Klassifikation

Wir haben zwei Mengen: S, die Menge von Objekten, und C, die Menge von Klassen.

Jedes Objekt ist dabei ein Vektor mit D reelen Zahlen - dessen Features.

Wir definieren die Fkt. $classify : S \to C$, sodass jedem Objekt die richtige Klasse zugeordnet wird. Diese Fkt. kennen wir nicht, wollen sie aber so gut wie möglich approximieren.

Stattdessen haben wir ein Modell: $model : \mathbb{R}^D \to C$, wobei D die Anzahl von Attributen/Features pro Objekt ist.

Außerdem definieren wir: $feature : S \times \{1, \dots, D\} \to \mathbb{R}$

Wir definieren die Hilsfkt. $card(M)$, welche die Anzahl von Elementen in der Menge M ausgibt.

Damit können wir die Genauigkeit unseres Modells bestimmen:

$$accuracy := \frac{card(\{o \in S | classify(o) = model(feature(o,1), \dots, feature(o,D))\})}{card(S)}$$

accuracy soll maximiert werden, kann aber nicht einfach abgeleitet werden (da die Fkt. nicht smooth ist), deswegen nutzen wir nachher Wahrscheinlichkeiten.

Aber zuerst eine geschickte Methode um das Maximum bei einer ablteitbaren Fkt. zu finden:

Gradient Ascent

Gegeben: $f : \mathbb{R}^n \to \mathbb{R}$

Außerdem haben wir die Fkt. argmax:

$$\hat{x} = \backslash\mathrm{argmax}_{x \in \mathbb{R}^n} f$$

Dabei ist $\hat{x}$ der Wert in $\mathbb{R}^n$ der $f(x)$ maximiert. Formal definiert:

$$\forall x \in \mathbb{R}^n : f(x) \leq f(\backslash\mathrm{argmax}_{x \in \mathbb{R}^n} f)$$

Die Ableitung von f ist $\nabla f :=< \frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} >$

Entsprechend gilt: $\nabla f(\hat{x}) = 0$

Der Gradient Ascent ist nun eine numerische Reihe, die uns zum Wert $\hat{x}$ führt:

$x_{n+1} := x_n + \alpha \cdot \nabla f(x_n)$ für alle $n \in \backslash\mathrm{N}_0$

Logistische Regression

Idee: Wir bestimmen die Wahrscheinlichkeit der Genauigkeit, da diese abgeleitet werden kann -> und wir dann den Hochpunkt (mittels Gradient Ascent) finden können.

Wir passen unser Modell an: $model(o, w)$, wobei w ein Vektor von Gewichten ist. Das Modell besteht nur aus den Gewichten im Vektor w.

Um die Wahrscheinlichkeit zu bekommen, brauchen wir die Sigmoid Fkt.:

$$S(t) = \frac{1}{1 + exp(-t)}$$

Zur Erinnerung, $exp(x) = e^x$

Eigenschaften der Sigmoid Fkt.:

- $lim_{t \to \infty} S(t) = 1$
- $lim_{t \to -\infty} S(t) = 0$
- $0 \leq S(t) \leq 1$
- $S(0) = \frac{1}{2}$

Wir beweisen: $S(-t) = 1 - S(t)$

1. $1 - S(t) = 1 - \frac{1}{1+exp(-t)}$
2. $1 - S(t) = \frac{1+exp(-t)}{1+exp(-t)} - \frac{1}{1+exp(-t)}$
3. $1 - S(t) = \frac{1+exp(-t)-1}{1+exp(-t)}$
4. $1 - S(t) = \frac{exp(-t)}{1+exp(-t)}$
5. $1 - S(t) = \frac{exp(-t)}{1+exp(-t)} \cdot 1$
6. $1 - S(t) = \frac{exp(-t)}{1+exp(-t)} \cdot \frac{exp(t)}{exp(t)}$
7. $1 - S(t) = \frac{exp(-t) \cdot exp(t)}{(1+exp(-t)) \cdot exp(t)}$
8. $1 - S(t) = \frac{exp(-t+t)}{1 \cdot exp(t) + exp(-t) \cdot exp(t)}$
9. $1 - S(t) = \frac{1}{exp(t)+1}$
10. $1 - S(t) = \frac{1}{1+exp(-(-t))}$
11. $1 - S(t) = S(-t)$

Jetzt beweisen wir: $exp(-t) = \frac{1-S(t)}{S(t)}$

1. $S(t) = \frac{1}{1+exp(-t)}$
2. $S(t) \cdot (1 + exp(-t)) = 1$
3. $1 + exp(-t) = \frac{1}{S(t)}$
4. $exp(-t) = \frac{1}{S(t)} - 1$
5. $exp(-t) = \frac{1}{S(t)} - \frac{S(t)}{S(t)}$
6. $exp(-t) = \frac{1-S(t)}{S(t)}$

Wir schauen uns jetzt den Kehrwert an, um das Inverse der Sigmoid Fkt. zu finden:

1. $exp(t) = \frac{S(t)}{1-S(t)}$
2. $t = ln(\frac{S(t)}{1-S(t)})$
3. $S^{-1}(y) = ln(\frac{y}{1-y})$

Jetzt leiten wir die Sigmoid Fkt. ab:

