}^M, \ x \mapsto \phi(x)$$

So $\phi$ takes in the entire vector $\mathbf{x}$ and maps it into a $M$-dimensional space. By identifying $\mathbf{z} := \phi(\mathbf{x})$ we receive the function

$$y(\mathbf{z}, \mathbf{w}) = \mathbf{w}^T \mathbf{z},$$

which is a linear function in $z$ as well as in $w$. Keep this in mind for later on. We will use this notation to keep things a little less cluttered.




### 1.2.2 Noise

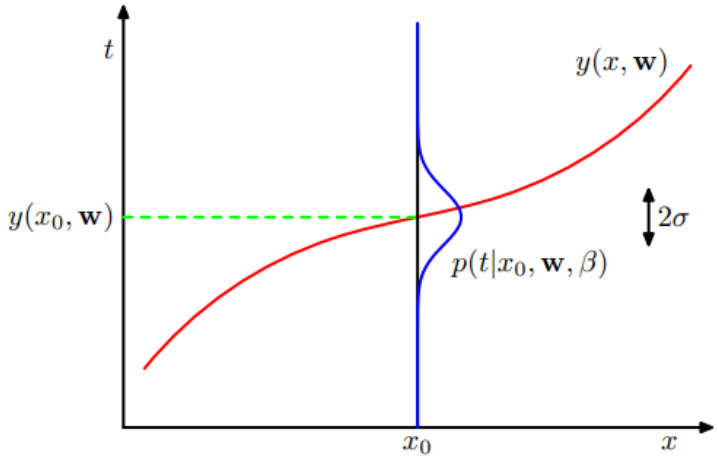
The problem with all regression models is, that we make certain assumptions about our data. First and foremost we consider our data to be true. Well, in reality our collected data are corrupted with noise. Visually this means, even if a have a fitting regression function $y(\mathbf{x}, \mathbf{w})$, the data points will most likely stray from this function. Mathematically this means:

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon,$$

Where $\epsilon$ is referred to as the noise.
This uncertainty about possible target values can be described using a conditional probability distribution $p(t|\mathbf{x})$. The most often used function to describe this probability is the Gaussian distribution $\mathcal{N}_{\mu,\sigma^2}$ where the mean is $\mu = y(\mathbf{x}, \mathbf{w})$ and the variance $\sigma^2$ is a fixed number. The probability densitty function for the Gaussian distribution is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{4}$$

In the following use the coditional probability $p(t|\mathbf{x}) = \mathcal{N}_{y(\mathbf{x},\mathbf{w}),\sigma^2}$.




### 1.2.3 Lossfunction

Let's proceed on the premise that $y(\mathbf{x}, \mathbf{w})$ describes the regression function we are looking for. Our ultimate goal is to find a function, that best fits our data points. Therefore we want to minimize the difference between our actual target values $t$ and the target vlaues $\hat{t}$ our function predicts. We do this utilizing a so called loss function, alternatively called error function. One commomly used loss function is the squared loss function. It is given by

$$L(t, y(\mathbf{x})) = (y(\mathbf{x}) - t)^2.$$

Its expected loss or average is

$$\mathbb{E}(L) = \iint (y(\mathbf{x}) - t)^2 \; p(\mathbf{x}, t) \; d\mathbf{x} \; dt.$$

The minimum of this function is obtained by calculating the derivative and its corresponding zero:

$$\frac{\delta \mathbb{E}(L)}{\delta y(\mathbf{x})} = 2 \int (y(\mathbf{x}) - t) \; p(\mathbf{x}, t) \; dt = 0$$

$$y(\mathbf{x}) = \frac{\int t p(\mathbf{x}, t) \; dt}{p(\mathbf{x})} = \int t p(t|\mathbf{x}) \; dt = \mathbb{E}(t|\mathbf{x}),$$

This implies, that the loss function is at its minimum, when $y(\mathbf{x}) = \mathbb{E}(t|\mathbf{x})$. We know about the Gaussian distribution $\mathcal{N}_{\mu,\sigma^2}$, that its average is $\mu$. Assuming $p(t|\mathbf{x})$ has Gaussian distribution with $\mu = y(\mathbf{x}, \mathbf{w})$ as noted above, we receive

$$\mathbb{E}(t|\mathbf{x}) = y(\mathbf{x}, \mathbf{w}).$$

As a result the regression function we are looking for simply is $y(\mathbf{x}, \mathbf{w})$.

