Computerübung - k-Nearest-Neighbors-Klassifikation

Daniel Stepanovic

Ausgangssituation

Zur Untersuchung gelangte eine Stichprobe mit 265 Beobachtungen und drei Variablen. Zwei metrische unabhängige Variablen, `k05` (Anzahl der Kinder im Alter von 0 bis 5 in der Familie) und `einkommen` (Jahresbruttoeinkommen ohne das Einkommen der Frau) sollen dazu verwendet werden, die Erwerbstätigkeit (`bam`) einer verheirateten Frau vorherzusagen. Die Variable `bam` ist die **Zielvariable** und gibt an, ob die Frau berufstätig ist (`ja`) oder nicht (`nein`). Es handelt sich dabei um eine **kategoriale Variable ohne Rangordnung** (also **nominalskaliert**).

Ziel ist es, das optimale k für den k-NN-Algorithmus zu bestimmen und die Erwerbstätigkeit für einen neuen Fall vorherzusagen: Eine Frau mit 3 kleinen Kindern und einem Familieneinkommen (ohne ihr eigenes) von 75.300 Euro.

Datenmanagement

Die Daten wurden mit `read.table()` eingelesen:

```
daten = read.table("wi23b095.txt", header = TRUE, sep = "|", na.strings = "NA")
daten_gesamt = nrow(daten)
```

Umwandlung der Zielvariable in einen Faktor

```
daten$bam = factor(daten$bam, levels = c("nein", "ja"))
```

Entfernen unvollständiger Beobachtungen

Die Datei enthielt fehlende Werte in den Variablen bam, k05 und/oder einkommen. Diese unvollständigen Beobachtungen wurden entfernt, um die k-NN-Klassifikation korrekt durchführen zu können.

```
daten = na.omit(daten[, c("bam", "k05", "einkommen")])
daten_bereinigt = nrow(daten)
```

```
cat("Stichprobengröße vor Bereinigung:", daten_gesamt, "\\n")
```

```
Stichprobengröße vor Bereinigung: 265 \n
```

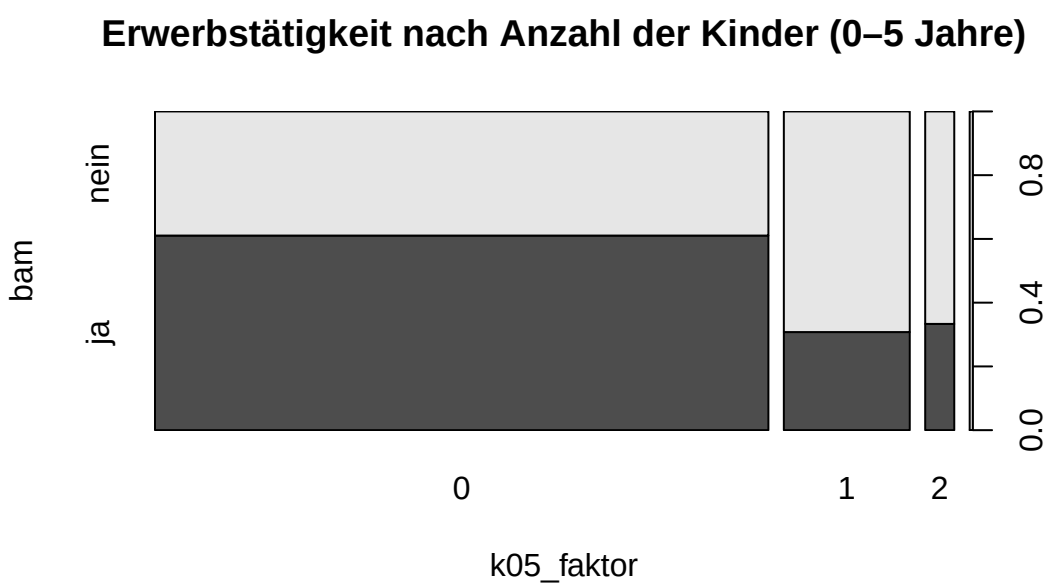
```
cat("Stichprobengröße nach Bereinigung:", daten_bereinigt)
```

```
Stichprobengröße nach Bereinigung: 239
```

Einfluss der Prädiktoren

Einfluss der Kinderanzahl

```
daten$k05_faktor = factor(daten$k05)
spineplot(bam ~ k05_faktor, data = daten,
          main = "Erwerbstätigkeit nach Anzahl der Kinder (0-5 Jahre)")
```



