Formelsammlung zu "Vertiefung in statistische Methoden"

Friedrich Leisch

7.1.2016

1 Einleitung

Diese Formelsammlung ist als einzige Unterlage bei der schriftlichen Prüfung aus „Vertiefung in statistische Methoden" zugelassen. Sie kann einseitig auf A4 ausgedruckt werden, auf Vorder- und Rückseite sind beliebige handschriftliche Notizen zugelassen. Stoff der Prüfung sind alle in der VU vorgetragenen Inhalte inklusive Hausübungen, nicht nur der Inhalt dieser Formelsammlung!

2 Zufällige Größen

Zufällige Größen

In der Statistik werden numerische Merkmale durch sogenannte *Zufallsvariablen* (zufällige Größen, ZG) X modelliert:

diskret: Wertebereich ist endliche Menge von Zahlen oder höchstens abzählbar unendlich, z.B. Zahlen $\{1, 2, 3, 4, 5, 6\}$, natürliche Zahlen, ganze Zahlen, ...: $\{x_1, x_2, \ldots\}$

stetig: Wertebereich ist überabzählbar unendlich, z.B. Intervall, positive reelle Zahlen oder alle reellen Zahlen.

Diskrete Zufallsgrößen

- Es gibt höchstens abzählbar unendlich viele Werte für X

$$M_X = \{x_1, x_2, \ldots\}$$

- *Wahrscheinlichkeitsfunktion (W-Fkt)*

$$p_X(x_i) = P(X = x_i)$$

- Die Summe der Wahrscheinlichkeiten aller möglichen Werte muß 1 ergeben:

$$\sum_{x_i \in M_X} p_X(x_i) = 1$$

Achtung: Die Wahrscheinlichkeiten der möglichen Werte sind hier nicht unbedingt alle gleich!

Stetige Zufallsgröße

- X kann (alle) Werte aus einem *Kontinuum* (z.B. Intervall) annehmen.
- Da es überabzählbar viele Punkte gibt, gilt für alle Punkte x aus dem Wertebereich von X

$$P(X = x) = 0$$

  Nur Intervalle haben positive Wahrscheinlichkeiten.

- *Dichtefunktion (DF)*:
  nichtnegative Funktion $f_X : \boldsymbol{R} \to \boldsymbol{R}_+$
- Es gilt:

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(t)\, dt$$

$$P(a < X \le b) = \int_a^b f_X(x)\, dx, \qquad \int_{-\infty}^{\infty} f_X(x)\, dx = 1$$

- Die Funktion $F_X(x) = P(X \le x)$ heißt *Verteilungsfunktion*, sie kann natürlich auch für diskrete Größen definiert werden.

Erwartungswert und Varianz

$$\mathrm{E}(X) = \mu_X = \begin{cases} \sum_{x_i \in M_X} x_i\, p_X(x_i) & \text{falls } X \text{ diskret} \\ \int_{-\infty}^{\infty} x\, f_X(x)\, dx & \text{falls } X \text{ stetig} \end{cases}$$

$$\mathrm{Var}(X) = \mathrm{E}((X - \mu_X)^2) = \begin{cases} \sum_{x_i \in M_X} (x_i - \mu_X)^2\, p_X(x_i) & \text{diskret} \\ \int_{-\infty}^{\infty} (x - \mu_X)^2\, f_X(x)\, dx & \text{stetig} \end{cases}$$

Univariate Normalverteilung

Eine stetige Zufallsvariable X heißt normalverteilt,
in Zeichen $X \sim N(\mu, \sigma^2)$, wenn sie die Dichte

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

besitzt.

$$\begin{aligned} \mathrm{E}(X) &= \mu \\ \mathrm{Var}(X) &= \sigma^2 \end{aligned}$$

Die spezielle Verteilung mit $\mu = 0$ und $\sigma^2 = 1$ heißt Standardnormalverteilung.

Lognormalverteilung

Eine stetige, nicht-negative Zufallsvariable X heißt logarithmisch normalverteilt, in Zeichen $X \sim LN(\mu, \sigma^2)$, falls die transformierte Zufallsvariable $Y = \log(X)$ $N(\mu, \sigma^2)$-verteilt ist. Die Dichte von X ist gegeben durch

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-(\log(x) - \mu)^2/2\sigma^2\right), \qquad x > 0.$$

$$\begin{aligned} \mathrm{E}(X) &= \exp(\mu + \sigma^2/2), \\ \mathrm{Var}(X) &= \exp(2\mu + \sigma^2) \cdot (\exp(\sigma^2) - 1). \end{aligned}$$

Konfidenzintervall für Mittelwert

Gegeben sei eine *beliebig verteilte* Stichprobe mit *bekannter Varianz* σ^2. Dann ist

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

approximativ standardnormalverteilt, die Schranken des entsprechenden *approximativen* Konfidenzintervalls sind

$$\begin{aligned} \theta_u &= \bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \\ \theta_o &= \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \end{aligned}$$

wobei z_α das α-Quantil der $N(0,1)$ ist. Bei unbekannter Varianz wird diese durch die Stichprobenvarianz $\hat{\sigma}^2$ ersetzt, das Konfidenzintervall sollte dann aber erst für größere n verwendet werden.

Konstruktion von Tests

Alle klassischen statistischen Tests basieren auf folgendem Grundprinzip:

1. Es wird eine sogenannte Nullhypothese H_0 und eine Alternative H_1 gebildet.
2. Mittels einer Teststatistik wird berechnet, wie wahrscheinlich die interessierende Eigenschaft der Stichprobe unter der Nullhypothese ist. Bildung eines Konfidenzintervalls für die Teststatistik, Reduktion des Entscheidungsproblems auf den Wert einer einzigen Zahl.
3. Falls die Teststatistik außerhalb des Konfidenzintervalls (=Annahmebereich) liegt, wird die Nullhypothese verworfen (zu unwahrscheinlich gegeben die Stichprobe).
4. Die Art der verwendeten Teststatistik bestimmt, welche Eigenschaft der Stichprobe untersucht wird (Lokation, Streuung, Verteilung, ...).

