Maschinelles Lernen

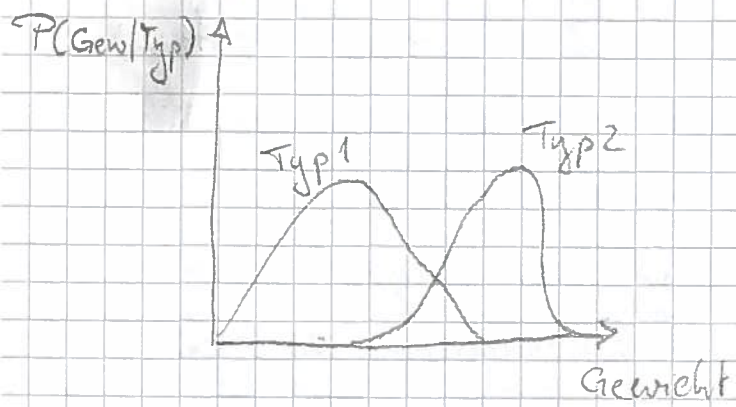
Wir wollen an Hand von Beispielen lernen, verschiedene Dinge voneinander zu unterscheiden.
Da die Problemstellung so allgemein ist, sind die Anwendungen vielfältig. Wir alle benutzen zum Beispiel Spamfilter, die E-Mails in erwünscht und unerwünscht klassifizieren können. Amazon unterscheidet zwischen Produkten, die uns interessieren und solchen, die es nicht tun. In industriellen Anwendungen möchte man Objekte auf Förderbändern nach ihrer Art sortieren, in der Videoüberwachung möchte man Gesichter vom Hintergrund unterscheiden.
Wir beginnen mit einem einfachen Beispiel und überlegen uns, wie man Entscheidungen so trifft, dass der ~~Fehler~~ Klassifizierungsfehler möglichst klein ist.

Fische unterscheiden, ohne zu lernen

Wir wollen eine Maschine bauen, die in einer Fischfabrik zwei Sorten Fisch voneinander unterscheidet.
Zunächst müssen wir uns fragen, welche Eigenschaften der Fische geeignet wären. Wir suchen Merkmale, die einfach zu messen sind und bei den beiden Arten möglichst verschieden.

Am einfachsten zu ermitteln – man muss nur zählen – ist die a-priori Wahrscheinlichkeit einen Typ 1 Fisch zu sehen. Haben wir sonst keine Informationen, sollten wir stets behaupten es handele sich um die a-priori wahrscheinlichere Art Fisch. Wenn etwa Typ 1 Fische 60% des Fangs ausmachen, so haben wir eine Trefferquote von 60%. Gelegentlich "Typ 2" zu sagen kann das nur verschlechtern!

Durch Messen weiterer Merkmale können wir natürlich besser antworten. Wiegen wir die Fische bekommen wir ein Bild wie dieses



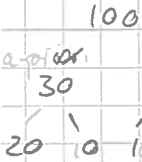
Wenn wir so die Wahrscheinlichkeit bestimmt haben, dass es sich um einen Typ 1 Fisch handelt, gegeben ein Gewicht von x kg, wollen wir natürlich wieder "1" sagen genau wenn

$$P(\text{Typ 1} \mid \text{Gewicht}=x) > P(\text{Typ 2} \mid \text{Gewicht}=x)$$

hier beachten wir aber noch die a-priori Wahrscheinlichkeit nicht.

Durch Wiegen der beiden Fischarten können wir die Gewichtsverteilung der beiden bestimmen.
Wir messen also $P(\text{Gewicht}=x \mid \text{Typ})$. Wir interessieren uns eigentlich für die umgekehrten Werte

also $P(\text{Typ} \mid \text{Gewicht}=x)$. Um das zu bekommen wenden wir die Formel von Bayes an



$$P(\text{Typ}=1 \mid \text{Gewicht}=x) = \frac{P(\text{Gewicht}=x \mid \text{Typ}=1) \cdot P(\text{Typ } 1)}{P(\text{Gewicht}=x \mid \text{Typ}=1) + P(\text{Gewicht}=x \mid \text{Typ } 2)}$$

Um zu klassifizieren prüfen wir einfach welche Wahrscheinlichkeit größer ist.
Wenn wir mehrere ~~Hofswerte~~ Merkmale haben, können wir die selbe Regel einfach mehrmals anwenden. ~~Ohne~~ Ohne Indizien haben wir $P(\text{Typ } 1)$ = a-priori Wahrscheinlichkeit, die wir dann nach und nach mit Bayes Regel anpassen.

