UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR ROBOTIK
UND KOGNITIVE SYSTEME

Künstliche Intelligenz
Sommersemester 2022
23. Mai 2022

Niclas Bockelmann
bockelmann@rob.uni-luebeck.de

Übungsblatt 6

Abgabe bis zum 31. Mai 2022, 10 Uhr, im Moodle.

Aufgabe 1: To Pass or not to Pass: That is the REAL question! (10 Punkte)

Nachdem Sie nun wissen, wie wahrscheinlich es ist, dass Sie Ihr KI-Übungsblatt bestehen, möchten Sie jetzt mithilfe des *ID3*-Algorithmus herausfinden, welche Attribute Ihr Bestehen am meisten beeinflussen.

Es sind wieder folgende Attribute vorhanden:

- Ihr **Wissen** zu dem Stoff auf dem Übungsblatt: gut, mittel, schlecht
- Ihre **Lust**, das Übungsblatt zu bearbeiten: keine, geht so
- Die **Deadline** zur Abgabe: < 2 Tage, ≥ 2 Tage
- **Rick and Morty**: "Mist, leider nur eine Wiederholung", "Yeah, endlich Staffel 6!"

Nehmen wir an, dass die Auswertung der ersten 11 Übungsblätter folgendes ergab:

Übungsblatt	**Wissen**	**Lust**	**Deadline**	**Rick and Morty**	Bestehen
1	gut	geht	≥ 2	Wiederholung	ja
2	gut	geht	≥ 2	Wiederholung	ja
3	mittel	keine	< 2	neue Folge	ja
4	schlecht	geht	< 2	neue Folge	nein
5	gut	geht	< 2	Wiederholung	ja
6	mittel	keine	≥ 2	neue Folge	nein
7	schlecht	geht	≥ 2	neue Folge	ja
8	schlecht	keine	≥ 2	neue Folge	ja
9	gut	geht	< 2	Wiederholung	ja
10	schlecht	geht	≥ 2	Wiederholung	ja
11	gut	keine	< 2	neue Folge	nein

Stellen Sie mithilfe des *ID3*-Algorithmus den Entscheidungsbaum zu den gegebenen Daten grafisch dar. Geben Sie für jeden Schritt die Menge S, $H(S)$, die einzelnen $|S_i|$ und $H(S_i)$ sowie für jedes Attribut A auch $G(S, A)$ an! (10 Punkte)

Aufgabe 2: Zu Vino sag ich nie no (10 Punkte)

Sie sind in einem Wein-Tasting-Wettbewerb und müssen drei verschiedene Weine erkennen und unterscheiden. Zum Glück ist es Ihnen erlaubt eine chemische Analyse aller Weine vorzunehmen. Daraus ergeben sich 13 verschiedene numerische Features:

- Alcohol
- Malic Acid
- Ash
- Alcalinity of Ash
- Magnesium
- Total Phenols
- Flavanoids
- Nonflavanoid Phenols
- Proanthocyanins
- Colour Intensity
- Hue
- OD280/OD315 of diluted wines
- Proline

Zur Klassifikation ziehen Sie das Random Forest Klassifikationsverfahren (RF) heran.

Laden Sie sich das juypter notebook aus dem Moodle herunter. Für diese Aufgabe benötigen Sie die Packages *scikit-learn*, *pandas*, *matplotlib* und *seaborn*.

a) Zunächst wollen Sie die Beziehungen der verschiedenen Features untereinander darstellen. Nutzen Sie dafür die *seaborn* Funktion `pairplot`. Achten Sie dabei drauf, dass die drei Klassen in dem Plot farblich voneinander zu unterscheiden sind. Speichern Sie die Figure als png-Datei. (2 Punkt)

b) Für das Training ihre RF nutzen Sie `RandomForestClassifier` von *scikit-learn*. Führen Sie eine 5-Fold Cross-Validation auf den Daten durch. Nutzen Sie hierfür die von *scikit-learn* gestellten Funktionen wie beispielsweise `cross_validate`. Trainieren Sie ihre RF-Modelle mit einer verschiedenen Anzahl von Bäumen von 1 bis 100 und tracken Sie jeweils die folgenden Metriken auf den Testdaten: F1, Accuracy, Recall und Precision. Nutzen Sie gerne den `scoring`-Parameter der `cross_validate`-Funktion. Beachten Sie, dass es sich um Multiklassen-Problem handelt. (4 Punkte)

c) Was sind die besten Ergebnisse die Sie erreichen können? Plotten Sie die vier Metriken (als Mittelwert mit Standardabweichung) über die Anzahl der Bäume in einer Figure mit der *seaborn* Funktion `relplot`. Eine beispielhafte Darstellung für die Accuracy-Metrik ist in Abbildung 1 zu sehen. Speichern Sie die Figure als png-Datei ab. Diskutieren Sie Ihre Ergebnisse im Code. (2 Punkte)

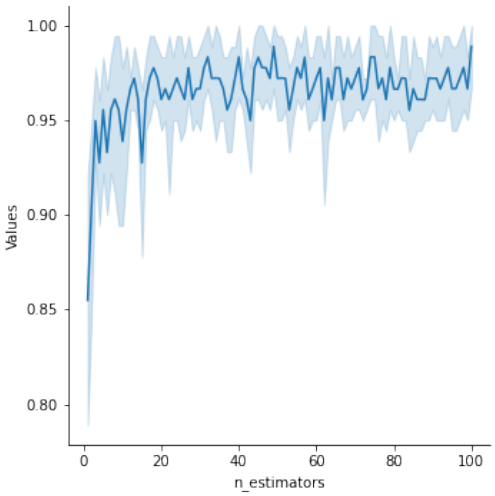

Abbildung 1 Accuracy als Mittelwert und Standardabweichung über Anzahl an Bäumen

d) Welches ist nun das wichtigste Feature für unseren RF? Dies können wir mit der Permutation der Feature überprüfen. Dabei wird nacheinander jedes Feature zufällig geshuffelt und neu trainiert. Die Veränderung der Ergebnisse zeigt dadurch wie wichtig das jeweilige Feature war. Hierfür können Sie die Funktion `permutation_importance` von *scikit-learn* verwenden. Führen Sie ein Training eines RF mit 100 Bäumen auf 80% der Trainingsdaten durch. Auf den 20% Testdaten sollen Sie die Permutation der Feature durchführen. Wiederholen Sie diese für jedes Feature 10 mal. Plotten Sie sich zum Schluss die Wichtigkeit der Feature als Barplot mit Mittelwert und Standardabweichung. Die Wahl der Plot-Funktion bleibt Ihnen selbst überlassen. Speichern Sie die Figure als png-Datei ab. (2 Punkte)

Geben Sie Ihre Implementierung und alle Figures als zip-Datei im Moodle ab.