Fehler 1. Art: Die Nullhypothese stimmt, aber der Test verwirft sie (=Größe oder Signifikanzniveau α des Tests).

Fehler 2. Art: Die Alternative stimmt, aber der Test akzeptiert die Nullhypothese.

	Realität	
Testresultat	Nullhypothese	Alternative
Nullhypothese	1− Größe	Fehler 2. Art
Alternative	Fehler 1. Art	Macht

Optimaler Test: Maximale Macht bei gegebener Größe, gleichzeitige Reduktion beider Fehlerarten nicht mehr möglich.

Die Gewichtung der beiden Fehlerarten hängt meist von der Anwendung ab.

Die wichtigsten Tests

t-Test: Test auf Mittelwert, 1 Stichprobe, 2 gepaarte oder ungepaarte Stichproben. Bei kleinen Stichproben Normalverteilungsannahme.

Rang-basierte Tests: nichtparametrische Tests auf Median, z.B. Wilcoxon. Annahme an Form der Verteilung.

mehr als 2 Stichproben: ANOVA (Normalverteilung), Kruskal-Wallis (Rang-basiert)

Varianzen: F-Test (2 Stichproben), Levene-Test (mehrere Stichproben)

Kreuztabelle: χ^2-Test auf Unabhängigkeit.

Schwankungsbreite bei Ja/Nein-Fragen

Wenn wir n Personen eine Ja/Nein-Frage stellen, ist die Anzahl der Ja's binomialverteilt mit Parametern n und p, wobei p der (in der Regel unbekannte) wahre Prozentsatz von Ja-Sagern ist.

Eine einzelne Person kann mit Ja (1) oder Nein (0) antworten, dies entspricht einer Zufallsvariable X mit

$$\begin{aligned} \mathrm{E}(X) &= p*1+(1-p)*0=p \\ \mathrm{Var}(X) &= \mathrm{E}(X^2)-(\mathrm{E}X)^2= \\ &= p*1^2+(1-p)*0^2-p^2= \\ &= p-p^2=p(1-p) \end{aligned}$$

Der Prozentsatz bei n befragten Personen ergibt sich als das Stichprobenmittel

$$\hat{p}=\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

Nach dem zentralen Grenzwertsatz ist dieser asymptotisch normalverteilt mit Erwartungswert p und Varianz $p(1-p)/n$:

$$z=\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \quad \sim \quad N(0,1)$$

Da 1.96 das 0.975-Quantil der Standardnormalverteilung ist, liegt z mit 95% Wahrscheinlichkeit im Intervall [-1.96,+1.96]. Dies kann benutzt werden um bei 1. gegebener Fallzahl die Schwankungsbreite, oder 2. gewünschter Schwankungsbreite die notwendige Stichprobengröße zu berechnen:

$$\begin{aligned} \Delta p &\approx \frac{1}{\sqrt{n}} \\ n &\approx \frac{1}{(\Delta p)^2} \end{aligned}$$

3 Das klassische lineare Modell

Daten: $(y_i, x_{i1}, \ldots, x_{ik})$, $i=1,\ldots,n$, zu einer metrischen Variablen y und metrischen oder (binär kodierten) kategorialen Regressoren $x_1,\ldots,x_k$.
Modell:

$$y_i=\beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik}+\varepsilon_i\,, \quad i=1,\ldots,n\,.$$

Die Fehler $\varepsilon_1,\ldots,\varepsilon_n$ sind unabhängig und identisch verteilt (u.i.v.) mit

$$\mathrm{E}(\varepsilon_i)=0\,, \quad \mathrm{Var}(\varepsilon_i)=\sigma^2.$$

Die geschätzte lineare Funktion

$$\hat{y}_i=\hat{f}(x_1,\ldots,x_k)=\hat{\beta}_0+\hat{\beta}_1 x_1+\ldots+\hat{\beta}_k x_k$$

kann als Schätzung $\widehat{\mathrm{E}}(y|x_1,\ldots,x_k)$ für den bedingten Erwartungswert von y bei gegebenen Kovariablen $x_1,\ldots,x_k$ angesehen und damit zur Prognose von y verwendet werden.

Parameterschätzungen und Residuen

Zur Schätzung der unbekannten Parameter β_i wird die Kleinste-Quadrate-Methode verwendet:

$$\mathrm{SQR}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}\Big(y_i-(\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\ldots\hat{\beta}_k x_{ik})\Big)^2 \quad \rightarrow \quad \min$$

Residuen: $\hat{\varepsilon}_i=y_i-\hat{y}_i$.

1. Die Residuen sind im Mittel Null:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0 \text{ bzw. } \bar{\hat{\varepsilon}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i = 0.$$

2. Der Mittelwert der geschätzten Werte $\hat{y}_i$ ist gleich dem Mittelwert der beobachteten Werte y_i:

$$\bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i = \bar{y}$$

.

3. Die Regressionshyperebene geht durch den Schwerpunkt der Daten:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \cdots + \hat{\beta}_k \bar{x}_k.$$

Linearität des Einflusses der Kovariablen

Auf den ersten Blick erscheint die lineare Modellierung des Einflusses der Kovariablen sehr restriktiv. Im Rahmen des linearen Modells können jedoch auch nichtlineare Beziehungen modelliert werden. Beispiel:

$$y_i = \beta_0 + \beta_1 \log(z_i) + \varepsilon_i,$$

Einfluss der erklärenden Variable z_i ist logarithmisch. Dies ist mit $x_i = \log(z_i)$ ein lineares Modell: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Im Allgemeinen lassen sich alle nichtlinearen Beziehungen auf ein einfaches lineares Modell zurückführen, solange diese *linear in den Parametern* sind.

