R-Tutorium
Übung 8

Wintersemester 2020/21
Christoph Herrmann

Aufgabe 1:
Laden Sie den Datensatz *biathlon4* aus der ILIAS Übungseinheit und alle weiteren Pakete, die Sie für die Bearbeitung der Aufgaben benötigen. Ziel der Aufgabe ist es, den gegebenen Datensatz deskriptiv zu analysieren und anschließend die Renntypen auf Basis der restlichen Variablen zu klassifizieren. Der folgende Datensatz beschreibt eine Zusammenstellung von Ergebnissen verschiedener Saisons und Rennen, dem Renntyp-abhängigen Streckendistanzen und der Anzahl an abgegeben Schüssen. Die Daten stammen von der Internetseite der Internationalen Biathlon Union (`https://www.biathlonworld.com/`). Biathlon ist ein vielseitiger Sport und sehr beliebt. Er kombiniert zwei Disziplinen (Langlauf und Schießen). Es gibt vier unterschiedliche Renntypen: Sprint, Einzelrennen, Massenstart und Verfolgungsrennen. Beim Sprint und beim Einzelrennen starten die Athleten in Intervallen, wobei die Reihenfolge gelost wird (außer für die besten Athleten, die können sich ihr Startintervall aussuchen). Beim Massenstart hingegen erfolgt, wie der Name schon sagt, der Start gleichzeitig. Das Verfolgungsrennen ist gekennzeichnet durch die Verfolgung des besten Athleten des vorherigen Sprintrennens, somit starten die Athleten in umgekehrter Reihenfolge und mit den Zeitrückständen auf den Bestplatzierten des Sprints. Tabelle 1 zeigt die Variablen des Datensatzes. Dieser besteht aus 3612 Beobachtungen und 21 Variablen.

(a) Betrachten Sie zunächst alle Variablen im Datensatz und analysieren Sie diese hinsichtlich Ihrer Lage und Streuung. Lassen sich irgendwelche Auffälligkeiten feststellen? Wenn ja sollten Sie überlegen, wie Sie diese bereinigen bzw. beseitigen könnten.

(b) Stellen Sie als nächstes die Variablen eindimensional graphisch dar. Nutzen Sie hierfür die unterschiedlichen Funktionen, die Sie in den Tutorials rund um den *tidyverse*-Approach kennengelernt haben. Betrachten Sie die Variablen zusätzlich im Zusammenhang. Gibt es hier Auffälligkeiten?

(c) Überlegen Sie im Anschluss, welche Merkmale für eine Klassifizierung geeignet sind (metrisches Skalenniveau). Erzeugen Sie in einem nächsten Schritt einen Datensatz, der nur die zur Analyse genutzten Variablen enthält und standardisieren Sie die Daten. Behalten Sie den Datensatz mit den nicht standardisierten Variablen bei und führen Sie die nächsten Schritte einmal für die standardisierten Daten und die original Daten durch

(d) Erzeugen Sie einen Trainings- und Validierungsdatensatz und führen Sie eine lineare Diskriminanzanalyse durch. Sagen Sie im Anschluss einmal die Renntypen für den Trainings- und Validierungsdatensatz vorher. Wie bewerten Sie die Ergebnisse?

(e) Erzeugen Sie als nächstes einen Trainings-, Validierungs- und Testdatensatz mit Hilfe der Funktion *h2o.splitFrame*. Vergessen Sie nicht, das System zu initialisieren und einen H2O Datensatz zu erzeugen. Nutzen Sie als nächstes die Funktion *h2o.randomForest* und klassifizieren Sie die Renntypen. Erstellen Sie den Forest auf Basis unterschiedlicher Anzahlen an Bäumen. Sagen Sie im Anschluss einmal die Renntypen für den Testdatensatz vorher. Unterscheiden sich die Ergebnisse? Betrachten Sie als nächstes die Variable Importance. Welche Variablen spielen eine bedeutende Rolle für die Klassifizierung?

(f) Nutzen Sie zum Abschluss die Funktion *h2o.deeplearning* und klassifizieren Sie die Renntypen. Erstellen Sie das neuronale Netz auf Basis unterschiedlicher Anzahlen an Neuronen und Hidden Layers. Sagen Sie im Anschluss einmal die Renntypen für den Testdatensatz vorher. Unterscheiden sich die Ergebnisse? Überlegen Sie, wie Sie die Vorhersage verbessern können.

Merkmal	Beschreibung
gender	beschreibt das Geschlecht des Biathleten und besitzt die Ausprägungen W für Woman und M für Man.
competition	beschreibt den Renntyp und hat die Ausprägungen I für Individual, S für Sprint, M für Mass Start und P für Pursuit.
type	beschreibt den Wettkampftyp und hat die Ausprägungen W für World Cup, C für Championship und O für Olympic Game.
nation	beschreibt die nationale Zugehörigkeit der Athleten und enthält die englischsprachigen Länderkürzel.
total.time	beschreibt die benötigte Zeit in Sekunden bis zum Ziel.
course.lap.1, course.lap.2, course.lap.3, course.lap.4, course.lap.5 und course.total	beschreiben für die jeweiligen Runden die benötigten Zeiten in Sekunden, die alleine für die Langlaufdisziplin benötigt wurden. Schießzeiten und der Weg vom Eingang bis zum Ende der Schießanlage sind nicht beinhaltet.
shoot.times.1, shoot.times.2, shoot.times.3, shoot.times.4 und shoot.times.total	beschreiben für die jeweiligen Schießdisziplinen die benötigten Zeiten in Sekunden. Laufzeiten und der Weg vom Eingang bis zum Ende der Schießanlage sind nicht beinhaltet.
fails.1, fails.2, fails.3, fails.4 und fails.total	beschreiben die Anzahl der Fehlschüsse bei den jeweiligen Schießdisziplinen.
max.climb	beschreibt den Anstieg mit der größten Höhe in m.
tot.climb	beschreibt die kompletten Höhenmeter, die während des Rennens absolviert werden müssen.
height.diff	beschreibt die Höhendifferenz in m zwischen höchstem und niedrigstem Punkt.

Tabelle 1: Variablenbeschreibung des Datensatzes