Fachhochschule Graubünden
University of Applied Sciences

Prof. Dr. Yves Staudt
Machine Learning

Introduction to Machine Learning

Musterlösung zum Aufgabenblatt

Machine Learning Grundbegriffe

Hinweis: Die folgenden Antworten sind Musterlösungen und zeigen das erwartete Niveau. Bei den offenen Fragen sind auch andere, gut begründete Antworten gültig.

Lern-Story: Bevor wir einzelne Algorithmen kennenlernen, brauchen wir ein gemeinsames Vokabular: Wann *lernt* ein System überhaupt aus Daten, und welche Arten des Lernens gibt es? Ein Spam-Filter, eine Kundensegmentierung und ein schachspielender Agent „lernen" auf sehr unterschiedliche Weise. In diesem Aufgabenblatt ordnen Sie die zentralen Grundbegriffe des Machine Learning ein — von der Art der Überwachung (supervised, unsupervised, semi-supervised, reinforcement) über die Art des Trainings (batch vs. online) bis zur Frage, wie ein Modell generalisiert (instanz- vs. modellbasiert, Over- vs. Underfitting). Ausserdem ordnen Sie Machine Learning in den grösseren Rahmen von *Künstlicher Intelligenz* und *Deep Learning* ein und unterscheiden die typischen Aufgaben (Regression, Klassifikation, Clustering).

Lernziel: Nach Abschluss dieses Teils kann der/die Absolvent:in die wichtigsten Grundbegriffe des Machine Learning erklären, voneinander abgrenzen und typischen Anwendungsfällen zuordnen, Machine Learning von Künstlicher Intelligenz und Deep Learning abgrenzen und typische ML-Aufgaben (Regression, Klassifikation, Clustering) unterscheiden.

Input:

- Folien *Introduction to Machine Learning*
- Videos zum Modul
- Begleitende Lektüre: Géron (2019), *Hands-On Machine Learning*, Kap. 1 (Typen von ML-Systemen)

Output:

- Sie können die Grundbegriffe des Machine Learning beschreiben, gegeneinander abgrenzen und an Beispielen anwenden.

Lernerfolg: Bestimmen Sie Ihren Lernerfolg.

Aufgabe 1: Wiederholung

Fassen Sie in eigenen Worten zusammen, was Sie in diesem Teil gelernt haben. Welche Grundbegriffe waren neu, und wie stufen Sie deren Wichtigkeit ein? Nennen Sie zu jeder Lernart (überwacht, unüberwacht, halbüberwacht, Reinforcement) ein eigenes Beispiel.

Musterlösung. Machine Learning bedeutet, dass ein System Muster *aus Daten* lernt, statt explizit programmiert zu werden. Die Grundbegriffe lassen sich auf drei Achsen ordnen:

- **Art der Überwachung:** *überwacht* (mit Zielvariable, z. B. Spam-Filter), *unüberwacht* (ohne Zielvariable, z. B. Kundensegmentierung), *halbüberwacht* (wenige beschriftete + viele unbeschriftete Daten, z. B. Foto-Tagging), *Reinforcement* (Agent lernt durch Belohnung, z. B. Schach).
- **Art des Trainings:** *Batch* (auf einem festen Datensatz) vs. *Online* (laufend mit neuen Daten).
- **Art der Generalisierung:** *instanzbasiert* (vergleicht mit gespeicherten Beispielen) vs. *modellbasiert* (abstrahiert ein Modell); dabei sind *Over-* und *Underfitting* die zentralen Risiken.

Zusätzlich ordnet sich ML in den Rahmen *KI* $\supset$ *ML* $\supset$ *Deep Learning* ein; eng verwandt mit halbüberwachtem Lernen ist das *self-supervised* Lernen, bei dem das Modell seine Labels aus den Daten selbst erzeugt (Grundlage von LLMs).

Aufgabe 2: Selbstwiederholung — Begriffe zuordnen

Ordnen Sie jedem Begriff (A–H) die passende Beschreibung (1–8) zu.

Begriff	**Beschreibung**
A Überwachtes Lernen	1 Maximiert eine kumulative Belohnung durch Interaktion mit einer Umgebung.
B Unüberwachtes Lernen	2 Trainiert auf dem vollständigen Datensatz; das Modell bleibt danach unverändert.
C Halbüberwachtes Lernen	3 Lernt aus Daten *mit* Zielvariable (Input + Output).
D Reinforcement Learning	4 Erstellt ein abstrahierendes Modell der Daten (z. B. lineare Regression).
E Batch Learning	5 Speichert Beispiele und vergleicht neue Eingaben mit ihnen (z. B. k-NN).
F Online Learning	6 Findet Strukturen in Daten *ohne* Zielvariable (z. B. Clustering).
G Instanzbasiertes Lernen	7 Kombiniert wenige Beobachtungen mit Zielvariable und viele ohne.
H Modellbasiertes Lernen	8 Aktualisiert das Modell laufend mit neu eintreffenden Daten.

