Multivariate Verfahren

Trees und Random Forests

Annika Hoyer

Sommersemester 2020

Trees und Random Forests - Inhalt

Classification and Regression Trees

Conditional Inference Trees

Random Forests

Classification and Regression Trees

Ausgangssituation

- Beobachtung von Merkmalsvektoren $\mathbf{x}_1, \ldots, \mathbf{x}_n$ der Objekte $a_1, \ldots, a_n$
- Ziel: von Merkmalsvektoren auf Zielvariable $Y_1, \ldots, Y_n$ schließen

1. Zufallsvariable Y kann stetig sein
   $\rightarrow$ Regressionsproblem
2. Zufallsvariable Y kann binär oder kategorial sein
   $\rightarrow$ Klassifikationsproblem

Im Fall eines Klassifikationsproblems ist die Fragestellung analog zur Diskriminanzanalyse.

Trees - Motivation

- Populärste Methode: Classification and Regression Trees (CART)
- Einführung von Breiman et al. (1984)
- Idee: Teile p-dimensionalen Merkmalsraum durch Splitten sukzessive in Teilmengen des Merkmalsraums auf
  $\rightarrow$ "Rekursives Partitionieren"
- Bilde Teilmengen des Merkmalsraums so, dass sie bezüglich der Zielvariable möglichst homogen sind

CART - Grundprinzip

- Betrachte in jedem Schritt binäre Splits, d.h. in jedem Schritt wird eine (bereits gebildete) Teilmenge weiter in genau zwei Teile aufgeteilt
- Betrachte in jedem Schritt genau eine Variable, die den neuen Split bestimmt
- Resultat: disjunkte Zerlegung des Merkmalsraums, die in einer Baumstruktur dargestellt werden kann

Beispiel: Studienanfänger Wirtschaftswissenschaften

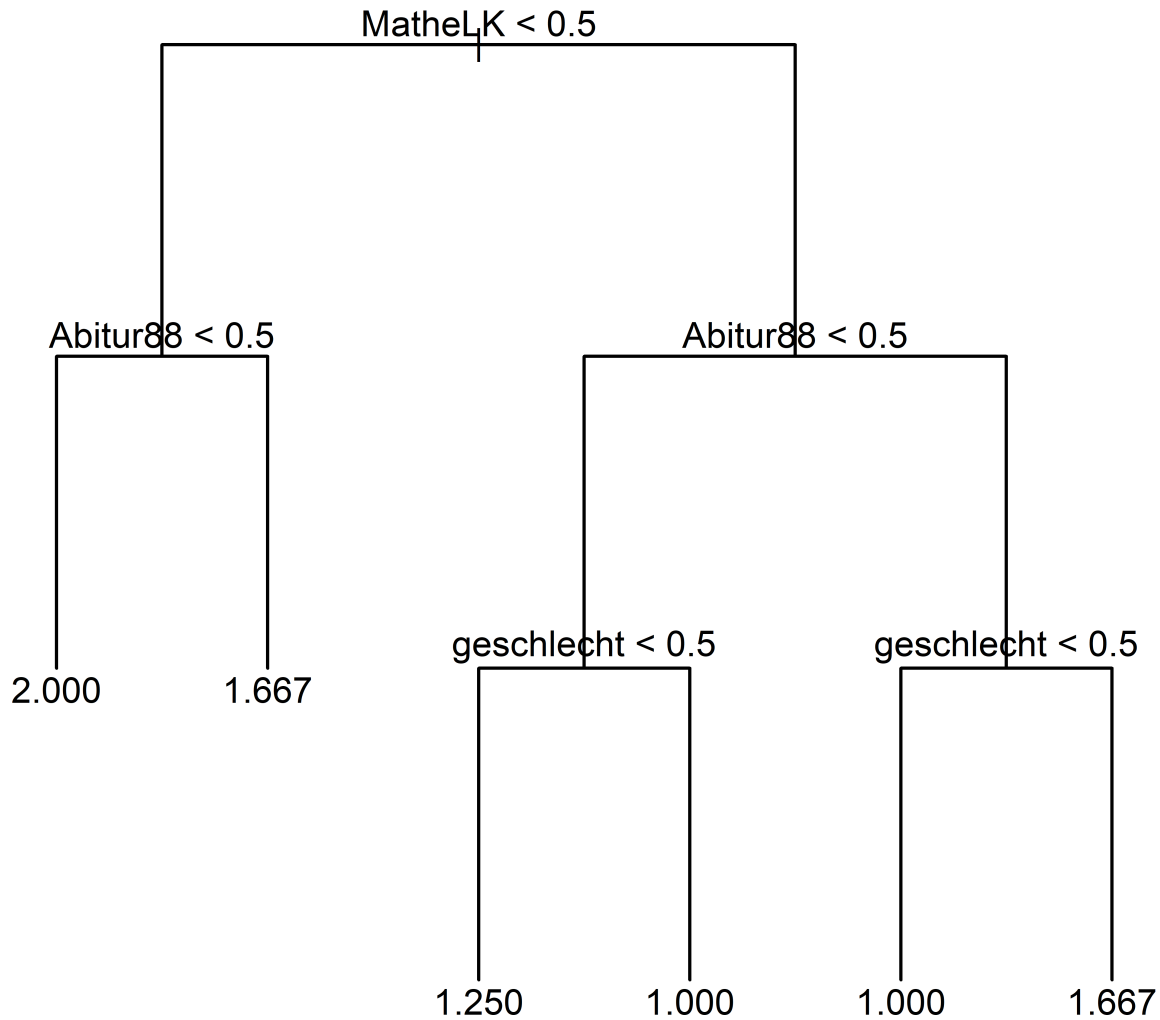
- Test zur Mittelstufenalgebra (26 Fragen) im Wintersemester 1988/89 bei Studienanfängern der Wirtschaftswissenschaften an der FU Berlin
- Variablen:
  - Geschlecht: w/m (1/0)
  - Besuch Leistungskurs Mathe: j/n (1/0)
  - Abitur im Jahr 1988: j/n (1/0)
  - Abinote Mathematik
  - Anzahl der im Test richtig gelösten Aufgaben
  - Gruppe 1: mindestens 14 Punkte erreicht $\rightarrow$ Test bestanden
  - Gruppe 2: weniger als 14 Punkte $\rightarrow$ Test nicht bestanden