- $S'(t) = -\frac{-exp(-t)}{(1+exp(-t))^2}$
- $S'(t) = -(-exp(-t)) \cdot \frac{1}{(1+exp(-t))^2}$
- $S'(t) = exp(-t) \cdot S(t)^2$
- $S'(t) = \frac{1-S(t)}{S(t)} \cdot S(t)^2$
- $S'(t) = (1 - S(t)) \cdot S(t)$
- $S'(t) = S(-t) \cdot S(t)$

Wir definieren außerdem die Logit Fkt. als den natürlichen Logarithmus der Sigmoid Fkt.:
$L(t) := ln(S(t))$

Die Ableitung der Logit-Fkt. ist $L'(t) = S(-t)$. Beweis:

- $L'(t) = \frac{S'(t)}{S(t)}$
- $L'(t) = \frac{S(-t) \cdot S(t)}{S(t)}$
- $L'(t) = S(-t)$

Jetzt können wir endlich die Wahrscheinlichkeit definieren:

Wir nehmen jetzt eine binäre Klassifikation an, damit wir sagen können:
$classify(\vec{x}) = y = +1$ oder -1

Damit gilt $P(y = +1|\vec{x}, \vec{w}) = S(\vec{x} \cdot \vec{w})$, dabei ist $\vec{x}$ das zu klassifizierende Objekt und w der Gewichtsvektor.

Dabei ist $\vec{x} \cdot \vec{w}$ das Skalarprodukt, also $\sum_{i=1}^{D} x_i \cdot w_i$.

Außerdem gilt:

$P(y = -1|\vec{x}, \vec{w}) = 1 - P(y = +1|\vec{x}, \vec{w}) = 1 - S(\vec{x} \cdot \vec{w}) = S(-\vec{x} \cdot \vec{w})$

Entsprechend können wir die beiden Gleichungen zusammenfassen:

$$P(y|\vec{x}, \vec{w}) = S(y \cdot (\vec{x} \cdot \vec{w}))$$

Wir haben eine Liste von N Trainingsdaten, die jeweils ein Objekt mit einer Klasse verbinden. Also haben wir eine Matrix X und einen Vektor $\vec{y}$

Wir suchen $\vec{w}$ sodass die likelihood $l(X, \vec{y}, \vec{w})$ maximiert wird.
Die likelihood Fkt. ist wie folgt definiert:

$$l(X, \vec{y}, \vec{w}) = \prod_{i=1}^{N} P(y_i | \vec{x}_i, \vec{w})$$

Da es zum Ableiten leichter wird, nutzen wir stattdessen den Logarithmus der Likelihood Fkt. - Die Log-Likelihood Fkt. ist dann:

$$ll(X, \vec{y}, \vec{w}) = ln(l(X, \vec{y}, \vec{w}))$$

Wir nutzen die Eigenschaft vom Logarithmus, dass $ln(a \cdot b) = ln(a) + ln(b)$

$$ll(X, \vec{y}, \vec{w}) = ln(\prod_{i=1}^{N} P(y_i | \vec{x}_i, \vec{w})) = \sum_{i=1}^{N} ln(P(y_i | \vec{x}_i \vec{w})) = \sum_{i=1}^{N} ln(S(y_i \cdot (\vec{x}_i \cdot \vec{w}))) = \sum_{i=1}^{N} L(y_i \cdot (\vec{x}_i \cdot \vec{w}))$$

Jetzt leiten wir den Log-Likelihood ab, um dessen Maximum bestimmen zu können:

$\frac{\partial}{\partial w_j} ll(X, \vec{y}, \vec{w})$ für alle $j \in \{1, \dots, D\}$

Wir definieren uns erst eine Hilfs-Fkt.:

$h(\vec{w}) := \vec{x}_i \cdot \vec{w} = \sum_{k=1}^{D} x_{i,k} \cdot w_k$

Diese Hilfs-Fkt können wir jetzt ableiten nach w_j:

$$\frac{\partial}{\partial w_j} h(\vec{w}) = \frac{\partial}{\partial w_j} \sum_{k=1}^{D} x_{i,k} \cdot w_k = \sum_{k=1}^{D} x_{i,k} \cdot \frac{\partial}{\partial w_j} w_k = \sum_{k=1}^{D} x_{i,k} \cdot \delta_{j,k} = x_{i,j}$$

Damit können wir nun den Log-Likelihood ableiten:

$$\frac{\partial}{\partial w_j} ll(X, \vec{y}, \vec{w}) = \frac{\partial}{\partial w_j} \sum_{i=1}^{N} L(y_i \cdot (\vec{x}_i \cdot \vec{w}))$$

$$= \sum_{i=1}^{N} (\frac{\partial}{\partial w_j} y_i \cdot (\vec{x}_i \cdot \vec{w})) \cdot \frac{\partial}{\partial w_j} L(y_i \cdot (\vec{x}_i \cdot \vec{w}))$$

$$= \sum_{i=1}^{N} (y_i \cdot \frac{\partial}{\partial w_j} h(\vec{w})) \cdot S(-y_i \cdot (\vec{x}_i \cdot \vec{w}))$$

$$= \sum_{i=1}^{N} y_i \cdot x_{i,j} \cdot S(-y_i \cdot (\vec{x}_i \cdot \vec{w}))$$

Wir haben dann eine Gleichung pro Gewicht im Vektor $\vec{w}$ (also D Gleichungen). Die können wir nicht direkt lösen, wir können aber die Methode des Gradient Ascents verwenden, um einen Wert für $\vec{w}$ zu finden, deer die Log-Likelihood (und damit auch die Likelihood) Fkt. maximiert.