Schätzung in R

In R werden lineare Modelle mit Hilfe der Funktion `lm` geschätzt:

```
> meinmodell <- lm(y ~ x1 + x2 + x3, data=meinedaten)
```

Der Teil mit der Tilde ist die sogenannte Modellformel. Auf der linken Seite steht der Name der abhängigen Variablen (Spalte im Datensatz), auf der rechten Seite alle Prädiktoren:

1. Mit Plus getrennte Variablen werden nicht addiert, sondern ins Modell aufgenommen.
2. Bei Termen der Form `x1 * x2` werden sowohl die sogenannten Haupteffekte wie auch eine Interaktion mit aufgenommen (kommt später).
3. Ein Punkt im Modell bezeichnet alle Variablen, die nicht auf der linken Seite stehen. Beispiel: `y ~ .`
4. Ein Minus vor einer Variablen signalisiert, daß diese nicht ins Modell aufgenommen wird. Beispiel: `y ~ . - x1`

Beispiel: Ein Aufruf der Form

```
> meinmodell <- lm(y ~ x1 + x2 * x3, data=meinedaten)
```

entspricht mathematisch dem Modell

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_2 x_3 + \epsilon$$

Elementare mathematische Funktionen können direkt in der Modellformel verwendet werden:

```
> meinmodell <- lm(log(y) ~ x1 + x2 * sin(x3), data=meinedaten)
```

entspricht mathematisch dem Modell

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \sin(x_3) + \beta_4 x_2 \sin(x_3) + \epsilon$$

Streuungszerlegung

Frage: Wie gut paßt die Regressionshyperebene zu den Daten? Maß für die Variabilität der abhängigen Variablen Y ist die Varianz:

$$\hat{s}_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Bei Regression betrachtet man üblicherweise die Quadratsumme

$$\mathrm{SQT} = (n-1)\hat{s}_Y^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

(SQT = „**S**um of s**Q**uares **T**otal“)

$$\mathrm{SQT} = \mathrm{SQE} + \mathrm{SQR}$$

mit

- **S**um of s**Q**uares **T**otal
$$\mathrm{SQT} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$
- **S**um of s**Q**uares **E**xplained
$$\mathrm{SQE} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$
- **S**um of s**Q**uares **R**esidual
$$\mathrm{SQR} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Erklärte Varianz

Bestimmtheitsmaß:

$$R^2 = \frac{\mathrm{SQE}}{\mathrm{SQT}} = 1 - \frac{\mathrm{SQR}}{\mathrm{SQT}} \quad \in [0,1]$$

Es gilt $R^2 = \mathrm{Cor}(y,\hat{y})^2 = r_{y\hat{y}}^2$

$R^2 \approx 0$**:** Varianz der Residuen identisch zur Varianz von Y, X hat keinen linearen (!) Einfluß auf Y

$R^2 \approx 1$**:** Varianz der Residuen fast 0, Daten liegen fast perfekt auf Hyperebene

Hypothesentests

- Die Parameterschätzer sind *Zufallsvariablen*, die von der Stichprobe abhängen.
- Falls die Störgrößen normalverteilt sind, so ist $\hat{\boldsymbol{\beta}}$ multivariat normalverteilt (mit unbekanntem wahrem Parameter als Erwartungswert und aus den Daten schätzbarer Kovarianzmatrix).
- Ohne Normalverteilungsannahme an die Störgrößen gilt die multivariate Normalverteilung approximativ für große Stichproben.

Die Normalverteilung der Parameter kann man für Inferenzstatistik auf dem linearen Modell benutzen:

- t-Test auf Mittelwert
- z-Test auf Mittelwert bei größeren Stichproben
- F-Test zum Vergleich von Varianzen

t-Test

Mit Hilfe des t-Tests kann überprüft werden, ob ein einzelner Koeffizient $\hat{\beta}_j$ sich signifikant von einem vorgegebenem Wert unterscheidet. Der bei weitem wichtigste Fall ist dabei Test auf $\beta_j = 0$, da dann die zugehörige Variable aus dem Modell entfernt werden könnte.

Sei se_j die geschätzte Standardabweichung von $\hat{\beta}_j$, dann gilt

$$t = \frac{\hat{\beta}_j}{se_j} \sim t_{n-p}$$

Die Nullhypothese $\beta_j = 0$ wird abgelehnt, falls

$$|t| > t_{n-p}(1 - \alpha/2)$$

Da die t-Verteilung für $(n-p) \to \infty$ gegen die Standardnormalverteilung konvergiert, ist der Test für große n (und fixes p) identisch zu einem z-Test.

Prognoseintervalle

Will man Konfidenzbereiche für die abhängige Variable y angeben, so sind zwei Fehlerquellen zu berücksichtigen:

- Der unbekannte wahre Vektor der Regressionskoeffizienten wird durch $\hat{\beta}$ approximiert.
- Die beobachtbaren Werte streuen mit Varianz σ^2 um die Regressionsgerade.