Beispiel: Studienanfänger Wirtschaftswissenschaften

Ergebnisse der Studienanfänger bei dem Mathetest

Geschlecht	MatheLK	MatheNote	Abitur88	Gruppe
0	0	3	0	2
0	0	4	0	2
0	0	4	0	2
0	0	4	0	2
1	0	3	0	2
...	...	...	...	...

Beispiel: Studienanfänger Wirtschaftswissenschaften



Trees - Notation

1. Wurzelknoten (root nodes): oberster Knoten
2. Entscheidungsknoten (children nodes): Beantwortung einer Frage mit ja/nein. Wird die Frage mit ja beantwortet, geht man zum linken Ast des Baumes zum nächsten Knoten, andernfalls zum rechten Ast.
3. Endknoten / Blätter (terminal nodes): unterster Knoten

Beispiel: Studienanfänger Wirtschaftswissenschaften

1. MatheLK < 0.5? bzw. Hat der Studierende den Mathematik-Leistungskurs nicht besucht?

$\rightarrow$ Beantwortung mit ”nein”: Weiter zum Knoten im rechten Ast

2. Abitur88 < 0.5? bzw. Hat der Studierende sein Abitur nicht 1988 gemacht?

$\rightarrow$ Beantwortung mit ”nein”: Weiter zum Knoten im rechten Ast

3. Geschlecht < 0.5? bzw. Ist der Studierende männlich?

$\rightarrow$ Beantwortung mit ”j”: Weiter zum linken Ast mit Endknoten

$\rightarrow$ Zugeordnete Zahl: 1 $\rightarrow$ Gruppe 1

CART - Binäre Splits

- Art des Splits hängt vom Skalenniveau der Merkmale $x_1, \ldots, x_p$ ab
- Aufteilung eines Knotens A in die Knoten A_1 und A_2 hat die folgende Form:
  1. Falls x_j metrisch oder ordinal, bildet man

$$A_1 = A \cap \{x_j \leq c_j\} \quad \text{und} \quad A_2 = A \cap \{x_j > c_j\},$$

     mit Splitpunkt c_j aus dem Wertebereich S_j von x_j
  2. Falls x_j nominal, bildet man

$$A_1 = A \cap B \quad \text{und} \quad A_2 = A \cap \overline{B},$$

     wobei $B \subset S_j$ und $\overline{B} = S_j \setminus B$.

