Vorlesung 12 Maschinelles Lernen

Was ist Maschinelles Lernen (ML)?

⇝ ML ist ein Teilgebiet der Datenanalyse.

⇝ Die grundlegende Idee in ML ist es Algorithmen zu entwickeln, die automatisch Informationen aus Daten berechnen. D.h. wir wollen Methoden entwickeln, die in jedem speziellen Szenario angewandt werden können, so dass wir nicht jedes mal per Hand entwickeln müssen.

⇝ Die Automatisierung eines Algorithmus geschieht durch das das Lernen von Daten. Die Daten werden in die automatische Entwicklung des Alg. miteinbezogen.

⇝ Daten werden hier als numerische Vektoren verstanden.

Die mathematische Abstraktion eines ML Problems ist:

> Gegeben seien $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^D \times \mathbb{R}^N$.
>
> Ziel: Finde eine Funktion $f: \mathbb{R}^D \to \mathbb{R}^N$, s.d.
>
> 1. $f(x_i) \approx y_i$, $1 \le i \le n$
> 2. für alle neuen Datenpunkte $(x, y) \in \mathbb{R}^D \times \mathbb{R}^N$: $f(x) \approx y$.

In vielen Datensätzen müssen wir Messfehler annehmen.

Verwendung von $\approx$ erlaubt uns Flexibilität.

$\approx$ wird im ML durch eine sogenannte Loss-Funktion modelliert (siehe unten).

Definition 12.1.

Seien $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^p \times \mathbb{R}^N$ Daten wie oben.

1. Die x_i heißen **Input-Variablen**
2. Die y_i heißen Labels, Output-Variablen oder **Response-Variablen**.

Beispiel 12.2

$x_i = (x_i^{(1)}, x_i^{(2)}, x_i^{(3)})$. $(D = 3)$

$x^{(1)}$ = Abschluss	$x^{(2)}$ = Wohnort	$x^{(3)}$ = Alter	y = Jahreseinkommen
MSc	Osnabrück	36	60145 €
PhD	Osnabrück	24	72 541 €
BSc	Hannover	31	58.901 €
MSc	Bremen	29	61.005 €

Wir können diese Daten in numerische Daten umwandeln:

$x^{(1)}$ = Abschluss	$x^{(2)}$ = Längengrad	$x^{(2)}$ = Breit.	$x^{(3)}$ = Alter	y = Jahreseinkommen
2	8,05	52,28	36	60145
3	8,05	52.28	24	72 541
1	9,73	52,38	31	58.901
2	8,8	53,1.	29	61.005

Definition 12.3

Variablen mit einem kontinuierlichen Werkbereich heißen **kontinuierliche** Var.
Variablen mit einem diskreten Werkbereich heißen **diskrete** Variablen oder **kategorische** Variablen.

Definition 12.4

Seien $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^D \times \mathbb{R}^N$ Datenpunkte.

1. Ein <u>deterministisches Modell</u> ist eine Funktion $f_\Theta : \mathbb{R}^D \to \mathbb{R}^N$, die von einem Parameter $\Theta \in \mathbb{R}^P$.
2. Ein <u>statistisches Modell</u> ist eine bedingte Wahrscheinlichkeitsverteilung für $(y|x)$, die von einem Parameter Θ abhängt. Wir nehmen dabei an, dass $(y|x)$ eine Dichte $p_\Theta(y|x)$ hat, die von $\Theta \in \mathbb{R}^P$ abhängt.