Ein *Konfidenzintervall für den Erwartungswert* $\mu_0 = E(y_0)$ einer neuen Beobachtung y_0 an der Stelle $\boldsymbol{x}_0$ zum Niveau $1 - \alpha$ ist gegeben durch

$$\hat{y}_0 \pm t_{n-p}(1 - \alpha/2)\hat{\sigma}\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}.$$

Ein *Prognoseintervall für eine neue Beobachtung* y_0 an der Stelle $\boldsymbol{x}_0$ zum Niveau $1-\alpha$ ist gegeben durch

$$\hat{y}_0 \pm t_{n-p}(1 - \alpha/2)\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}.$$

Kategoriale Kovariablen: Dummy-Codierung

Zur Modellierung des Effekts einer c-kategorialen Kovariable $x \in \{1, \ldots, c\}$ mit Hilfe der Dummy-Kodierung werden die $c - 1$ Dummy-Variablen

$$x_{i1} = \begin{cases} 1 & x_i = 2 \\ 0 & \text{sonst} \end{cases} \quad \ldots \quad x_{i,c-1} = \begin{cases} 1 & x_i = c \\ 0 & \text{sonst} \end{cases}$$

für $i = 1, \ldots, n$ definiert und als erklärende Variablen ins Regressionsmodell aufgenommen:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{i,c-1} x_{i,c-1} + \ldots + \varepsilon_i$$

Aus Gründen der Identifizierbarkeit wird für eine Kategorie von x, hier die erste Kategorie, keine Dummy-Variable ins Modell mit aufgenommen. Diese Kategorie wird als Referenzkategorie bezeichnet. Die Schätzergebnisse werden dann jeweils im Vergleich zu der weggelassenen Kategorie interpretiert.

Vergleich ganzer Modelle

Zwei Modelle werden als *geschachtelt* bezeichnet, wenn alle Variablen des kleineren Modells auch im größeren Modell enthalten sind. Dabei kann der Fehler des kleineren Modells auf den Trainingsdaten nie kleiner als der des größeren Modells sein.

Geschachtelte Modelle können mit Hilfe der Varianzanalyse verglichen werden (F-Test): Ist die Varianz des größeren Modells signifikant kleiner gegeben die zusätzliche Anzahl von geschätzten Parametern?

Die beiden wichtigsten Verfahren:

Anova Typ 1: Geschachtelte Modelle mit Variablen in derselben Reihenfolge wie in der Modellformel.

Anova Typ 2: Vergleich volles Modell mit allen Modellen, aus denen jeweils eine Variable entfernt wurde.

Das korrigierte Bestimmtheitsmaß

Zum Vergleich verschiedener Modelle ist das Bestimmtheitsmaß R^2 nur bedingt geeignet, da es automatisch größer wird, wenn eine neue Kovariable ins Modell aufgenommen wird.

Dieses Problem soll mit dem sogenannten *korrigierten Bestimmtheitsmaß* verhindert werden, indem eine Korrektur für die Anzahl der Parameter angeführt wird, so dass das Gütemaß nicht notwendigerweise größer wird wenn eine zusätzliche Kovariable aufgenommen wird:

$$\bar{R}^2 = 1 - \frac{n-1}{n-p}(1 - R^2).$$

Sehr populär, wird in allen statistischen Programmpaketen standardmäßig ausgegeben, bestraft aber zu wenig für neue Kovariablen.

Informationskriterium AIC

Das im Rahmen der Maximum-Likelihood-Inferenz am häufigsten verwendete Modellwahlkriterium ist das Informationskriterium nach Akaike (AIC). Allgemein ist das AIC definiert durch

$$\text{AIC} = -2 \cdot l(\hat{\beta}, \hat{\sigma}^2) + 2\,|p|$$

wobei $l(\hat{\beta}, \hat{\sigma}^2)$ der maximale Wert der Log-Likelihood ist, und p die Anzahl der Parameter des Modells. Die Herleitung des AIC basiert auf einer Taylorreihenentwicklung des erwarteten Fehlers und ignoriert konstante Terme der Likelihood. Das AIC kann daher nur benutzt werden, um *geschachtelte Modelle derselben Modellfamilie* zu vergleichen, es ist eine relative Größe. Modelle mit kleinerem AIC werden bevorzugt.

Informationskriterium BIC

Das Bayesianisches Informationskriterium BIC ist allgemein gegeben durch

$$\text{BIC} = -2 \cdot l(\hat{\beta}, \hat{\sigma}^2) + \log(n)\,|p|,$$

Es werden wieder Modelle mit kleinerem BIC bevorzugt. AIC und BIC sind auf sehr unterschiedliche Art motiviert, aus praktischer Sicht ist der Hauptunterschied, dass das BIC komplexe Modelle deutlich stärker bestraft als das AIC (wegen $\log(8) = 2.079 > 2$ ab $n = 8$).

4 Generalisierte lineare Modelle

Binäre Regressionsmodelle

Ziel: Modellierung und Schätzung des Effekts der Kovariablen auf die (bedingte) Wahrscheinlichkeit

$$\pi_i = \mathrm{P}(y_i = 1 \,|\, x_{i1}, \ldots, x_{ik}) = \mathrm{E}(y_i \,|\, x_{i1}, \ldots, x_{ik})$$

für das Auftreten von $y_i = 1$, gegeben die Kovariablenwerte $x_{i1}, \ldots, x_{ik}$. Zielvariablen werden dabei als (bedingt) unabhängig angenommen. Lösungsansatz in allen üblichen binären Regressionsmodellen: Verknüpfung der Wahrscheinlichkeit π_i durch eine Beziehung der Form

$$\pi_i = h(\eta_i) = h(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}),$$

mit dem linearen Prädiktor η_i.

- *Responsefunktion* (Antwortfunktion): h ist eine auf der ganzen reellen Achse streng monoton wachsende Funktion mit $h(\eta) \in [0, 1]$, $\forall \eta \in \mathbb{R}$. Insbesonders können daher viele Verteilungsfunktionen als Responsefunktion verwendet werden.

- *Linkfunktion* (Verknüpfungsfunktion): Inverse $g = h^{-1}$ der Responsefunktion, es gilt daher $\eta_i = g(\pi_i)$.

