

Basic Linear Regression

Notes from 14.Nov.11, slide 1-11

The term "linear regression" refers to a numerical method that adapts the best possible function to a set of measured values. A synonym can also be the term "method of least square error" to describe the same process.

Theory and practice

What is to do to fit a line as closely as possible to a set of measurements? Of course: we must try to minimize the deviations between the predicted model and the data points.

Slide 2: the green function ($h(x_i)$) is the actual signal, function, value... which is unknown. The blue dots are samples at some state x (with an additional zero mean gaussian noise from the measurement process) which are the targets (z).

Slide 3: First we try to fit a 0th order polynomial (with only one parameter w_0) to the given data. Obviously not the best choice.

Slide 4: Now we tune our so called model $y(x, \mathbf{w})$ with another parameter. At this point we modeled a optimal linear function for the dataset but its still not satisfactory.

Slide 5: The 3rd order polynomial could be the right choice for the model. But $h(x_i)$ is still unknown and so we move on to hit all our targets with one function.

Slide 6: In this slide the function fits perfectly the target points, but shows as well a typical problem in linear regression. The high order polynomial is *overfitting* the data and will not predict further data sufficiently.

Slide 7: The simple model shown on this slide consists of a basis function ϕ and a parameter vector \mathbf{w} . The parameter w_0 allows to fix an offset in the data and is sometimes called a bias parameter. w_1, w_2, \dots, w_{M-1} can be seen as multipliers for the basis function. These variable linear parameters in combination with an basis function (typical ones are shown on the next slide) enables to form or reproduce the desired function. By using nonlinear basis functions, we allow the function $y(\mathbf{x}, \mathbf{w})$ to be a non-linear function of the input vector \mathbf{x} . Functions of the form in eq. 3 are called linear models because this function is linear in \mathbf{w} . This linearity will greatly simplify the analysis of our model.

Slide 10: Now consider a data set of inputs $\mathbf{X} = (x_1, \dots, x_N)$ with corresponding target values z_1, \dots, z_N . We group the target variables into a column vector that we denote by \mathbf{z} . Making the assumption that these data points are drawn independently from the distribution, we obtain the expression in eq. 5 for the likelihood function, which

is a function of the adjustable parameters w and σ^2 . Where the σ^2 is related to our gaussian distributed measurement noise ϵ .

To see where the formula came from see following analog computation:

$$\begin{aligned} p((z_1, x_1), (z_2, x_2)) &= p(z_1, x_1)p(z_2, x_2) = p(z_1|x_1)p(x_1)p(z_2|x_2)p(x_2) \\ &\propto p(z_1|x_1)p(z_2|x_2) = \mathcal{N}(z_1|y(x_1, w), \sigma^2)\mathcal{N}(z_2|y(x_2, w), \sigma^2) \end{aligned}$$

Slide 11: After inducing the likelihood function, we can use maximum likelihood to determine w and σ . We see that maximization of the likelihood function under a conditional Gaussian noise distribution for a linear model is the same as minimizing a sum-of-squares error function given by $E_D(w)$ which can be seen in eq. 8 or in the rear part of eq. 6. So far so good, now we need to set the gradient (eq. 7) equals zero and we obtain the eq. 9 which is known as the normal equation for the least squares problem.