Wir nehmen außerdem an, dass die n Datenpunkte unabhängig voneinander gezogen wurden, so dass

$$p_\Theta((y_1, \dots, y_n) \mid (x_1, \dots, x_n)) = \prod_{i=1}^{n} p_\Theta(y_i \mid x_i).$$

Das Ziel in ML ist es in einem Modell einen Parameter $\Theta^* \in \mathbb{R}^P$ zu wählen, der die Daten "gut" beschreibt. Außerdem soll Θ^* auch unbeobachtete Daten "gut" beschreiben. Dazu teilen wir die Daten in <u>Trainingsdaten</u> und <u>Testdaten</u>. Die Trainingsdaten werden verwendet um Θ^* zu <u>lernen</u>. Die Testdaten simulieren unbeobachtete Daten und wir evaluieren unser berechnetes Modell auf ihnen. Dieser letzte Schritt heißt <u>Validierung</u>.

Insgesamt haben wir folgende Schritte zum Lösen eines ML Problems:

1. Wahl des Modells.
2. Aufteilen der Daten in Training- und Testdaten
3. Lernen der Parameter $\Theta \in \mathbb{R}^P$
4. Validierung.

Für Schritt 4 verwenden wir eine Qualitätsfunktion um die Güte des Modells zu beurteilen. Zwei konkrete Qualitätsfunktionen sind wie folgt:

Definition 12.5

Gegeben seien Daten $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^D \times \mathbb{R}^N$ und ein Modell f_θ oder P_θ. Sei $\ell : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}_{\geq 0}$ eine Loss-Funktion.

Das empirische Risiko bzgl. ℓ ist

$$R(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)). \quad \text{bzw.}$$

$$R(\theta) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\hat{y}_i \sim P_\theta(y|x_i)} \ell(y_i, \hat{y}_i).$$

Der Root-Mean-Square-Error (RMSE) ist:

$$RMSE(\theta) = \sqrt{R(\theta)} \quad \text{für } \ell(y, \hat{y}) = \|y - \hat{y}\|^2.$$

Der einfachste Weg ein Modell zu wählen, ist die Daten zufällig nach Training- und Testdaten zu teilen, dann einen Parameter θ^* zu berechnen und das Ergebnis durch eine Qualitätsfunktion $Q(\theta)$ zu beurteilen. Das Modell, für welches der Unterschied in $Q(\theta)$ zwischen Training- und Testdaten am kleinsten ist, wird gewählt. Die zufällige Wahl hier und in Punkt 2. sollen unabhängig sein. Alternativ können wir Kreuz-Validierung benutzen.

Für Schritt 2 können wir die Daten zufällig aufteilen (unabh. von Schritt 1). Eine übliche Aufteilung sieht 50% - 80% der Daten für Training.

Zu 3. Wie lernen wir Parameter?

Im deterministischen Modell verwenden wir Empirical Risk Minimization (ERM). D.h. wir berechnen $\Theta \in \mathbb{R}^p$ mit

$$\Theta \in \operatorname{argmin}_{\Theta \in \mathbb{R}^p} R(\Theta).$$

für eine gewählte Loss-Funktion und unter Verwendung der Trainingsdaten.

Im statistischen Modell verwenden wir Maximum-Likelihood (ML) oder Maximum a-Posteriori (MAP) Schätzung. Diese sind gegeben durch Maximierung folgender Funktionen.

Definition 12.6.

Seien $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^D \times \mathbb{R}^N$ Daten und P_Θ ein stat. Modell

1. Die Likelihood-Funktion ist die Wkt. von $(y_1, \dots, y_n)$ gegeben $(x_1, \dots, x_n)$: $L(\Theta) = \prod_{i=1}^{n} P_\Theta(y_i | x_i)$. Die log-Likelihood Funktion ist dann

$$\ell(\Theta) = \log L(\Theta) = \sum_{i=1}^{n} \log P_\Theta(y_i | x_i).$$

2. Sei $X = [x_1 - x_n]^T \in \mathbb{R}^{n \times D}$, $Y = [y_1, \dots, y_n]^T \in \mathbb{R}^{n \times N}$.
Ang. Θ ist selbst zufällig und dass (Θ, X, Y) eine Dichte hat. Die posteriori Funktion ist.

$$\alpha(\Theta) = P(\Theta | X, Y).$$