Logit-Modell:

$$\pi = \frac{\exp(\eta)}{1 + \exp(\eta)} \qquad \Longleftrightarrow \qquad \log \frac{\pi}{1 - \pi} = \eta.$$

Probit-Modell:

$$\pi = \Phi(\eta) \qquad \Longleftrightarrow \qquad \Phi^{-1}(\pi) = \eta.$$

Komplementäres log-log-Modell:

$$\pi = 1 - \exp(-\exp(\eta)) \qquad \Longleftrightarrow \qquad \log(-\log(1 - \pi)) = \eta.$$

Interpretation des Logit-Modells

Mit dem linearen Prädiktor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

gilt für die *Chance (odds)*

$$\frac{\pi_i}{1 - \pi_i} = \frac{\mathrm{P}(y_i = 1 \,|\, \boldsymbol{x}_i)}{\mathrm{P}(y_i = 0 \,|\, \boldsymbol{x}_i)}$$

das multiplikative Modell

$$\frac{\mathrm{P}(y_i = 1 \,|\, \boldsymbol{x}_i)}{\mathrm{P}(y_i = 0 \,|\, \boldsymbol{x}_i)} = \exp(\beta_0) \cdot \exp(x_{i1}\beta_1) \cdot \ldots \cdot \exp(x_{ik}\beta_k).$$

Wird z.B. x_{i1} um 1 auf $x_{i1}+1$ erhöht, so gilt für das Verhältnis der Chancen

$$\frac{\mathrm{P}(y_i=1\,|\,x_{i1}+1,\dots)}{\mathrm{P}(y_i=0\,|\,x_{i1}+1,\dots)} \Big/ \frac{\mathrm{P}(y_i=1\,|\,x_{i1},\dots)}{\mathrm{P}(y_i=0\,|\,x_{i1},\dots)} = \exp(\beta_1).$$

$$\begin{aligned}
\beta_1 > 0 \quad &: \quad \text{Chance } \mathrm{P}(y_i=1)/\mathrm{P}(y_i=0) \text{ wird größer,} \\
\beta_1 < 0 \quad &: \quad \text{Chance } \mathrm{P}(y_i=1)/\mathrm{P}(y_i=0) \text{ wird kleiner,} \\
\beta_1 = 0 \quad &: \quad \text{Chance } \mathrm{P}(y_i=1)/\mathrm{P}(y_i=0) \text{ bleibt gleich.}
\end{aligned}$$

Maximum-Likelihood-Schätzung

Unabhängig von der Link-Funktion kann die Likelihood des Modells nicht geschlossen nach $\boldsymbol{\beta}$ aufgelöst werden, es muß numerisch optimiert werden → Fisher Scoring. Es lässt sich zeigen: für $n \to \infty$ existiert der ML-Schätzer asymptotisch und ist sowohl konsistent als auch asymptotisch normalverteilt (Stichprobenumfang $n \to \infty$ genügt).

Signifikanz der Parameter

Mit dem Test

$$H_0 : \beta_j = 0 \quad \text{gegen} \quad H_1 : \beta_j \neq 0, \tag{1}$$

wobei β_j ein Element von $\boldsymbol{\beta}$ ist, testet man wieder, ob diese Variable aus dem Modell entfernt werden kann. Da $\hat{\beta}$ nur asymptotisch normalverteilt ist, wird immer mit der Standardnormalverteilung verglichen → „z-Wert“, „z-Test“.

Linear trennbare Klassen

Eine unangenehme Eigenschaften der binären Regression (egal ob Logit, Probit, ...) ist, daß der einfachste Fall linear trennbarer Gruppen zu „unendlichen“ Koeffizienten führt, der ML-Schätzer liegt bei $\hat{\beta} = \pm\infty$.

Das numerische Maximieren der Likelihood wird in allen (vernünftig implementierten) Paketen nach einer Maximalanzahl von Fisher Scoring Iterationen mit einer Warnung abgebrochen. Die geschätzten Parameter sind dann einfach nur „sehr groß“'.

In diesem Fall liefert z.B. die Fisher'sche Diskriminanzanalyse ein geeignetes Modell und insbesonders die trennende Hyperebene.

Poisson-Regression für Zähldaten

Die Zielvariablen y_i nehmen Werte aus $\{0, 1, 2, \dots\}$ an und sind bei gegebenen Kovariablen $x_{i1}, \dots, x_{ik}$ (bedingt) unabhängig.

Log-lineares Poisson-Modell: $y_i \,|\, \boldsymbol{x}_i \sim Po(\lambda_i)$ mit

$$\lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \quad \text{bzw.} \quad \log \lambda_i = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Modell mit Überdispersion:

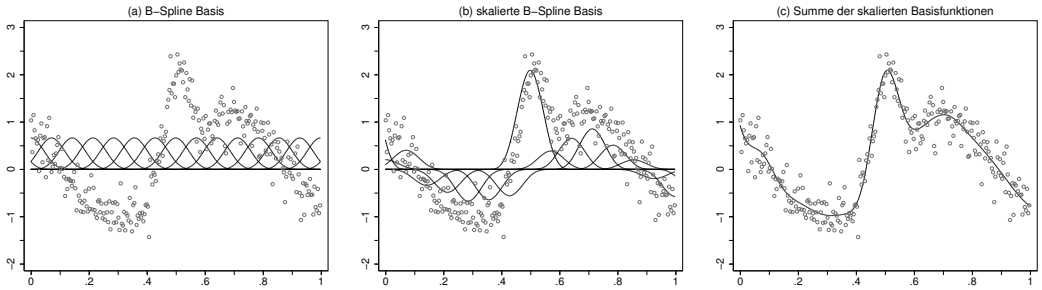
$$\mathrm{E}(y_i \,|\, \boldsymbol{x}_i) = \lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}), \quad \mathrm{Var}(y_i \,|\, \boldsymbol{x}_i) = \phi\lambda_i$$

mit Überdispersions-Parameter ϕ. Die normale Poissonverteilung hat nur einen einzigen Parameter λ, dieser ist zugleich Mittwelwert und Varianz der Verteilung:

$$P\{X = k\} = \frac{\lambda^k e^{-\lambda}}{k!}, \quad EX = \mathrm{Var}(X) = \lambda$$

