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 = "|")
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 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
```

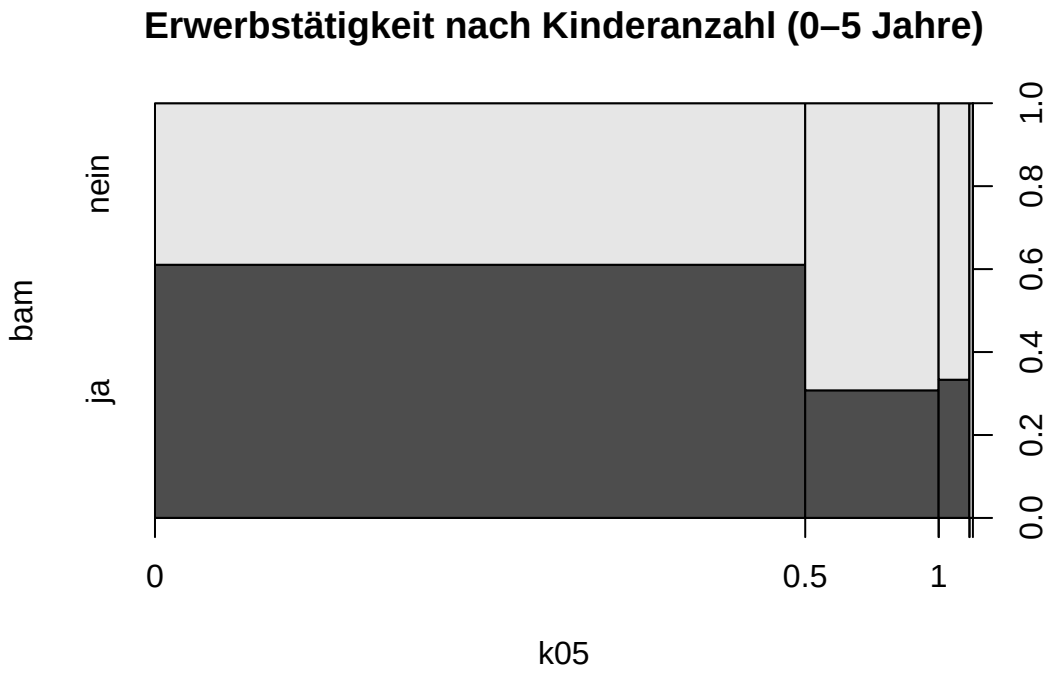
```
cat("Stichprobengröße nach Bereinigung:", daten_bereinigt)
```

```
Stichprobengröße nach Bereinigung: 239
```

Einfluss der Prädiktoren

Einfluss der Kinderanzahl

