Data Mining & Grundlagen Maschinelles Lernen 1

Wintersemester 2025/26

Übungsprojekt: Klassifizierung

1 Zweck und Scope

Im Modul Data Mining & Grundlagen Maschinelles Lernen 1 sollen Sie unter anderem lernen, mathematische Vorhersagemodelle für Klassifizierungs- oder Regressionsprobleme zu entwickeln und zu bewerten. Sie sollen dabei die Kenntnisse und Techniken, die im Laufe der Vorlesung vermittelt wurden, auf ein konkretes Problem anwenden, um ein Modell zu entwickeln, das auf unbekannten Daten möglichst gute Vorhersagen liefert und in einen Anwendungskontext eingebunden ist.

Die einzelnen benötigten Techniken wurden im Laufe des Moduls bereits in der Vorlesung und in kleineren Übungsaufgaben angewendet. Die Lernziele der vorliegenden Aufgabe sind:

- Das Erlernte mit relativ wenigen Vorgaben für ein neues Problem einzusetzen und dabei erlerntes Wissen aus verschiedenen Einheiten zu verknüpfen und in einem Anwendungskontext einzusetzen.
- Sauber strukturierten und kommentierten Code zu erzeugen, der von anderen Personen übernommen und ggf. weiterentwickelt werden könnte.

2 Organisation und einzureichende Dokumente

Die Übung ist durch Projektgruppen bestehend aus **bis zu vier** Studierenden umzusetzen. Die Gruppen müssen in ILIAS registriert sein.

- **Start**: 20.11.2025
- **Abgabe Jupyter Notebook (Detailhinweise s.u.): 10.12.2025, 12 Uhr.** Über ILIAS muss ein lauffähiges Jupyter Notebook bis zur Deadline eingereicht werden (ggf. mit Hinweisen auf benötigte Pakete/Versionen).
- **Präsentation zur Aufgabe (Detailhinweise s.u.): 10.12.2025, 12 Uhr.** Foliensatz ebenfalls über ILIAS einzureichen.
- **Präsentation der Ergebnisse (Detailhinweise s.u.): 11.12.2025.** In der Vorlesung bzw. Übung.

3 Datensatz

Ein erheblicher Teil des Gesamtenergieverbrauchs weltweit entfällt heutzutage auf Gebäude, insbesondere auf Heizung, Lüftung und Klimatechnik. Hier bieten sich durch intelligente Steuerung große Einsparpotenziale. Ein Baustein dafür sind Sensoren, die erkennen, ob sich Menschen in bestimmten Teilen des Gebäudes aufhalten und diese Information an die Gebäudesteuerung weiterleiten, sodass Energieverbraucher dynamisch zu- und abgeschaltet werden können. Der vorliegende Datensatz (*Datensatz-Sensor.csv*) enthält Messdaten, die von einem Sensor in einem Bürogebäude aufgezeichnet wurden. Eine Zeile beschreibt die Erfassung verschiedener Messwerte zu einem bestimmten Zeitpunkt an einem festen Ort (d.h., der Sensor wurde nicht bewegt). Leider kann der Sensor aktuell noch nicht messen, ob sich Personen im Raum befinden. Für den vorliegenden Zeitraum wurde die Anwesenheit einer Person aber manuell kontrolliert und den Messdaten hinzugefügt. Insgesamt existieren folgende Felder in den Daten:

- Datum: Zeitstempel im Format JJJJ-MM-DD hh:mm:ss
- Temperatur: Temperatur in Grad Celsius (°C)

- Luftfeuchtigkeit: relative Luftfeuchtigkeit in %
- CO2: CO2 Gehalt in ppm
- Wassergehalt: Verhältnis von Gewicht des verdunsteten Wassers zum Gewicht der trockenen Luft in kg Wasserdampf / kg Luft (abgeleitet aus Temperatur und Luftfeuchtigkeit)
- Anwesenheit: Ist eine Person anwesend (1) oder nicht (0).

Es gibt einen zweiten verborgenen Datensatz, mit welchem nach Ihrer Einreichung überprüft wird, wie gut Ihre Modelle auf unbekannten Daten funktionieren.