5 Generalisierte Additive Modelle

Interpolation mit Splines



Splines benutznen regelmäßige einfache nichtlineare Funktionen, die sogenannte Spline-Basis, die in Kombination fast beliebige nichtlinare Verläufe modellieren können:

$$\hat{f}(x) = \sum \alpha_n b_n(x)$$

Z.B. kann die x-Achse in disjunkte Intervalle zerteilt werden und in jedem Intervall eine Polynom niederer Ordnung angepaßt werden.
Kubische Splines:

$$\xi_{i-1} \leq x \leq \xi_i : f(x) = a_i + b_i x + c_i x^2 + d_i x^3$$

Verschiedene Basisfunktionen haben numerisch unterschiedliche Eigenschaften → B-Splines, natürliche Splines, ...

Glättende Splines

Suche f mit $f'' \in L^2$, sodaß für $\lambda \geq 0$

$$\mathrm{RSS}(f, \lambda) = \sum_{n=1}^{N} (y_n - f(x_n))^2 + \lambda \int (f''(t))^2 \, dt$$

minimal wird.

$\lambda = 0$**:** Lösung ist jede Funktion, die die Daten interpoliert

$\lambda \to \infty$**:** Gerade (2.Ableitung 0)

Additive Regression

Beobachtungen: $(y_i, x_{i1}, \ldots, x_{ik}, z_{i1}, \ldots, z_{iq})$, $i = 1, \ldots, n$

- x_{ij}: linearer Einfluß auf y
- z_{ij}: (potentiell) nichtlinearer Einfluß

Additive Modelle erweitern das lineare Modell zu

$$\begin{aligned} y_i &= f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i \\ &= f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \eta_i^{lin} + \varepsilon_i \\ &= \eta_i^{add} + \varepsilon_i \end{aligned}$$

mit

$$\eta_i^{lin} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}, \qquad \eta_i^{add} = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \eta_i^{lin}.$$

Identifikationsproblem: Addiert man etwa zur Funktion $f_1(z_1)$ eine Konstante $c \neq 0$ und subtrahiert c gleichzeitig von einer zweiten Funktion $f_2(z_2)$, so bleibt die Summe

$$f_1(z_1) + f_2(z_2) = f_1(z_1) + c + f_2(z_2) - c$$

unverändert, d.h. durch den Übergang von $f_1(z_1)$ zu $\tilde{f}_1(z_1) = f_1(z_1) + c$ und von $f_2(z_2)$ zu $\tilde{f}_2(z_2) = f_2(z_2) - c$ ändert sich der Wert des Prädiktors nicht. Lösung: zentrieren der f_j

$$\sum_{i=1}^{n} f_1(z_{i1}) = \ldots = \sum_{i=1}^{n} f_q(z_{iq}) = 0$$

Für die Fehlervariablen ε_i werden im Standardmodell der additiven Regression die gleichen Annahmen wie im klassischen linearen Modell getroffen, d.h. die ε_i sind u.i.v. mit $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ und gegebenenfalls normalverteilt. Wie im linearen Modell übertragen sich die Annahmen über die Fehlervariablen ε_i auf die Zielgrößen y_i, d.h. die y_i sind für gegebene Kovariablenwerte (bedingt) unabhängig mit

$$\begin{aligned} \mathrm{E}(y_i) &= \mu_i = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \\ \mathrm{Var}(y_i) &= \sigma^2 \end{aligned}$$

und gegebenenfalls normalverteilt mit

$$y_i \sim \mathrm{N}(\mu_i, \sigma^2).$$

6 Hauptkomponentenanalyse (PCA)

Matrizen

Matrizen können beliebige Dimensionen haben.

$$\begin{pmatrix} 1 & 3 \\ 4 & 2 \\ 2 & 5 \end{pmatrix}, \begin{pmatrix} 1 & 3 & 4 \\ 4 & 2 & 2 \\ 2 & 5 & 9 \end{pmatrix}, \begin{pmatrix} 1 & 3 & 4 \\ 4 & 2 & 2 \end{pmatrix}$$

In der Mathematik gelten bestimmte Rechenregeln für Matrizen.

$$\begin{pmatrix} 1 & 3 \\ 4 & 2 \end{pmatrix} + \begin{pmatrix} 2 & 2 \\ 3 & 1 \end{pmatrix} = \begin{pmatrix} 1+2 & 3+2 \\ 4+3 & 2+1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 3 \\ 4 & 2 \\ 2 & 5 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 1 \times \beta_0 + 3 \times \beta_1 \\ 4 \times \beta_0 + 2 \times \beta_1 \\ 2 \times \beta_0 + 5 \times \beta_1 \end{pmatrix}$$

etc.

Lineares Gleichungssysteme

Daten: $(y_i, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, n$,

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \ldots + \beta_k x_{1k} + \varepsilon_1 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \ldots + \beta_k x_{nk} + \varepsilon_n \end{aligned}$$

In Matrix-Schreibweise sieht das Gleichungssystem wie folgt aus

$$\begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_k \end{pmatrix}$$

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}$$

$\boldsymbol{y}$ ist eine Linearkombination von $\mathbf{X}$.