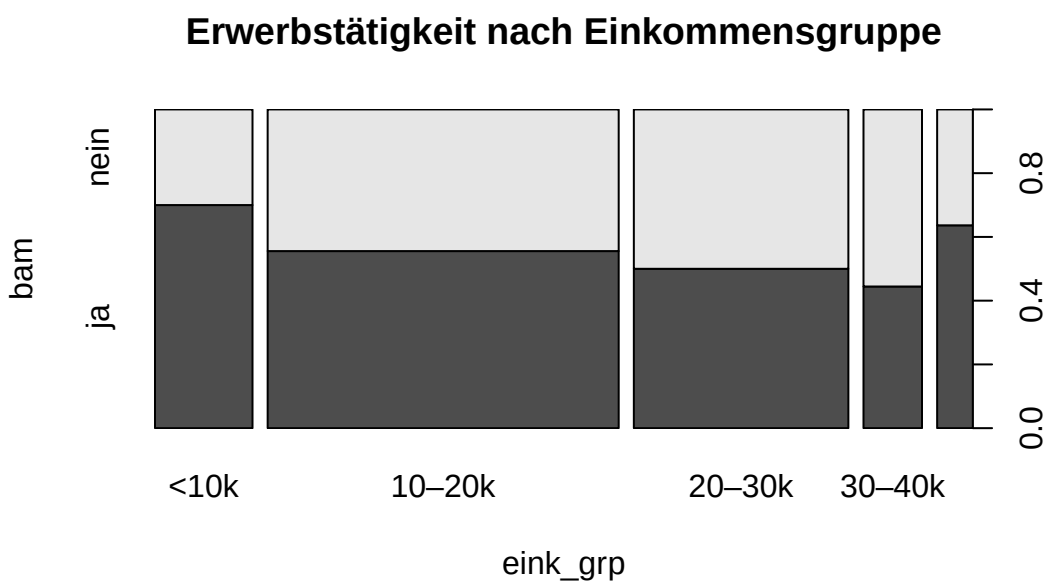
Das Spineplot zeigt deutlich, dass der Anteil berufstätiger Frauen mit zunehmender Anzahl kleiner Kinder abnimmt. Während bei kinderlosen Frauen eine Erwerbstätigkeit häufiger ist, sinkt dieser Anteil bei Frauen mit einem oder zwei kleinen Kindern merklich. Der Anteil erwerbstätiger Frauen ist bei Müttern mit zwei kleinen Kindern besonders gering.

Die Balkenbreite ist in dieser Darstellung gleich, da die Kinderanzahl als kategoriale Variable behandelt wurde. So wird die Verteilung der Erwerbstätigkeit unabhängig von der Gruppengröße übersichtlich vergleichbar dargestellt.

Einfluss des Familieneinkommens

```
# Einkommen in Kategorien einteilen
daten$eink_grp = cut(daten$einkommen,
                     breaks = c(0, 10000, 20000, 30000, 40000, 60000),
                     labels = c("<10k", "10-20k", "20-30k", "30-40k", "40-60k"),
                     include.lowest = TRUE)

# Spineplot mit kategorisiertem Einkommen
spineplot(bam ~ eink_grp, data = daten,
          main = "Erwerbstätigkeit nach Einkommensgruppe")
```

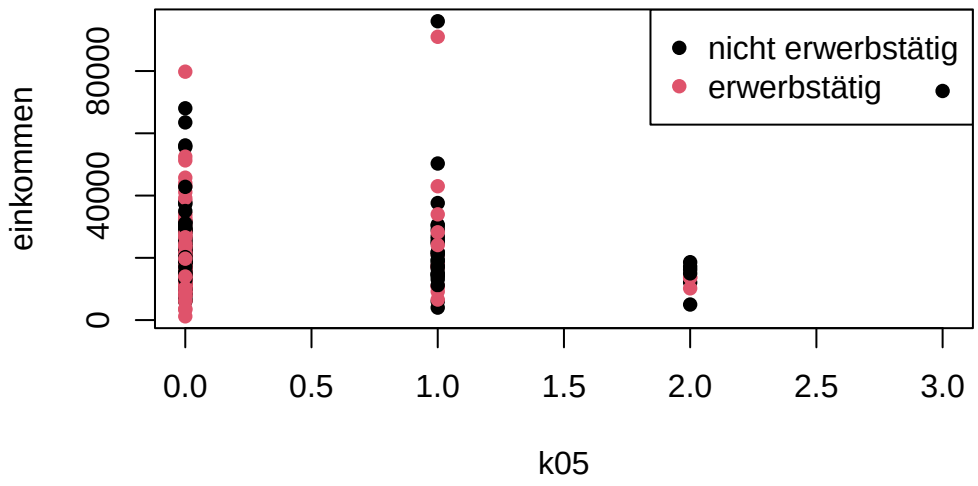


Das Diagramm zeigt den Zusammenhang zwischen dem Familieneinkommen (ohne Einkommen der Frau) und der Erwerbstätigkeit verheirateter Frauen. Dabei wurde das Einkommen in fünf Gruppen unterteilt. Es zeigt sich, dass bei niedrigem Einkommen (<10.000 €) der Anteil erwerbstätiger Frauen deutlich höher ist als bei hohem Einkommen. Mit zunehmendem Einkommen nimmt die Erwerbstätigkeit tendenziell ab. Besonders in der Einkommensgruppe von 30.000 bis 40.000 € ist der Anteil nicht erwerbstätiger Frauen am höchsten. Die Breite der Balken zeigt außerdem, wie viele Fälle in jeder Gruppe vorhanden sind in den höchsten Einkommensgruppen gibt es deutlich weniger Beobachtungen.