A	B	C	D	E	F	G	H
3	6	7	1	2	8	5	4

A $\to$ *3 überwacht · B* $\to$ *6 unüberwacht · C* $\to$ *7 halbüberwacht · D* $\to$ *1 Reinforcement · E* $\to$ *2 Batch · F* $\to$ *8 Online · G* $\to$ *5 instanzbasiert · H* $\to$ *4 modellbasiert.*

Aufgabe 3: Single Choice

Kreuzen Sie pro Frage die **eine** richtige Antwort an.

1. Was ist die Hauptcharakteristik des überwachten Lernens?

   ☐ Es benötigt keine Zielvariable für die Trainingsdaten.

   ■ Es nutzt eine Zielvariable, um ein Modell zu trainieren.

   ☐ Es basiert ausschliesslich auf der Interaktion mit einer Umgebung.

   ☐ Es verwendet Clustering-Methoden zur Datenanalyse.

2. Welche der folgenden Aufgaben wird typischerweise im unüberwachten Lernen durchgeführt?

   ☐ Klassifikation

   ☐ Regression

   ■ Clustering

   ☐ Reinforcement

3. Welche Aussage trifft auf halbüberwachtes Lernen zu?

   ☐ Es benötigt für jede Beobachtung eine Zielvariable.

   ■ Es verwendet eine Kombination aus Beobachtungen mit und ohne Zielvariable.

   ☐ Es basiert ausschliesslich auf unstrukturierten Daten.

   ☐ Es verwendet keine Zielvariablen.

4. Was ist das Ziel des Reinforcement Learnings?

   ☐ Das Finden verborgener Strukturen in Daten.

   ☐ Das Minimieren von Underfitting durch Datenaugmentation.

   ■ Das Maximieren einer kumulativen Belohnung durch Interaktion mit der Umgebung.

   ☐ Das Erstellen eines Modells auf Basis einer Zielvariable.

5. Wann ist Online Learning besonders nützlich?

   ■ Bei sehr grossen Datensätzen, die nicht vollständig auf einmal verarbeitet werden können.

   ☐ Wenn alle Beobachtungen eine Zielvariable enthalten.

   ☐ Wenn das Modell nicht kontinuierlich aktualisiert werden soll.

   ☐ Wenn ein statisches Modell bevorzugt wird.

6. Wie verhalten sich Künstliche Intelligenz (KI), Machine Learning (ML) und Deep Learning (DL) zueinander?

   ■ DL ist ein Teilgebiet von ML, das wiederum ein Teilgebiet der KI ist.

   ☐ KI ist ein Teilgebiet von ML.

   ☐ ML und DL bezeichnen dasselbe.

   ☐ KI, ML und DL sind voneinander unabhängige Felder.

7. Welche der folgenden Aufgaben ist eine *Regression*?

   ■ Vorhersage des Hauspreises in CHF.

   ☐ Einteilung von E-Mails in „Spam" / „kein Spam".

   ☐ Gruppierung von Kund:innen in Segmente.

   ☐ Erkennen der Ziffer auf einem Bild.

8. Wozu dient die Aufteilung der Daten in Trainings- und Testdaten?

   ■ Um den Generalisierungsfehler auf ungesehenen Daten zu schätzen.

   ☐ Um das Modell auf denselben Daten zu bewerten, auf denen es trainiert wurde.

   ☐ Um die Trainingszeit künstlich zu verlängern.

   ☐ Um die Zielvariable aus den Daten zu entfernen.

9. Was besagt das No-Free-Lunch-Theorem (Wolpert, 1996)?

   ■ Ohne Annahmen über die Daten ist kein Modell grundsätzlich besser als ein anderes.

   ☐ Komplexe Modelle sind immer besser als einfache.

   ☐ Mehr Daten führen immer zu Overfitting.

   ☐ Deep Learning schlägt jedes andere Verfahren.

10. Was ist ein typisches Merkmal von Deep Learning gegenüber klassischem ML?

    ■ Es lernt Merkmale (Features) automatisch über mehrere Schichten.

    ☐ Es benötigt grundsätzlich keine Daten.

    ☐ Es kommt ohne jegliche Parameter aus.

    ☐ Es kann ausschliesslich Regressionsaufgaben lösen.