CART - Klassifikationsregel (2-Klassen-Fall)

- Gegeben: Baum mit Q Endknoten $A_1, \ldots, A_Q$
- Geschätzte Wahrscheinlichkeiten in jedem Knoten:

$$\hat{p}_{q1} = \frac{1}{n_q} \sum_{a_i \in A_q} I(Y_i = 1) \quad \text{und}$$

$$\hat{p}_{q2} = \frac{1}{n_q} \sum_{a_i \in A_q} I(Y_i = 2), \quad q = 1, \ldots Q\,.$$

Ordne alle Objekte $a_i \in A_q$ in Klasse 1 zu, falls $\hat{p}_{q1} > \hat{p}_{q2}$ (oder äquivalent falls $\hat{p}_{q1} > 0.5$), ansonsten zu Klasse 2.

Beispiel: Studienanfänger Wirtschaftswissenschaften

- p_{11}: Wahrscheinlichkeit im Wurzelknoten zu Gruppe 1 zu gehören
- $p_{11} > 0.5$: Gruppe 1
- $p_{11} < 0.5$: Gruppe 2
- $p_{11} = 0.5$: zufällige Zuteilung
- $\hat{p}_{11} = 0.45 \rightarrow$ ordne zufällig ausgewählten Studierenden Gruppe 2 zu

Unreinheit/Unsicherheit der Entscheidung

- Je näher p_{q1} an 0.5, desto fehlerhafter Entscheidung
- Quantifizierung durch Maßzahl?
- Betrachte Zufallsvariable Y:

$$Y = \begin{cases} 1 & \text{falls der Studierende zur Gruppe 1 gehört} \\ 0 & \text{falls der Studierende zur Gruppe 2 gehört} \end{cases}$$

- Y ist bernoulliverteilt mit Parameter p_{q1} und Varianz $p_{q1}(1 - p_{q1})$

$\rightarrow$ Je näher p_{q1} an 0.5, umso größer Varianz

$\rightarrow$ Varianz minimal, wenn p_{q1} 0 oder 1 ist

CART - Splitkriterien (2-Klassen-Fall)

- Betrachte in jedem Schritt ”Unreinheit” der bereits gebildeten Knoten
- Sei p_{q1} der Anteil von Klasse 1 im Knoten A_q, so kann man folgende Maße betrachten:

1. Fehlklassifikation:

$$F(A_q) = 1 - \max(p_{q1}, 1 - p_{q1})$$

2. Gini Index:

$$G(A_q) = 2\, p_{q1}(1 - p_{q1})$$

3. Entropy:

$$E(A_q) = -p_{q1} \log(p_{q1}) - (1 - p_{q1}) \log(1 - p_{q1})$$

Beispiel: Studienanfänger Wirtschaftswissenschaften

- Wurzelknoten: $p_{11} = 0.45$

$\rightarrow$ $G(A_1) = 2 \cdot 0.45(1 - 0.45) = 0.495$

$\rightarrow$ Unreinheit des Wurzelknotens groß, da beide Gruppe nahezu gleich häufig in der Population vertreten

- Zerlegung in Knoten A_2: 1 von 9 Studierenden, der keinen MatheLK besucht hat, ist in Gruppe 1 $\rightarrow p_{21} = \frac{1}{9}$

$\rightarrow$ $G(A_2) = 2 \cdot \frac{1}{9}\left(1 - \frac{1}{9}\right) = 0.1975$

- Zerlegung in Knoten A_3: 8 von 11 Studierenden, die einen MatheLK besucht haben, sind in Gruppe 1 $\rightarrow p_{31} = \frac{8}{11}$

$\rightarrow$ $G(A_3) = 2 \cdot \frac{8}{11}\left(1 - \frac{8}{11}\right) = 0.3967$

- Unreinheit in den Knoten A_2 und A_3 kleiner als im Wurzelknoten

$\rightarrow$ Aufteilung in Teilpopulationen, die hinsichtlich des Merkmals ”bestanden” homogener sind