### 1.2.4 Maximum-Likelihood Estimation (MLE)

We figured the regression function fitting our data will be given by $y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T\phi(\mathbf{x})$ with priorly chosen basisfunctions $\phi_i,\ i \in \{0, ..., M-1\}$. But we still do not know what the weights $\{w_j\}, j \in \{0, ..., M-1\}$ will look like. We will calculate them using maximum-likelihood estimation. Maximum-Likelihood estimation (MLE) is a mathematical procedure to estimate a certain parameter of a given probability distribution. This is done by calculating a likelihood-function, which describes the probablity an event, that has this distribution, occurs depending on the choice of parameter of the distribution. This function then will be maximized to ensure the highest probability for the regarded event.
Our "event" of interest is the target variable $\mathbf{t}$ and the parameter we want to adjust is $\mathbf{w}$. So what we want to do is to find such $\mathbf{w}$ that increases the probability for our observed data. Therefore we will once again look at the conditional distribution of $\mathbf{t}$. The resulting likelihood-function is

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}) = \prod_{n=1}^{N} p(t_n|y(x_n, \mathbf{w})) = \prod_{n=1}^{N} \mathcal{N}_{y(x_n,\mathbf{w}),\sigma^2}(t_n) = \prod_{n=1}^{N} \mathcal{N}_{\mathbf{w}^T\phi(x_n),\sigma^2}(t_n).$$

Our goal will be to maximize this function. Often one looks at the logarithm of the likelihood functions because it is more simple to work with. It suffices to maximise the logarithm beacause the result is the same as when maximizing the original function. Here the logarithm of the likelihood function is given by

$$ln(p(\mathbf{t}|\mathbf{x}, \mathbf{w})) = ln(\prod_{n=1}^{N} \mathcal{N}_{\mathbf{w}^T\phi(x_n),\sigma^2}(t_n)) = ln(\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(t_n - \mathbf{w}^T\phi(x_n))^2})$$

$$= N\ ln(\frac{1}{\sqrt{2\pi\sigma^2}}) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\phi(x_n))^2.$$

We now want to maximize this function with respect to $\mathbf{w}$. This means, that we want to minimize the term

$$\sum_{n=1}^{N}(t_n - \mathbf{w}^T\phi(x_n))^2. \tag{5}$$

From linear algebra, we know the equality $\sum_{i=1}^{n} a_i^2 = a^T a$ where $a$ is a $n$-dimensional vector. By applying this to (5) we obtain

$$(\mathbf{t} - \mathbf{\Phi w})^T(\mathbf{t} - \mathbf{\Phi w})$$

where $\mathbf{\Phi}$ is the so called desgin matrix comprising the rows $\phi(x_n)^T$ given by

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_{M-1}(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_N) & \phi_1(x_N) & \cdots & \phi_{M-1}(x_N) \end{pmatrix}.$$

As mentioned, we want to minimize $M(\mathbf{w}) := (\mathbf{t} - \mathbf{\Phi w})^T(\mathbf{t} - \mathbf{\Phi w})$. With some linear algebra we obtain the gradient $\nabla_{\mathbf{w}} M = -\mathbf{\Phi}^T\mathbf{t} + \mathbf{\Phi}^T\mathbf{\Phi w}$. By calculation the zero of $\nabla_{\mathbf{w}} M$ we receive

$$\mathbf{w}_{MLE} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\ \mathbf{\Phi}^T\mathbf{t}.$$

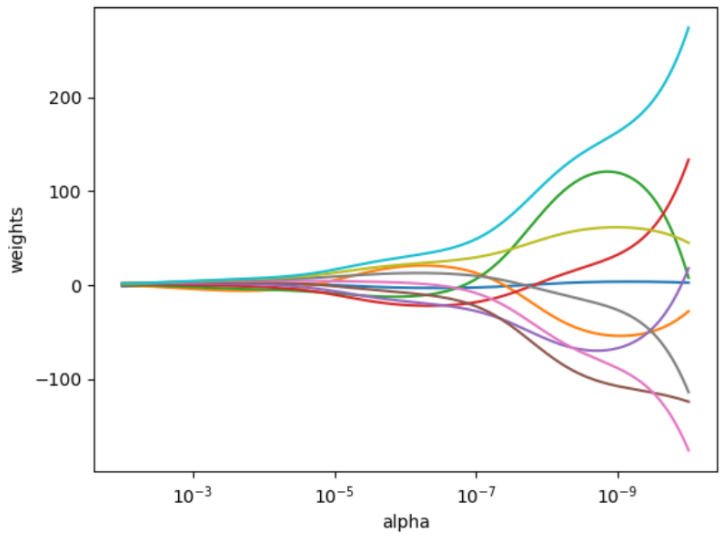
This leaves us with

$$y(\mathbf{x}, \mathbf{w}) = \mathbf{w}_{MLE}^T\mathbf{x}.$$

### 1.2.5 Regularized least squares

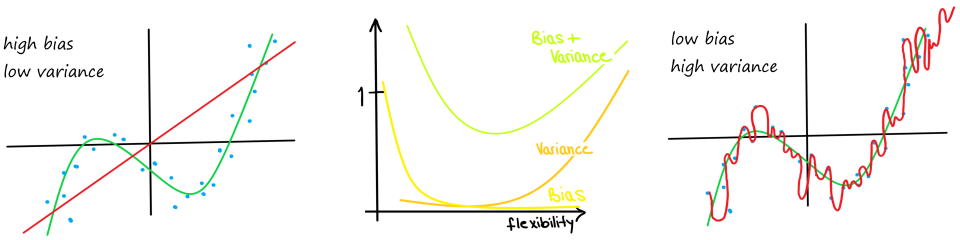
A frequent problem in regression analysis is over-fitting. Also the lossfunctions tends to get more complex than we need it to be, resulting in the need for high computational power. We can regularize the loss function by adding a regularization term. This term will drive weight values towards zero when they are not important for the model. One example for a regularization term is $\alpha E_W(\mathbf{w}) = \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$. By adding this term to our squared loss function we obtain $E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - \mathbf{w}^T\phi(x_n))^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$. And our weights will be given by

