Hoofdstuk 1

Machine Learning

1.1 Wat is machine learning?

Machine learning is een tak binnen het gebied van de artificiële intelligentie, waarbij we een model trainen op basis van een gegeven dataset. Naarmate de training vordert, zal het model bepaalde verbanden beginnen leggen en bepaalde structuren beginnen herkennen in de data waarop het getraind wordt. In principe neemt de kwaliteit van het model toe wanneer we de data, waarop het model zich baseert, toeneemt.

In ons geval zullen we dus een model trainen dat verbanden zal zoeken in de dataset met tumoren van borstkankerpatiënten. Het model zal trachten een link te vinden tussen de verschillende kenmerken van deze tumoren en de klasse waartoe deze tumoren behoren; die van de goedaardige of die van de kwaadaardige.

1.2 Terminologie en notatie

De dataset beschikt n verschillende datapunten. Deze individuele datapunten x_i hebben elk een eindig aantal kenmerken, die we *features*, *inputs* of *onafhankelijke variabelen* zullen noemen. We duiden de hoeveelheid *features* aan met p. Elk individueel datapunt x_i uit onze set van datapunten $x_1, x_2, ..., x_n$ is dus van de vorm $x_i(x_1, x_2, ..., x_p)$ (met $1 \leq i \leq n$). Deze features van de datapunten kunnen we voorstellen aan de hand van een matrix X, waarvoor geldt:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{np} \end{bmatrix}$$

Hierbij is X_{ij} de meetwaarde van de j-e feature voor het i-e datapunt met met $i = 1, 2, 3, ..., n$ en $j = 1, 2, 3, ..., p$. De tumordataset bevat gegevens over 569 tumoren, die elk 30 kenmerken hebben. In de dataset is dus $n = 569$ en $p = 30$.

Aangezien we aan *supervised learning* doen (hier gaan we in Hoofdstuk **??** verder op in), hebben we voor elke input x_i in de dataset ook een bijhorende *output* y_i. We hebben dus n y-waarden $y_1, y_2, ..., y_n$, die we - net zoals de features - matricieel kunnen voorstellen als volgt:

$$Y = \begin{bmatrix} y_1 & y_2 & y_3 & \dots & y_n \end{bmatrix}$$

Elke tumor is ofwel kwaadaardig, ofwel goedaardig, dus elke individuele output y_i is ofwel gelijk aan 1 (goedaardig), ofwel gelijk aan -1 (kwaadaardig).

1.3 Predictie

Na het trainen van ons model, hebben we hopelijk een verband gevonden tussen de verschillende *features* van de tumoren en hun aard. Indien we ook het validatieproces (zie **??**) doorlopen, kunnen we ons model toepassen op een nieuwe *batch* datapunten om zo voorspellingen te kunnen doen over de aard van deze - door het model nog niet eerder geziene - tumoren. We noemen dit *predictie*.