4 Aufgaben

Das Ziel der Übung ist es, ein Klassifikationsmodell auf den Sensordaten zu trainieren, sodass der Sensor erkennen kann, ob sich Personen im Raum aufhalten oder nicht (wenn mathematische Modelle eingesetzt werden, um neue, virtuelle "Messwerte" aus physikalischen Messungen zu generieren, spricht man auch von einem Softsensor). Hierzu wird eine möglichst gute Performance angestrebt. Es soll der gesamte ML-Workflow durchlaufen werden, von der Datenvorverarbeitung über das Modelltraining verschiedener Modelle bis zur Auswertung der Modellgüte. Bearbeiten Sie dazu folgende Aufgaben:

1. Laden Sie die Daten und verschaffen Sie sich einen Überblick (Wertebereich einzelner Features, mögliche Korrelationen, Ausreißer, fehlende Werte).
2. Führen Sie geeignete Vorverarbeitungsschritte durch, z.B. Behandlung von Ausreißern und fehlenden Werten, Skalierung der Features, evtl. Generierung neuer Features. Denken Sie daran, dass die Vorverarbeitungsschritte auch für den verborgenen Datensatz automatisiert durchgeführt werden und schließlich auch für den Einsatz im Sensor funktionieren müssen.
3. Trainieren Sie drei Klassifikationsmodelle, die in der Vorlesung behandelt wurden.
4. Optimieren Sie die Hyperparameter der Modelle mittels Suche und Kreuzvalidierung. Überlegen Sie, welche Hyperparameter des entsprechenden Modells mit Hilfe der Vorlesungsfolien und der Dokumentation der Methoden in scikit-learn sind.
5. Erstellen Sie einen ROC-Plot, in dem Sie die verschiedenen Modelle vergleichen.
6. Nehmen Sie vereinfachend an, dass der Sensor Heizung, Lüftung und Klimaanlage aktiviert, sobald Anwesenheit von Personen festgestellt wird. Dadurch entstehen Kosten von 5€ pro Stunde. Umgekehrt ist es für Angestellte nicht akzeptabel, wenn die Heizung bzw. Klimaanlage während ihrer Anwesenheit mehr als eine Stunde pro Tag nicht funktioniert, weil das Modell keine Anwesenheit erkennt. Der Einfachheit halber können Sie annehmen, dass sich die Angabe "eine Stunde pro Tag" nur im Durchschnitt über den Betrachtungszeitraum gelten muss, d.h. an einzelnen Tagen darf der Wert höher sein, wenn er an anderen Tagen niedriger ist. Insgesamt steht die Zufriedenheit der Mitarbeitenden an erster Stelle. Küren Sie als Bestes Ihrer Modelle dasjenige, das voraussichtlich die geringsten Kosten verursacht. Wie sollten Sie die Modelle unter den gegebenen Bedingungen verbessern? Auf welche Metriken achten Sie?

5 Anforderungen

5.1 Jupyter Notebook

Das Notebook sollte sauber strukturiert und lauffähig sein (ggf. Hinweise auf verwendete Pakete und Versionsnummern). Pakete sollten entsprechend importiert werden. Benutzen Sie sowohl Kommentare im Code als auch Markdown-Zellen, um Ihr Vorgehen zu erläutern. Achten Sie auch auf sinnvolle Bezeichner für die Variablen. Grundsätzlich gehe ich davon aus, dass Sie die in der Vorlesung behandelten Pakete wie **scikit-learn**, **Pandas** und **NumPy** für die Analysen benutzen. Falls Sie gänzlich andere Pakete bevorzugen, sprechen Sie dies bitte mit mir ab. Grundsätzlich gilt: Benutzen Sie nur Methoden, die Sie auch erklären können.

5.2 Vorstellung der Analysen

Jede Gruppe soll ihre Lösung im Rahmen der Vorlesung vorstellen. In der Vorstellung sollen Sie Ihre Analysen und Ihr Vorgehen erläutern. Bezüglich des Materials können Sie sich auf Ihr (wohlorganisiertes und -kommentiertes) Jupyter Notebook und eine Präsentation der wichtigsten Ergebnisse stützen. Ziel der Vorstellung ist, dass die Zuhörer verstehen, welche Analysen bzw. Verfahren Sie eingesetzt und wie Sie diese im Code umgesetzt haben. Ich gehe davon aus, dass jedes Teammitglied gleichermaßen mit dem abgegebenen Code vertraut ist und werde ggf. Verständnis durch Rückfragen überprüfen. Als Richtwert für die Vorstellung plane ich ungefähr 15 Minuten pro Gruppe ein (mit 5 Minuten für Rückfragen).