$$\mathbf{w}_{MLE} = (\mathbf{\Phi}^T\mathbf{\Phi} + \alpha I)^{-1}\ \mathbf{\Phi}^T\mathbf{t}.$$




### 1.2.6 Bias-Variance trade-off

As mentioned before, the choice of basisfunction highly impacts the complexity of our regression function and can easily lead to over-fitting. Regularizing the loss function helps prevent over-fitting, but it is not clear how to choose the regularization coefficient $\alpha$. We will not go into detail on how to choose the correct basisfunction or a suitable $\alpha$. Nonetheless we will take a short look on how the complexity of our regression function impacts the overall expected loss.
The expected loss is calculated with three different types of error: error due to noise, error due to bias and error due to variance. The error resulting from noise is irreducible, there just is inherent randomnes is our collected data. The error due to bias describes the difference between our regression function and our collected data. Choosing a less complex function might result in not matching a lot of data points and therefore a high bias. Error caused by variance describes the amout by which the predicted function will change if we change training data. If we choose a very complex function, the variance will be very high. In order to find a good balance, one consideres the function adding the loss due to variance and the loss due to bias and searches for this function's minimum.




## 1.3 Applications of linear regression

Linear regression is broadly used in machine learning. Most often it is utilized for predictions and forecasting, but one could also use it in order to analyze the impact a certain variable has on an other. There are many fields that make use of regression algorithms, for example in business, in medicine or in environmental models. Business prediction models include predicting future prices/costs (e.g. material prices or labor costs) or on the other hand expected revenues in dependency of for example advertising. When it comes to comparison of different marketing strategies or product lines, linear regression mind also find an useful application. In medicine one could investigate the effect the consumption of certain drugs have on the human body and maybe even whether a combination of different drugs could be more effective or not. Imagine you are in agriculture and you want your cow to produce as much milk as possible. In that case you could also use a linear regression model, considering different factors as for example the amount and quality of food you give your cow, the size of your cow herd, how much space your cow has or the greenness of the grass it walks on. Governments could also use regression to analyze domestic immigration related to income, crime rates and education. Even in environmental studies, linear regression finds application as researcher use it in air pollution models (average fine particle and nitrogen dioxide concentration). So as you can see there are many areas in which regression analysis, especially linear regression is used. But why is this such a popular model? Linear regression is a pretty simple model, so it is easily implemented. It is straightforward in its interpretation, and the results are comprehensible. Linear regression can be used with a wide variety of data, it is not really specified, so it can be retrained depending on the data you put in.

# References


[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[2] Friedman, Jerome, Trevor Hastie and Robert Tibshirani. *The elements of statistical learning.* https://christophm.github.io/interpretable-ml-book/limo.html

[3] scikit learn. *Linear Models.* https://scikit-learn.org/stable/modules/linear_model.html

[4] mathematicalmonk. *Machine Learning.* https://www.youtube.com/playlist?list=PLD0F06AA0D2E8FFBA

[5] towardsdatascience. *Introduction to Machine Learning Algorithms: Linear Regression.* https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a

[6] Hendrik I. Christensen. *Linear Models for Regression.* https://www.cc.gatech.edu/ hic/8803-Fall-09/slides/8803-09-lec03.pdf

[7] ML Glossary. *Linear Regression.* https://ml-cheatsheet.readthedocs.io/en/latest/linear_regression.html

[8] Wikipedia. *Regression analysis.* https://en.wikipedia.org/wiki/Regression_analysis

[9] Risi Kondor. *Regression by linear combination of basis functions* http://www.cs.columbia.edu/ jebara/4771/tutorials/regression.pdf

[10] Dan Xiang and Tyler Dae Devlin. *Lecture2.* http://cs.brown.edu/people/pfelzens/engn2520-2017/CS1420_Lecture_2.pdf

[11] Francesco Corona *Linear basis function models* https://fkorona.github.io/ATAI/2015_2/Lecture_notes/03_1_Linear_basis_function_models_draft_SEP24.pdf

[12] Keboola. *The Ultimate Guide to Linear Regression for Machine Learning.* https://www.keboola.com/blog/linear-regression-machine-learning

[13] Elle knows machines. *Lear Regression: The beginner's Machine Learning algorithm.*
https://elleknowsmachines.com/linear-regression/

[14] Environmental International. *A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide.*
https://www-sciencedirect-com.proxy-ub.rug.nl/science/article/pii/S0160412019304404