2.1.4 Nonlinear Regression

The approximation of data points (x_i, y_i) by a linear function of the form $\hat{f}(x) := wx + b$ is parametric, meaxning that the sought linear function depends only on certain parameters w, b that need to be determined. Once w and b are known, $wx + b$ can be easily computed for any input x.
However, if data pairs do not follow the model $y_i = wx_i + b + \epsilon_i$, but more generally $y_i = f(x_i) + \epsilon_i$ for $i = 1, \dots, N$ with a nonlinear function f, linear regression is not a good model

Polynomial Regression

One possible solution is polynomial regression, meaning one chooses the basis function

$$\hat{f}(x) := \sum_{j=0}^{p} \beta_j x^j$$

(where p is the polynomial degree)
and tries to determine β_j. x_j are the polynomial terms(e.g. $x^0 = 1, x^1 = x, x^2$ etc) The corresponding least squares problem still has the form

$$\min_{\beta \in \mathbb{R}^{p+1}} \frac{1}{2} ||X\beta - y||^2$$

(p+1 because "+1" comes from the **constant term** (β_0) which corresponds to x^0. Even if the polynomial degree is p, we still need to include the constant term, making the total number of coefficients $p + 1$)
with a matrix X of the form

$$X = \begin{pmatrix} 1 & x_1 & \dots & x_1^p \\ \vdots & & \vdots & \\ 1 & x_N & & x_N^p \end{pmatrix}$$

Each row corresponds to a data points (x_i, y_i) and each column corresponds to a polynomial term x^j. Since there are $p + 1$ polynomial terms ($x^0, x^1, \dots, x^p$), the matrix X has $p + 1$ columns The matrix X is known as a Vandermonde matrix and has the following properties:

1. If $N = p + 1$ (number of data points equals number of coefficients), then X is invertible if and only if all x_i are different
2. If $N > p + 1$ (more data points than coefficients), then $X^T X$ is invertible if and only if there are p different x_i
3. If $N < p + 1$ (fewer data points than coefficients), then $X^T X$ is not invertible

Mathematically, polynomial regression is thus very similar to linear regression. Disadvantages of polynomial regression are that it tends to suffer from overfitting, meaning that if the polynomial degree p is too high, noise in the data is exactly represented by the polynomial. Additionally, generalization to inputs $x_i \in \mathbb{R}^d$ is non-trivial

Next: 2.1.4.1 Neural Networks