CART - Baumkonstruktion

- Wähle in jedem Schritt der Baumkonstruktion Knoten, Variable x_j und Splitpunkt c_j aus, die Unreinheit in neuen Knoten minimiert
- Teilt man Knoten A_1 in Knoten A_2 und A_3 (unter Verwendung des Gini Index), so minimiert man

$$G(A_2, A_3) = n_2 \, G(A_2) \; + \; n_3 \, G(A_3) \, ,$$

wobei n_2 und n_3 Anzahl der Beobachtungen in den beiden neuen Knoten

Beispiel: Studienanfänger Wirtschaftswissenschaften

- MatheLK:

$$\begin{aligned} G(A_2, A_3) &= n_2 G(A_2) + n_3 G(A_3) \\ &= 1 \cdot 0.1975 + 8 \cdot 0.3967 \\ &= 3.3711 \end{aligned}$$

Merkmal	$G(A_2, A_3)$
Geschlecht	4.9
MatheLK	3.3711
Abitur88	4.4514

$\rightarrow$ Start mit MatheLK, dann Abitur88, dann Geschlecht

CART - Pruning

- (theoretische) Fortsetzung der Baumkonstruktion bis alle Knoten vollkommen homogen

$\rightarrow$ sehr starke Gefahr des Overfitting

Definiere ein geeignetes Stopkriterium, um die Größe des Baumes (d.h. die Anzahl an Splits) zu kontrollieren (”Pruning”).

Mögliche Stopkriterien:

- minimale Anzahl an Beobachtungen pro Blatt (Endknoten)
- minimale Verbesserung des Unreinheitsmaßes
- maximale Anzahl an Stufen/Levels des Baumes

Cost-Complexity Pruning

- Starte bei Blättern des Baumes und fasse diejenigen Knoten wieder zusammen, die zur kleinsten Verschlechterung der Klassifikationsgenauigkeit führen

→ Genestete Sequenz an Unterbäumen, die im Bezug auf die Klassifikationsgenauigkeit für eine gegebene Baumgröße optimal sind

- Bestimme aus Sequenz den optimalen Baum durch Minimierung einer Kosten-Komplexitäts-Funktion

→ Führt zu Kompromiss zwischen Klassifikationsgenauigkeit und Baumgröße

Kosten-Komplexitäts-Funktion

- Bestimme optimalen Baum durch Minimierung der Funktion

$$R_\alpha(T) = R(T) + \alpha\,|T|\,.$$

- $R(T)$ Unreinheit des Baumes, z.B. über den Gini Index
- $|T|$ Größe des Baumes, d.h. Anzahl an Blättern
- α Tuning-Parameter, der die Baumgröße steuert

$\rightarrow$ Bestimmung von α durch Resampling-Methoden, z.B. durch Kreuzvalidierung

k-fache Kreuzvalidierung

1. Setze α fest
2. Teile den Datensatz zufällig in k gleich große Teile auf
3. Für $k = 1, \ldots, K$:
   - Fitte den Baum T_k auf allen Daten außer Teil k
   - Berechne das Kostenkomplexitätskriterium $R_\alpha(T_k)$ auf Teil k
4. Berechne das Gesamtkriterium $R_\alpha(T) = \frac{1}{k} \sum_{r=1}^{k} R_\alpha(T_k)$
5. Wiederhole die Schritte 1 bis 4 für verschiedene Werte von α
6. Bestimme den Wert von α, der das Gesamtkriterium in Schritt 4 minimiert

Trees - Variablenselektion

- Durch Pruning kommen möglicherweise nicht alle Merkmale $x_1, \ldots, x_p$ im resultierenden Baum vor
- Baum selektiert automatisch die informativen Variablen und schließt gleichzeitig die nicht informativen Variablen für die Klassifikation aus
- Problem: Variablen mit einer großen Anzahl an Ausprägungen (möglichen Splits) werden bevorzugt ausgewählt.

