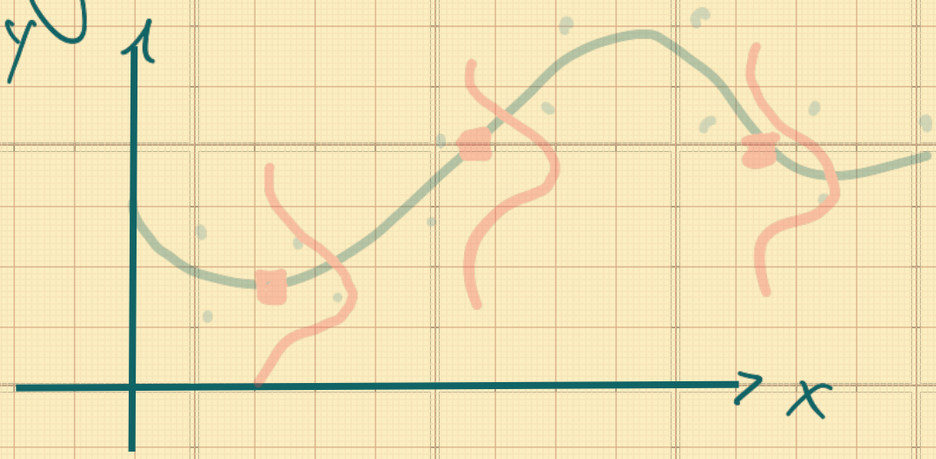
Linear Least Squares

Probalistic p.o.v.: signal/data contains noise, e.g. measurement error

→ regression with likelihood:

$$p(y|x) = \mathcal{N}(y \mid f(x), \sigma^2)$$

here, x is again the input, y is the output/target and we have $y = f(x) + \varepsilon$, where ε is an i.i.d. Gaussian noise with zero μ and variance σ^2



- Note: $f(x)$ is the hidden generating function, that we can only observe through samplings
- Note: we assume that σ^2 is known and constant
- Linear regression means linear in the parameters ☺! (features can be combined non-linear(y):

$$f(x) \approx w_1 x + w_0 \cdot 1 = x^T \theta$$

$$= \begin{pmatrix} x \\ 1 \end{pmatrix}^T \begin{pmatrix} w_1 \\ w_0 \end{pmatrix}$$

features → x, parameters → θ

so: $p(y|x,\theta) = \mathcal{N}(y|x^T\theta, \sigma^2)$, *

$y = f(x) + \varepsilon = x^T\theta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

linear in θ

Remark: if $\sigma^2 \to 0$: $p(y|x,\theta) \to \delta$

How is (*) useful in finding a model?

=> Classic estimator: MLE (see lecture statistics II)

maximizing the likelihood / prediction of the training data $\mathbb{D}$ given the unknown, but fixed parameters θ:

$$\theta_{MLE} := \arg\max_{\theta} p(\mathbb{D}, \theta)$$

so more specifically:

$$\theta_{MLE} = \arg\max_{\theta} p(\underline{Y} | \underline{X}, \theta)$$

feature

set of N training samples $\{$ $(y_1, x_{21} \dots y_N)^T$ $\rightarrow$ $\begin{pmatrix} x_1^0 & x_1^1 \\ x_2^0 & x_2^1 \\ \vdots & \\ x_N^0 & x_N^1 \dots \end{pmatrix}$

Since we assume the samples to be i.i.d., we can factorize the likelihood as:

$$p(\underline{Y} | \underline{X}, \theta) = \prod_{n=1}^{N} p(y_n | x_n, \theta)$$

$$= \prod_{n=1}^{N} \mathcal{N}(y_n | x_n^T \theta, \sigma^2) \quad **$$

Remark: $p(\underline{Y} | \underline{X}, \Theta)$ is not a pdf in Θ, but it is one in $\underline{Y}$

Instead of maximizing (**), it is numerically easier to minimize the log-likelihood. Since $\log(x)$ is strictly monotone, the extrema of x are identical to those of $\log(x)$!

[MMLB, 9.2]

NLL: $$NLL(\Theta) := -\log p(\underline{Y} | \underline{X}, \Theta)$$
$$= -\log \prod_{n=1}^{N} p(y_n | x_n, \Theta)$$
$$= -\sum_{n=1}^{N} \log p(y_n | x_n, \Theta)$$

since we have $p(y_n | x_n, \theta) = \mathcal{N}(y_n \; x^T\theta, \sigma^2)$:

$$NLL(\theta) = -\sum_{n=1}^{N} \log \mathcal{N}(y_n | x_n^T \theta, \sigma^2)$$

$$= -\sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \left(\frac{(y_n - \underbrace{x_n^T\theta}_{\mu})^2}{2\sigma^2} \right)$$

$$= -\sum_{n=1}^{N} \log \exp \left(\frac{-(y_n - x_n^T\theta)^2}{2\sigma^2} \right) + \text{const.}$$

↓ no influence on minimum location!

$$= \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - x_n^T\theta)^2 + \text{cons}'$$

so, we need to find the minimum of

$$NLL^*(\theta) = \sum_{n=1}^{N} (y_n - x_n^T \theta)^2$$

This is simply an L_2-error / quadratic form minimization as we have seen before!

To make the connection clearer, rewrite NLL^* in vector form:

$$NLL^*(\theta) = \underbrace{(\bar{Y} - \bar{X}\theta)^T (\bar{Y} - \bar{X}\theta)}_{r^T r}$$

We have seen before that this can be seen as a quadratic form, for which the minimum is found by solving the associated linear system, which leads again to the

normal equation of LLS:

$$\theta_{MLE} = \frac{\bar{X}^T \bar{y}}{\bar{X}^T \bar{X}}$$

→ see talk on normal form.

→ More on MLE and MAP estimator + regularized least squares: Bishop Sec 3.1.