Aufgabe 4: Offene Fragen

1. Was versteht man unter überwachtem Lernen, und wie wird es angewendet?

Musterlösung. Ein Ansatz, bei dem ein Modell mit Daten trainiert wird, die sowohl Eingaben als auch die zugehörige Zielvariable (Output) enthalten. Ziel ist eine Funktion, die Eingaben korrekt auf Ausgaben abbildet. Typische Aufgaben: Klassifikation (z. B. Spam-Filter) und Regression (z. B. Immobilienpreise).

2. Wie unterscheidet sich unüberwachtes Lernen vom überwachten Lernen, und wofür wird es verwendet?

Musterlösung. Unüberwachtes Lernen arbeitet mit Daten *ohne* Zielvariable und sucht Muster/Strukturen (z. B. Clustering zur Kundensegmentierung, Dimensionsreduktion). Überwachtes Lernen braucht eine Zielvariable und ist auf konkrete Vorhersagen ausgerichtet.

3. Was ist halbüberwachtes Lernen, und warum ist es nützlich?

Musterlösung. Kombiniert wenige Beobachtungen *mit* und viele *ohne* Zielvariable. Nützlich, weil das Beschriften von Daten teuer/zeitaufwendig ist: Die beschrifteten Daten geben eine Grundstruktur vor, die unbeschrifteten verbessern das Modell (z. B. Foto-Tagging).

4. Was sind die Hauptziele des Reinforcement Learnings, und wie unterscheidet es sich von anderen Ansätzen?

Musterlösung. Ein Agent lernt durch *Interaktion* mit einer Umgebung, Aktionen zu wählen, die eine kumulative Belohnung maximieren. Anders als beim überwachten Lernen gibt es keine Zielvariable je Eingabe, sondern Feedback über Belohnung/ Bestrafung (z. B. Navigation eines Roboters, Schach/Go).

5. Was versteht man unter Batch Learning, und wann wird es eingesetzt? Worin unterscheidet es sich von Online Learning?

Musterlösung. *Batch*: Training auf dem vollständigen Datensatz; das Modell bleibt danach unverändert, bis es neu trainiert wird — gut bei statischen Daten. *Online*: Das Modell wird laufend mit neu eintreffenden Daten aktualisiert — gut bei dynamischen/sehr grossen Datenströmen (z. B. Empfehlungssysteme).

6. Was ist instanzbasiertes Lernen, und wie unterscheidet es sich vom modellbasierten Lernen?

Musterlösung. *Instanzbasiert*: speichert Trainingsbeispiele und entscheidet über Ähnlichkeit zu neuen Eingaben (z. B. k-NN). *Modellbasiert*: lernt ein abstrahierendes Modell der Beziehung Eingabe → Ausgabe (z. B. lineare Regression); meist speichereffizienter und schneller in der Vorhersage.

7. Was sind Over- und Underfitting?

Musterlösung. *Underfitting*: Modell zu einfach, schlechte Leistung schon auf den Trainingsdaten — Gegenmittel: komplexeres Modell, mehr/bessere Features, längeres Training. *Overfitting*: Modell zu stark an Trainingsdaten angepasst, generalisiert schlecht.

8. Wie hängen Künstliche Intelligenz, Machine Learning und Deep Learning zusammen? Grenzen Sie die drei Begriffe voneinander ab.

Musterlösung. *Künstliche Intelligenz (KI)* ist der Oberbegriff für Systeme, die Aufgaben lösen, die Intelligenz zu erfordern scheinen. *Machine Learning* ist ein Teilgebiet der KI, in dem Systeme Muster *aus Daten* lernen, statt feste Regeln zu programmieren. *Deep Learning* ist wiederum ein Teilgebiet des ML, das tiefe neuronale Netze nutzt und Merkmale automatisch lernt. Es gilt also KI $\supset$ ML $\supset$ Deep Learning.

9. Worin unterscheiden sich Regression und Klassifikation?

Musterlösung. *Regression* sagt einen *kontinuierlichen* Wert voraus (z. B. Hauspreis in CHF) — Methode: lineare Regression. *Klassifikation* sagt eine *Klasse/Kategorie* voraus (z. B. Spam vs. kein Spam)

10. Wozu dient ein Train/Test-Split, und was versteht man unter dem Generalisierungsfehler?

Musterlösung. Die Daten werden in einen *Trainingsdatensatz* (zum Anpassen des Modells) und einen *Testdatensatz* (zur Bewertung auf ungesehenen Daten) geteilt. Die Testdaten werden verwendet um zu messen wie gut man generalisieren kann. Somit wird der Fehler auf dem Testdatensatz auch Generalisierungsfehler genannt. Der *Generalisierungsfehler* ist die Fehlerrate auf neuen, nicht im Training gesehenen Daten; er schätzt, wie gut das Modell in der Praxis generalisiert. Wichtig: nie auf den Trainingsdaten bewerten — das Ergebnis wäre zu optimistisch.