$\rightarrow$ kann zu einer Verzerrung in der Variablenselektion führen

Mögliche Alternative

- Devianz eines Knotens A_q:

$$D_q = -2[n_{1q} \ln(p_{1q}) + (n_q - n_{1q}) \ln(1 - p_{1q})]$$

mit

- n_q: Anzahl der Beobachtungen im Knoten A_q
- n_{1q}: Anzahl der Beobachtungen im Knoten A_q, die zu Gruppe 1 gehören
- p_{1q}: Wahrscheinlichkeit, dass Beobachtung in Knoten A_q zu Gruppe 1 gehört

- Schätze $p_{1q} = n_{1q}/n_q$

$\rightarrow$ Geschätzte Devianz:

$$\hat{D}_q = -2[n_{1q} \ln(n_{1q}/n_q) + (n_q - n_{1q}) \ln(n_{1q}/n_q)]$$

$\rightarrow$ Wähle Merkmal zur Verzweigung, bei dem Verminderung der Devianz

$$\hat{D}_1 - \hat{D}_2 - \hat{D}_3$$

am größten ist

Beispiel: Studienanfänger Wirtschaftswissenschaften

- MatheLK: Wurzelknoten mit $n_q = 20$ und $n_{1q} = 9$

$$\hat{D}_1 = -2[9 \ln(9/20) + (20 - 9) \ln(1 - 9/20)] = 27.53$$

- 9 Studierende haben keinen MatheLK besucht, davon einer in Gruppe 1

$$\hat{D}_2 = -2[1 \ln(1/9) + (9 - 1) \ln(1 - 1/9)] = 6.28$$

- 11 Studierende haben MatheLK besucht, davon 8 in Gruppe 1

$$\hat{D}_3 = -2[8 \ln(8/11) + (11 - 8) \ln(1 - 8/11)] = 12.89$$

- Verminderung der Devianz:

$$\hat{D}_1 - \hat{D}_2 - \hat{D}_3 = 27.53 - 6.28 - 12.89 = 8.36$$

- Verminderung bei Geschlecht: 0.21
- Verminderung bei Abitur88: 0.03

Conditional Inference Trees

Conditional Inference Trees

- CART selektieren Variablen, jedoch steht kein Konzept der statistischen Signifikanz dahinter
- Problem addressiert durch Konzept der Conditional Inference Trees (Hothorn et al, 2006)
- Grundidee: Verwende p-Werte von Signifikanztests zur Selektion der Splits

Conditional Inference Trees - Konzept

- Teste in jedem Schritt der Baumkonstruktion (in jedem Knoten) die globale Nullhypothese H_0 der Unabhängigkeit zwischen der Zielvariable Y und allen Merkmalen $x_1, \ldots, x_p$
- Annahme: Verteilung der Zielvariable $D(Y)$ gegeben die Merkmale kann über Baumstruktur beschrieben werden, d.h.

$$D(Y|x_1, \ldots, x_p) = D(Y|\text{tr}(x_1, \ldots, x_p))$$

- Notwendig: Annahme über bedingte Verteilung der Zielvariable

Conditional Inference Trees - Algorithmus

1. Teste in jedem bereits gebildeten Knoten die Nullhypothese

$$H_0 = \bigcap_{j=1}^{p} H_0^j, \quad \text{wobei} \quad H_0^j : D(Y|x_j) = D(Y)$$

$\rightarrow$ Falls H_0 nicht abgelehnt wird, wird der jeweilige Knoten als Endknoten deklariert

2. Selektiere die Variable x_j^* mit der stärksten Assoziation
3. Bestimme zur Variable x_j^* den optimalen Splitpunkt c_j^* und führe den Split durch
4. Wiederhole die Schritte 1 bis 3 bis alle Knoten als Endknoten deklariert wurden

Conditional Inference Trees - Details

- In Schritt 2 des Algorithmus werden *p*-Werte der Tests verwendet

$\rightarrow$ unverzerrte Variablenselektion

- Approximation der wahren Verteilung der Teststatistiken nach Konzept basierend auf Permutationstests (Strasser und Weber, 1999)
- Da in jedem Schritt *p* Hypothesen gleichzeitig getestet werden, ist eine Adjustierung des Signifikanzniveaus α notwendig, z.B. über Bonferroni-Korrektur

Random Forests

Random Forests - Motivation

- Nachteil von Bäumen: hohe Instabilität, d.h. kleine Änderungen in Daten können zu sehr unterschiedlichen Bäumen führen

$\rightarrow$ Zwar einfach interpretierbar, aber für verlässliche Vorhersage oft ungeeignet

$\rightarrow$ Stabilisiere die Ergebnisse einzelner Bäume durch Betrachtung eines Ensembles vieler Bäume