```
spineplot(bam ~ k05, data = daten,
          main = "Erwerbstätigkeit nach Kinderanzahl (0-5 Jahre)")
```



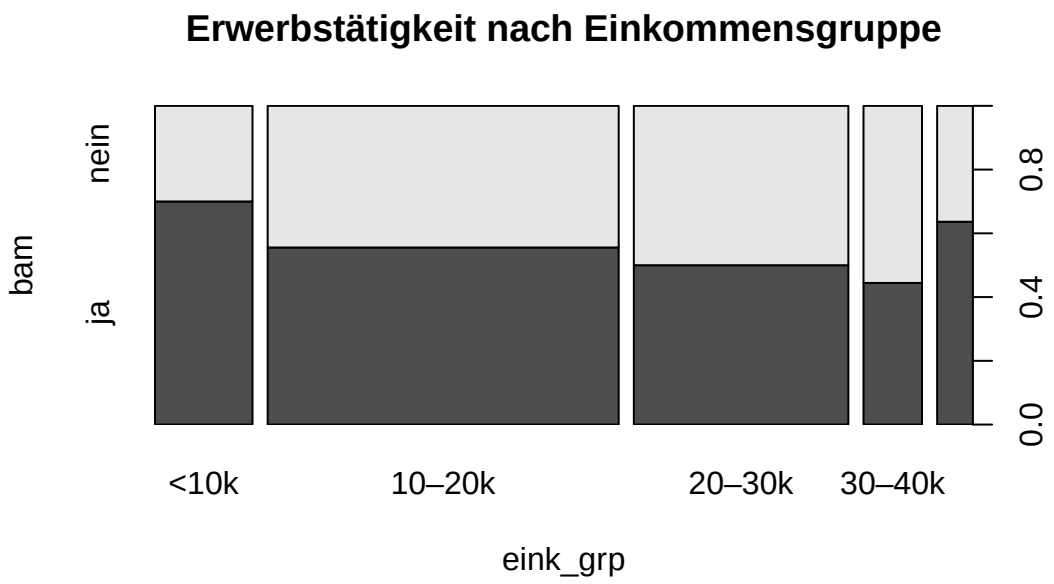
Das Spineplot zeigt deutlich, dass der Anteil berufstätiger Frauen mit zunehmender Anzahl kleiner Kinder abnimmt. Während bei kinderlosen Frauen eine Erwerbstätigkeit häufiger ist, sinkt dieser Anteil bei Frauen mit einem oder zwei kleinen Kindern merklich. Der Anteil erwerbstätiger Frauen ist bei Müttern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

Die Balkenbreite zeigt die Häufigkeit der jeweiligen Kinderanzahl und macht die Gruppengrößen sichtbar.

Einfluss des Familieneinkommens

```
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(bam ~ eink_grp, data = daten,
          main = "Erwerbstätigkeit nach Einkommensgruppe")
```



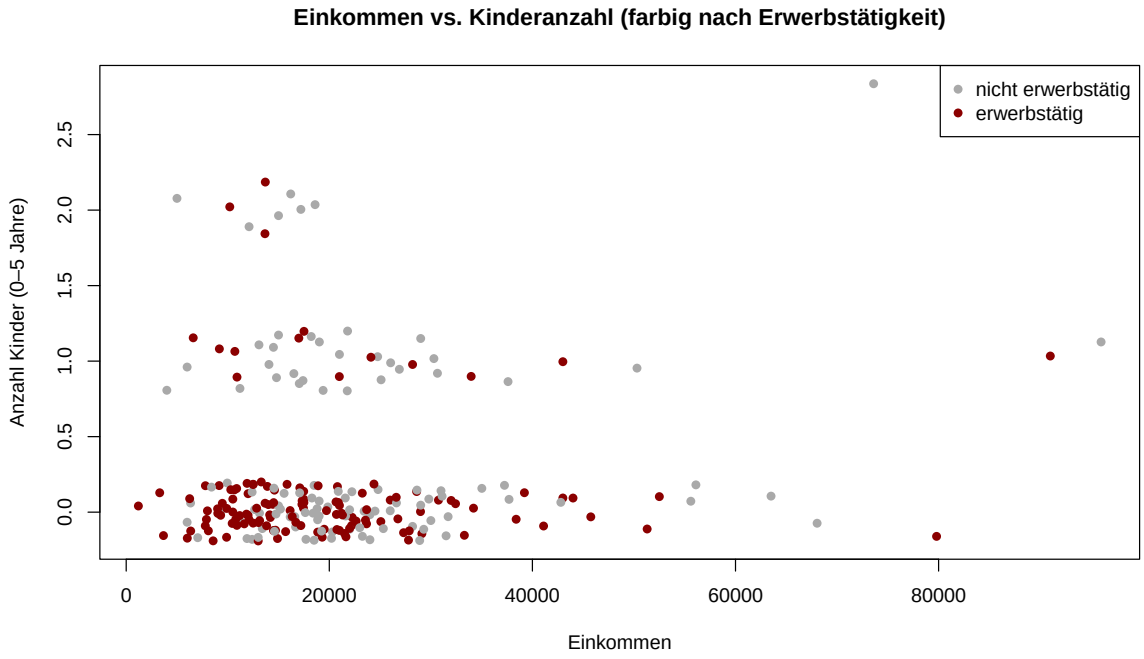
Das Diagramm zeigt den Zusammenhang zwischen dem Familieneinkommen (ohne Einkommen der Frau) und der Erwerbstätigkeit verheirateter Frauen. Dabei wurde das Einkommen in fünf Kategorien gruppiert.

Es ist erkennbar, dass bei niedrigem Familieneinkommen (unter 10.000€) der Anteil erwerbstätiger Frauen deutlich höher ist als bei mittleren oder hohen Einkommen. Mit steigendem Einkommen nimmt die Erwerbstätigkeit tendenziell ab. Besonders im Bereich von etwa 30.000 bis 40.000€ ist der Anteil nicht erwerbstätiger Frauen am höchsten.

Die Breite der Balken spiegelt die Anzahl der Fälle in den jeweiligen Einkommensbereichen wider. In den höheren Einkommensgruppen gibt es deutlich weniger Beobachtungen, was die Interpretation dieser Gruppen einschränkt.

Gemeinsamer Einfluss