Bsp 2 >
Dafür ist es natürlich wichtig, dass die Merkmale unabhängig voneinander sind. Sowohl Gewicht als auch Größe der Fische zu messen gibt weniger Informationen als etwa Gewicht und Farbe.

Spam
Spamerkennung funktioniert nach diesem Prinzip.
Die Programmierer haben eine Reihe von Merkmalen festgelegt, z.B. die Häufigkeitsverteilung von verschiedenen Worten, oder die Anzahl der Links im Text.
Jedes Mal wenn eine E-Mail von Ihnen als Spam oder nicht-Spam markiert wird, ~~zählt~~ aktualisiert das E-Mail Programm diese Wahrscheinlichkeits-verteilungen.
Mit etwas raffinierteren Tricks wird dafür gesorgt, dass falsch geschriebene Worte den Filter nicht so leicht trügen.

Nächste Nachbarn

Neben diesen statistischen Methoden gibt es auch andere Verfahren. Hierbei werden die ~~Daten~~ Objekte als Punkte in einem hochdimensionalen Raum interpretiert, das heißt jedes gemessene Merkmal wird mit einer Koordinate in diesem Raum identifiziert. Hat man nur zwei Merkmale kann man das also aufzeichnen

Die Aufgabe ist dann einen neuen Punkt einer Klasse zuzuordnen. Die Intuition ist hierbei, dass Punkte der selben Klasse nah beieinander sind und weit entfernt von den Punkten anderer Klassen.
Daraus ergibt sich sofort ein einfaches Verfahren um Punkte zu klassifizieren:

Man berechnet die Abstände zu allen anderen Punkten, sucht so den nächsten Nachbarn und gibt dem neuen Punkt die selbe Klasse.
Da Meßwerte aber typischerweise verrauscht sind, kann es sein, dass der nächste Nachbar eine ~~Fehl~~klassifikation hervorruft. Dieses Problem kann man abmildern, indem man sich die nächsten k-Nachbarn, für $k > 1$ ansieht und dann einen Mehrheitsentscheid macht. k darf natürlich auch nicht zu groß sein, weil man sonst Punkte aus einer anderen Klasse mit ~~bt~~ betrachtet.

Ein großer Nachteil dieses Verfahrens ist, dass der Rechenaufwand enorm groß ist. Für jeden zu klassifizierenden Punkt muss man sich die gesamte Datenmenge ansehen. Außerdem passiert hier im Prinzip kein "Lernen" die Daten werden vollständig gespeichert, es gibt keinen Abstraktionsschritt.

Nimmt man an, dass die Punktewolke jeder Klasse ungefähr kugelig ist, so kann man sich einen großen Teil des Rechenaufwands sparen, wenn man einmal das Zentrum der Kugel ausrechnet und anschließend alle anderen Punkte wegschmeißt.
Dann braucht man einen neuen Punkt nur noch mit den Zentren zu vergleichen.

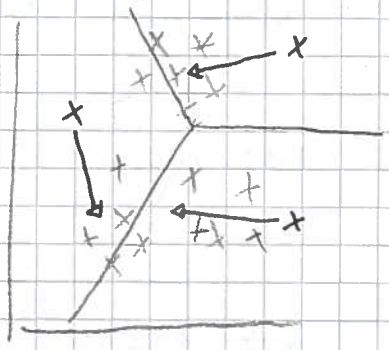
Raffiniertere Verfahren versuchen ~~Linien~~ Kurven durch den Raum zu finden, die die Punktklassen möglichst gut voneinander trennen. Dabei muss man aber darauf achten, dass die Kurven einfach bleiben und nicht jeden (verrauschten!) Punkt beachten.

Lernen ohne Lehrer

Die Verfahren, die wir bisher betrachtet haben, gingen davon aus, dass wir Zugang zu einer möglichst großen Zahl von fertig klassifizierten Beispielen haben, In der Praxis ist das oft nicht gegeben.

Es gibt verschiedene Verfahren, die Gruppen von ähnlichen Punkten automatisch erkennen können. Ein einfaches Verfahren ist "k-Means"

Es funktioniert, indem man sich k zufällige Punkte wählt. Jetzt rechnet man für alle Trainingsdaten aus, welcher von den k-Punkten ihr nächster Nachbar ist. Anschließend werden die k-Punkte in das Zentrum der Punktemenge geschoben, die sie als nächsten Nachbarn hat.



Das wiederholt man so lange bis sich die Punkte nur noch wenig bewegen. Mit ein wenig Glück sitzen sie dann in den Zentren der Punktewolken, die zu den verschiedenen Klassen gehören.