h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbmn
FACHBEREICH MATHEMATIK
UND NATURWISSENSCHAFTEN

Gruppe 42
Carl Friedrich Gauß (161803)
Hypatia von Alexandria (314159)
Leonhard Euler (271828)

Data Mining 1

Booklet von Gruppe 42

Inhaltsverzeichnis

Regelungen

Für das Booklet in der Lehrveranstaltung *Data Mining 1* gelten folgende Regelungen:

- Mindestens auf der Titelseite sollte der Gruppenname und die Namen aller Personen (incl. Maktrikelnummer) aufgeführt werden
- Einseitige Zusammenfassung (Executive Summary) mit den wichtigsten Ergebnissen am Anfang des Booklets
- Zu Beginn ein Inhaltsverzeichnis
- Ein Literaturverzeichnis nach dem Anhang
- Maximal 12 Seiten Inhalt (Mindestens Schriftgröße 11)
- Als Inhalt wird mindestens erwartet:
  - ⇒ Beschreibung der Vorgehensweise bei den einzelnen Aufgaben (Algorithmus, Parameterwahl, etc.), Bewertung möglicher Alternativen
  - ⇒ Übersichtliche Darstellung der quantitativen Ergebnisse
  - ⇒ Interpretation der Ergebnisse bzgl. Aussagekraft, Anwendbarkeit der Methode, etc.
- Code im Anhang
- Erstellung des Booklets mit LaTeX
- Abgabe als PDF

Zusammenfassung

1 Aufgabe 1: Entscheidungsbäume

1.1 Feature Engineering

1.1.1 Fehlende Werte

Zunächst wurde der zu untersuchende Datensatz auf fehlende Werte untersucht. Im ersten Schritt wurden insgesamt 834 der 34236 Beobachtungen aus dem Datensatz entfernt, in denen die Zielvariable *RainTomorrow* nicht vorhanden ist.
In einem nächsten Schritt werden die Spalten aus dem Datensatz entfernt, in denen mehr als 40% der beinhaltenden Variablen fehlen. Namentlich werden somit die Spalten *Evaporation, Sunshine, Cloud9am sowie Cloud3pm* aus dem Datensatz entfernt. Der Schwellwert von 40% wurde empirisch festgelegt und hat zu den besten Klassifizierungsergebnissen geführt.
Des Weiteren werden Beobachtungen aus dem Datensatz entfernt, von denen mehr als 50% der Variablen fehlen.
Die übrig gebliebenen Datensätze sind noch immer nicht frei von fehlenden Werten. Um diese zu ersetzen werden für kategorische und numerische Variablen verschiedene Strategien verfolgt. Für die *Imputation* von numerischen Werten wird der Median der jeweiligen Variablen verwendet. Für kategorielle Variablen hingegen wird der am öftesten vorkommende Wert verwendet. Wichtig bei der Ermittlung des Medians bzw. des meist vorkommenden Wertes ist, dass dieser ausschließlich auf der Basis der Trainigsdaten ermittelt wird. Die Ermittlung auf Grund des gesamten Datensatzes würde eine *Test-Train-Leakage* darstellen und ist zu vermeiden. Es muss davon ausgegangen werden, dass die Testdaten nicht bekannt sind. Die ermittelten Werte werden dann auf die Trainings und Testdaten angewendet.

1.1.2 Kategorische Werte

Um kategorische Werte behandeln zu können müssen diese in numerische Werte umkodiert werden. Hierbei werden für die binäre Variablen *RainToday und RainTomorrow* alle Werte in Boolean umgewandelt. Alle anderen kategorischen Variablen werden mit Hilfe des *One-Hot-Encoding-Verfahrens* in numerische Werte umgewandelt. Eine einfache Zuteilung eines Zahlenwertes pro auftretender Variablenausprägung hat den Nachteil, dass dadurch eine metrische numerische Variable für ein gegebenenfalls nicht metrisch interpretierbares Merkmal entsteht. Für Entscheidungsbäume wäre dies kein Problem, jedoch sollte das Feature Engineering weitgehend unabhängig von dem verwendeten Klassifizierungsalgorithmus durchgeführt werden.

1.1.3 Merkmalsaufteilung

Das Feld Datum wurde aufgeteilt.

1.1.4 Merkmalserstellung

Eine weit verbreitete Technik des Feature Engineerings ist die zusätzliche Erstellung von Merkmalen. Somit wurde die Variable *MinMaxDif* erstellt. *PressureDiff, HumidityDiff und WindSpeedDiff auch.*

1.1.5 Diskretisierung

Eine Diskretisierung von Variablen hat den Vorteil, ... Das Merkmal Monat wurde auf die verschiedenen Jahreszeiten diskretisiert.

1.1.6 Bereinigung von Ausreissern

1.1.7 Auswahl von Variablen

1.2 Entscheidungsbäume

2 Aufgabe 2: Musterthema$_2$ in Musterprogramm$_2$

2.1 Unterabschnitt 1

2.2 Unterabschnitt 2

Anhang

Quellcode zu Aufgabe 1

Quellcode zu Aufgabe 2

Literatur