$\rightarrow$ random forests

Random Forests - Grundprinzip

- Basieren auf Konzept des Bagging (Bootstrap Aggregating)
- Betrachtung von B Bootstrap-Stichproben (Ziehen mit Zurücklegen), auf denen jeweils ein Baum gefittet wird
- Klassifikation über Aggregation der Ergebnisse der B Bäume

Random Forests - Algorithmus

1. Ziehe Bootstrap Stichprobe aus dem originalen Datensatz
2. Fitte Baum (CART oder Conditional Inference Tree), wobei bei jedem Split nur $m < p$ zufällig ausgewählte Merkmale zur Auswahl herangezogen werden

$\rightarrow$ Verringerung der Korrelation zwischen den Bäumen

3. Wiederhole die Schritte 1 und 2 B mal.

Random Forests - Klassifikationsregel (2-Klassen-Fall)

- Algorithmus liefert geschätzte Wahrscheinlichkeiten $\hat{p}_{i1}^b$ und $\hat{p}_{i2}^b$ für alle Beobachtungen $i = 1, \ldots, n$ und alle Wiederholungen $b = 1, \ldots, B$

Averaging

Berechne die mittleren Wahrscheinlichkeiten

$$\hat{p}_{i1} = \frac{1}{B} \sum_{b=1}^{B} \hat{p}_{i1}^b \quad \text{und} \quad \hat{p}_{i2} = \frac{1}{B} \sum_{b=1}^{B} \hat{p}_{i2}^b \, .$$

Ordne das Objekt a_i in Klasse 1 zu, falls $\hat{p}_{i1} > \hat{p}_{i2}$, ansonsten zu Klasse 2.

Random Forests - Klassifikationsregel (2-Klassen-Fall)

- Algorithmus liefert geschätzte Wahrscheinlichkeiten $\hat{p}_{i1}^b$ und $\hat{p}_{i2}^b$ für alle Beobachtungen $i = 1, \ldots, n$ und alle Wiederholungen $b = 1, \ldots, B$

Majority Vote

Bestimme die Klassifikation für alle B Wiederholungen.

Ordne das Objekt a_i in Klasse 1 zu, falls a_i häufiger in Klasse 1 zugeordnet wurde, ansonsten zu Klasse 2.

Random Forests - Variablenwichtigkeit

- Einzelne Bäume: einfach interpretierbar
- Random forests: nicht einfach interpretierbar

$\rightarrow$ Einfluss einzelner Merkmale schwierig zu evaluieren

- Lösung: Variablenwichtigkeitsmaße (basierend auf Permutationen der Daten oder basierend auf Splitkriterium)

Permutation-Variablenwichtigkeit

- $i_b = 1, \ldots, n_b, \ b = 1, \ldots, B,$: Beobachtungen der Bootstrap Stichproben
- $i_{\bar{b}} = 1, \ldots, n_{\bar{b}}$: Beobachtungen, die jeweils nicht zur Baumkonstruktion verwendet wurden (”out of bag” Beobachtungen)
- Bestimme für jeden der B Bäume und jede Variable x_j die Differenz der Klassifikationsgenauigkeit

$$VI^b(x_j) = \frac{\sum_{i_{\bar{b}}=1}^{n_{\bar{b}}} I\left(Y_{i_{\bar{b}}} = \hat{Y}^b_{i_{\bar{b}}}\right)}{n_{\bar{b}}} - \frac{\sum_{i_{\bar{b}}=1}^{n_{\bar{b}}} I\left(Y_{i_{\bar{b}}} = \hat{Y}^b_{i_{\bar{b}},j}\right)}{n_{\bar{b}}}.$$

- $\hat{Y}^b_{i_{\bar{b}}}$ geschätzte Klasse auf der originalen Bootstrap Stichprobe
- $\hat{Y}^b_{i_{\bar{b}},j}$ geschätzte Klasse nach Permutation von Variable x_j

Permutation-Variablenwichtigkeit

- Berechne jeweils die mittlere Differenz

$$VI(x_j) = \frac{1}{B} \sum_{b=1}^{B} VI^b(x_j), \; j = 1, \ldots, p \,,$$

  als Maß für die Variablenwichtigkeit
- Beachte: $VI^b(x_j) = 0$ falls x_j im b-ten Baum nicht vorkommt