PCA: Hauptkomponentenanalyse

- engl. Principal Components Analysis - PCA
- Transformation der ursprünglichen Variablen in eine neue Menge unkorrelierter Variablen auf orthogonalen Koordinatenachsen
- Hauptkomponenten sind Linearkombinationen der ursprünglichen Variablen
- Hoffnung, dass wenige Hauptkomponenten für den grössten Teil der Variation in den Originaldaten verantwortlich sind

$X^t = (X_1, ..., X_m)$ ist ein ein m-dimensionaler Zufallsvektor.
Finde Hauptkomponenten $Z_1, \ldots, Z_m$ unkorreliert und mit fallender Varianz, so dass

$$Z_j = a_{1j}X_1 + a_{2j}X_2 + \ldots + a_{mj}X_m = a_j^t X$$

unter der Nebenbedingungen $a_j' a_j = 1$ (um eine willkürliche Skalierung zu vermeiden). Diese Normierung bewirkt, dass die gesamte Transformation der Daten orthogonal wird (Abstände bleiben erhalten). Die wichtigste Bedingung ist jedoch, dass die Varianz der Z_j also $Var(Z_j) = Var(a_j^t X)$ maximal wird. Dadurch erhält man ein Maximierungsproblem für das es eine eindeutige Lösung gibt.

Wie findet man die richtige Anzahl an HK?

- Reduzierung der Hauptkomponenten führt zur Reduzierung der wiedergegebenen Varianz.
- Hauptkomponenten so auswählen, dass Sie die Dimensionen reduzieren, aber trotzdem möglichst viel der Varianz wiedergeben.

7 Clusteranalyse

Gegeben: Multivariater Datensatz mit N Beobachtungen von p Variablen

Aufgabe: Finde Gruppen, die in sich möglichst homogen sind und sich gleichzeitig möglichst stark voneinander unterscheiden.

3 wichtige Gruppen von Verfahren:

- *Hierarchisches Clustern*
- *Partitionierendes Clustern*
- Modellbasiertes Clustern

plus unendlich viele weitere Algorithmen ...

Hierarchisches Clustern

- **Divisive Clusterverfahren**
  Start mit nur einem Cluster
  Schrittweises Aufteilen in immer kleinere Cluster
  Stopp wenn jedes Cluster nur noch aus einem Objekt besteht
- **Agglomerative Clusterverfahren (häufiger)**
  Jedes Objekt in eigenem Cluster
  Schrittweises Zusammenfassen zu immer größeren Clustern
  Stopp wenn alle Objekte zu einem Cluster gehören

Das Aufteilen bzw. Zusammenfassen der Cluster ergibt sich aus den *Distanzen* zwischen den Beobachtungen. Es ensteht ein Baum (Dendrogramm).

Distanzmaße

Der Abstand zwischen 2 Punkten kann auf viele Arten definiert werden.
2 Datenpunkte: $\mathbf{x} = (x_1, \ldots, x_p)'$, $\mathbf{y} = (y_1, \ldots, y_p)'$

* Euklidisch (L^2):

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

* Quadratisch-Euklidisch (L^2):

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{p}(x_i - y_i)^2$$

* Manhattan (L^1, absolut):

$$d(x, y) = \sum_{i=1}^{p}|x_i - y_i|$$

* Maximum:

$$d(x, y) = \max_{i=1,\ldots,p}|x_i - y_i|$$

* Nominal:

$$d(x, y) = \frac{\#\{i : x_i \neq y_i\}}{p}$$

Die am häufigsten verwendeten Distanzmaße sind euklidisch und Manhattan:

Euklidisch: Erzeugt kugelförmige Cluster.

Manhattan: Erzeugt würfelförmige Cluster, robuster gegen Ausreißer.

Distanzen - zwischen Clustern

Distanz d zwischen Datenpunkten a und b vs. Distanz D zwischen Clustern A und B:

Single Linkage:	$D(A, B) = \min\, d(a, b)$
Average Linkage:	$D(A, B) = \text{mean}\, d(a, b)$
Complete Linkage:	$D(A, B) = \max\, d(a, b)$
Centroid:	$D(A, B) = \|\|\bar{a} - \bar{b}\|\|$
Ward:	$D(A, B) = 2\|A\|\|B\| * \|\|\bar{a} - \bar{b}\|\|/(\|A\| + \|B\|)$

mit $\bar{a}$ Mittelwert von Cluster A.

Partitionierende Verfahren

Im Gegensatz zum hierarchischen Clustern werden nicht Lösungen für $k = 1, \ldots, N$ Cluster erzeugt, sondern eine einzige Partition der Daten (disjunkte Zerlegung der Daten in Segmente). Clusterzentren sind meistens die Zentroide.

- Vorgabe einer bestimmten Anzahl von Clustern
- Suche Lösung, wo jeder Punkt möglichst nahe bei „seinem“ Clusterzentrum liegt (= kleiner Radius)
- Bekannteste Methode ist der Klassiker k-Means.

KCCA: k-Centroid Cluster Analysis

Optimierungsproblem: Gegeben Datenpunkte $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, suche für *vorgegebenes* k eine Menge von Zentroiden $C_k = \{\mathbf{c}_1, \ldots, \mathbf{c}_k\}$ so dass der Abstand jedes Punktes x zum nächsten Zentroid möglichst klein ist. Sei $c(x)$ der nächste Zentroid:

$$c(x) = \mathbf{c}_i \quad \Leftrightarrow \quad d(\mathbf{x}, \mathbf{c}_i) \leq d(\mathbf{x}, \mathbf{c}_j)$$

Daraus entsteht das Optimierungsproblem:

$$\sum_{n=1}^{N} d(x_n, c(x_n)) \quad \rightarrow \quad \min_{\{\mathbf{c}_1,\ldots,\mathbf{c}_k\}}$$

Geschlossene Lösung existiert nicht, daher iterative Lösungsverfahren : k-Means, Learning Vector Quantization (LVQ), ...