```
plot(jitter(k05) ~ einkommen, data = daten, col = ifelse(bam == "ja", "darkred", "darkgray")
     pch = 16, main = "Einkommen vs. Kinderanzahl (farbig nach Erwerbstätigkeit)",
     xlab = "Einkommen", ylab = "Anzahl Kinder (0-5 Jahre)")
legend("topright", legend = c("nicht erwerbstätig", "erwerbstätig"),
       col = c("darkgray", "darkred"), pch = 16)
```



Das Punktdiagramm zeigt den Zusammenhang zwischen Familieneinkommen (ohne Einkommen der Frau) und Anzahl der Kinder (0–5 Jahre), farblich codiert nach Erwerbstätigkeit.

Es lässt sich erkennen, dass Frauen ohne Kinder häufiger berufstätig sind, insbesondere bei niedrigem Familieneinkommen. Mit zunehmender Kinderanzahl verschiebt sich das Verhältnis deutlich zugunsten der Nicht-Erwerbstätigkeit. Ab zwei kleinen Kindern treten erwerbstätige Frauen nur noch vereinzelt auf.

Gleichzeitig ist auch ein leichter Trend sichtbar, dass Erwerbstätigkeit eher bei niedrigeren bis mittleren Einkommen vorkommt. In höheren Einkommensgruppen sind deutlich weniger Fälle vorhanden, was die Interpretierbarkeit einschränkt.

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
```

Zur Bestimmung des optimalen Parameters k für das k-NN-Klassifikationsverfahren wurde das Modell 20 Mal mit Kreuzvalidierung getestet. In jedem Durchlauf wurde das beste k im Bereich von 1 bis 5 bestimmt.

Die Ergebnisse zeigen, dass in 13 von 20 Fällen ein Wert von $k = 5$ die beste Klassifikationsgüte erzielt hat. Der Wert $k = 4$ wurde 5-mal als optimal bestimmt, $k = 3$ nur 2-mal. Die kleineren Werte $k = 1$ und $k = 2$ lieferten in keinem Fall die beste Performance.

Daraus lässt sich schließen, dass ein $k = 5$-Modell in dieser Stichprobe als stabil und geeignet angesehen werden kann und daher für die nachfolgende Klassifikation verwendet wird.