Gemeinsamer Einfluss

```
plot(einkommen ~ k05, data = daten, col = bam,
     main = "Erwerbstätigkeit in Abhängigkeit von Kinderanzahl und Einkommen",
     pch = 16)
legend("topright", legend = c("nicht erwerbstätig", "erwerbstätig"), col = 1:2, pch = 16)
```

werbstätigkeit in Abhängigkeit von Kinderanzahl und Einko



Klassifikation mit k-NN

Bestimmung des optimalen k

```
library(e1071)
set.seed(1)

X = scale(daten[, c("k05", "einkommen")])
y = daten$bam

best.k = replicate(20, tune.gknn(y ~ ., data = data.frame(y, X), k = 1:5)$best.parameters$k)
table(best.k)
```

```
best.k
 3  4  5
 2  5 13
```

Modellgüte

```
library(caret)
```

```
Lade nötiges Paket: ggplot2
```

```
Lade nötiges Paket: lattice
```

```
set.seed(4)

n = nrow(daten)
testind = sort(sample(n, size = floor(n / 3)))

knn.mod = gknn(bam ~ k05 + einkommen, data = daten[-testind, ], k = 1)
knn.pred = predict(knn.mod, daten[testind, ])
confusionMatrix(knn.pred, ref = daten$bam[testind], mode = "prec_recall")
```

```
Confusion Matrix and Statistics

          Reference
Prediction nein ja
      nein   21 12
      ja     22 24

               Accuracy : 0.5696
                 95% CI : (0.4533, 0.6806)
    No Information Rate : 0.5443
    P-Value [Acc > NIR] : 0.3686

                  Kappa : 0.1516

 Mcnemar's Test P-Value : 0.1227

              Precision : 0.6364
                 Recall : 0.4884
                     F1 : 0.5526
             Prevalence : 0.5443
         Detection Rate : 0.2658
   Detection Prevalence : 0.4177
      Balanced Accuracy : 0.5775
```

```
       'Positive' Class : nein
```

Bewertung für Kategorie “ja”

```
confusionMatrix(knn.pred, ref = daten$bam[testind], mode = "prec_recall", positive = "ja")
```

```
Confusion Matrix and Statistics

          Reference
Prediction nein ja
      nein   21 12
      ja     22 24

               Accuracy : 0.5696
                 95% CI : (0.4533, 0.6806)
    No Information Rate : 0.5443
    P-Value [Acc > NIR] : 0.3686

                  Kappa : 0.1516

 Mcnemar's Test P-Value : 0.1227

              Precision : 0.5217
                 Recall : 0.6667
                     F1 : 0.5854
             Prevalence : 0.4557
         Detection Rate : 0.3038
   Detection Prevalence : 0.5823
      Balanced Accuracy : 0.5775

       'Positive' Class : ja
```

Klassifikation eines neuen Falls

```r
neuer_fall = data.frame(k05 = 3, einkommen = 75300)
vorhersage = predict(knn.mod, neuer_fall)
cat("Vorhersage für neue Beobachtung:", as.character(vorhersage))
```

```
Vorhersage für neue Beobachtung: ja
```

Fazit

- Die Variable bam ist nominalskaliert, da sie zwei Kategorien ohne natürliche Rangordnung enthält.
- Die ursprüngliche Stichprobe umfasste 265 Beobachtungen, von denen aufgrund fehlender Werte r daten_bereinigt vollständig für die Analyse verwendet werden konnten.
- Die Anzahl kleiner Kinder steht in negativem Zusammenhang mit der Erwerbstätigkeit.
- Auch ein hohes Familieneinkommen (ohne das Einkommen der Frau) ist tendenziell mit geringerer Erwerbstätigkeit verbunden.
- Das k-NN-Modell lieferte eine brauchbare Klassifikation. Der optimale Wert für k wurde mithilfe wiederholter Kreuzvalidierung bestimmt.
- Für den neuen Fall mit 3 kleinen Kindern und einem Familieneinkommen von 75.300 € ergibt sich folgende Vorhersage: r vorhersage