Verallgemeinertes k-Means

1. Wähle k zufällige Cluster-Zentren als Initialisierung (normalerweise k zufällige Beobachtungen).
2. Ordne jede Beobachtung $\mathbf{x}_n$ dem nächstgelegenen Zentroiden $c(x)$ zu $\rightarrow$ Partition der Daten in k Cluster.
3. Update der Zentroide getrennt für jeden Cluster:

$$\mathbf{c}_k := \text{argmin}_{\mathbf{c}} \sum_{n:c(\mathbf{x}_n)=\mathbf{c}_k} d(\mathbf{x}_n, \mathbf{c}), \qquad k = 1, \ldots, K.$$

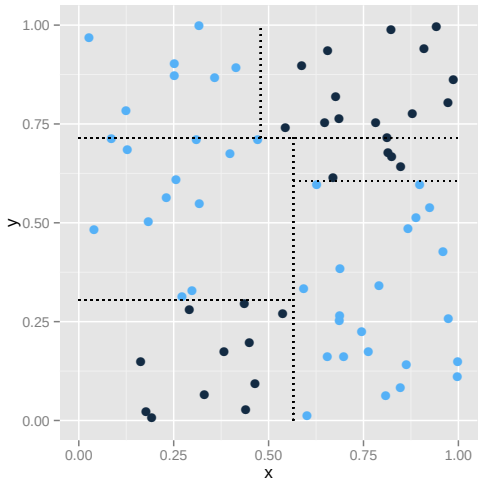
4. Wiederhole Schritte 2 und 3 bis zur Konvergenz (Iteration).

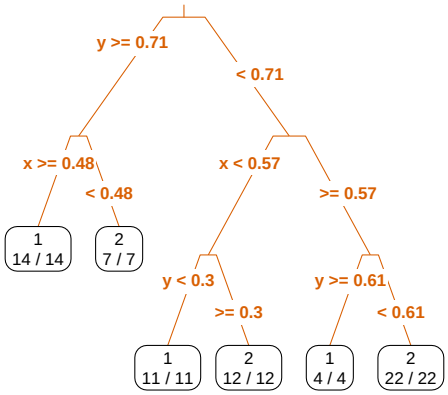
Für Euklidische und Manhatten-Distanz ist Schritt 3 sehr einfach, die Zentroide entsprechen den Mittelwerten bzw. Medianen der Cluster.
$\rightarrow$ k-Means, k-Medians.

8 Baumverfahren

Funktionsweise des Entscheidungsbaumes:

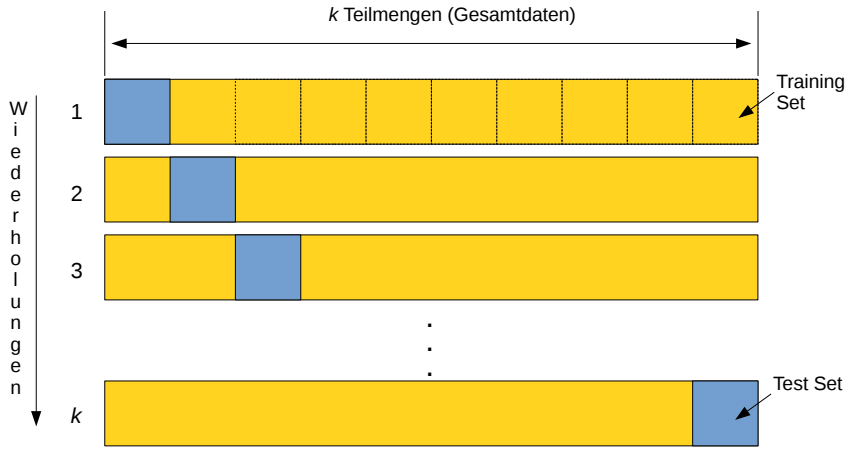
- rekursives Teilen des Variablenraumes in jedem Schritt
- Klassifikation wird mittels Durchlaufen des Baumes vom Wurzel (root) Knoten bis zum Blatt (leaf) erreicht





Kreuzvalidierung

Zur Evaluierung/Optimierung von Decision Trees wird die **Kreuzvalidierung** verwendet, die das Paket `rpart` automatisch durchführt. Hier wird der originale Datensatz zufällig in das so genannte **Training Set** und **Test Set** gespalten. Die prozentuelle Aufteilung richtet sich nach der Anzahl der Wiederholungen der Kreuzvalidierung. Mit dem Training Set wird der Baum erstellt, mit dem Test Set wird die Performance evaluiert bzw. Fehlerrate errechnet. Dies wird eine gewisse Anzahl an Wiederholungen durchgeführt und danach die durchschnittliche Fehlerquote ermittelt.



Würde man den Baum zu genau konstruieren - mit sehr kleinen Endknoten, wo nur wenige Beobachtungen enthalten sind - entsteht ein zu stark auf die vorhandenen Daten (dem Training Set) angepaßter Baum, der auf neuen Daten (dem Test Set) falsche Vorhersagen liefern würde (Overfitting). Deshalb ist es notwendig den Baum nicht überanzupassen (in den vorhandenen Daten könnten z. B. Ausreißer und/oder durch Meßfehler verzerrte Daten sein). Standardmäßig sind daher in `rpart` und `party` Mindestgrößen für Endknoten bzw. für Knoten die gesplittet werden sollen vorgesehen.

Random Forest

Ein Wald besteht ja bekanntlich aus vielen Bäumen → viele Decision Trees bilden daher einen Random Forest. Es werden n Entscheidungsbäume folgendermaßen erstellt:

- Aus dem Datensatz wird mit Zurücklegen ein Datensatz selber Größe gezogen (Bootstrapping)
- Für jeden Knoten im Baum wird ein Subset an Variablen zufällig ausgewählt mit denen der beste Split berechnet wird.
- Der Baum wird vollständig angepaßt (kein Pruning)

Ein neuer Datensatz wird dann durch jeden Baum des Forests geschickt. Die Mehrheit entscheidet über die endgültige Gruppenzuweisung.