Modellgüte

Um die Qualität des verwendeten 5-NN-Modells genauer einschätzen zu können, wurde die Stichprobe zufällig in einen Trainings- und einen Testdatensatz (jeweils zwei Drittel zu einem Drittel) aufgeteilt. Die Modellgüte wurde anhand der Konfusionsmatrix und abgeleiteter Kennzahlen beurteilt:

```
library(caret)
```

```
Lade nötiges Paket: ggplot2
```

```
Lade nötiges Paket: lattice
```

```r
set.seed(4)
n = nrow(daten)
testind = sort(sample(n, size = floor(n / 3)))
knn.mod = gknn(bam ~ k05 + einkommen, data = daten[-testind, ], k = 5)
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   20 10
      ja     23 26

               Accuracy : 0.5823
                 95% CI : (0.4659, 0.6923)
    No Information Rate : 0.5443
    P-Value [Acc > NIR] : 0.28704

                  Kappa : 0.182

 Mcnemar's Test P-Value : 0.03671

              Precision : 0.6667
                 Recall : 0.4651
                     F1 : 0.5479
             Prevalence : 0.5443
         Detection Rate : 0.2532
   Detection Prevalence : 0.3797
      Balanced Accuracy : 0.5937

       'Positive' Class : nein
```

- Accuracy liegt bei ca. 58%, was nur geringfügig über der No-Information-Rate (NIR = 54%) liegt. Das Modell ist damit nur leicht besser als reines Raten.
- Der p-Wert für den Vergleich Accuracy > NIR beträgt 0,287 – somit ist das Ergebnis nicht signifikant.
- Die Precision (Trefferquote für Klasse „nein“) liegt bei 66,7%, dass heißt, von allen als „nicht erwerbstätig“ klassifizierten Frauen sind etwa zwei Drittel korrekt vorhergesagt worden.
- Der Recall beträgt 46,5%, d.h. nur knapp die Hälfte der tatsächlich nicht erwerbstätigen Frauen wurde vom Modell richtig erkannt.
- Der F1-Score (Kompromiss aus Precision und Recall) liegt bei 54,8% und deutet auf eine mäßige Modellleistung hin.
- Der Kappa-Wert liegt bei nur 0,182, was auf eine geringe Übereinstimmung zwischen Vorhersage und tatsächlicher Klasse hinweist (jenseits des Zufalls).
- Die Balanced Accuracy von 59% zeigt, dass das Modell beide Klassen insgesamt nur schwach unterscheidet.

Vorhersage eines neuen Falls

Nun wird die Erwerbstätigkeit einer verheirateten Frau mit 3 kleinen Kindern und einem Familieneinkommen (ohne ihr Einkommen) von 75.300€ mithilfe des zuvor bestimmten 5-NN-Modells vorhergesagt:

```
# Neuer Fall
neuer_fall = data.frame(k05 = 3, einkommen = 75300)

# Vorhersage mit 5-NN
vorhersage = predict(knn.mod, neuer_fall)
vorhersage
```

```
   1
nein
Levels: nein ja
```

Das 5-NN-Modell sagt voraus, dass die verheiratete Frau mit 3 kleinen Kindern und einem hohen Familieneinkommen von 75.300 € nicht erwerbstätig ist.

Das Ergebnis ist nachvollziehbar: Viele Frauen mit mehreren kleinen Kindern und hohem Einkommen arbeiten laut den Daten ebenfalls nicht. Das Modell erkennt diese Muster und nutzt sie für die Vorhersage.

Fazit

Die k-NN-Analyse zeigt, dass die Erwerbstätigkeit verheirateter Frauen stark mit der Anzahl kleiner Kinder und dem Familieneinkommen zusammenhängt. Frauen mit mehr kleinen Kindern und höherem Einkommen sind deutlich seltener erwerbstätig.

Das Modell konnte mit einfachen Mitteln eine Vorhersage treffen, die zur Datenlage passt. Die Modellgüte ist jedoch nur mittelmäßig, was auf mögliche Einflüsse weiterer Variablen hinweist, die im Datensatz nicht enthalten sind.

Insgesamt eignet sich das k-NN-Verfahren in diesem Fall gut zur explorativen Analyse. Für präzisere Vorhersagen wäre jedoch ein erweitertes Modell mit mehr Einflussfaktoren sinnvoll.