Klausur zur Vorlesung

Maschinelles Lernen: Symbolische Ansätze

Prof. J. Fürnkranz
Technische Universität Darmstadt — Wintersemester 2010/11
Termin: 2. 3. 2011

Name: **Vorname:** **Matrikelnummer:**

Fachrichtung:

Punkte: (1) .... (2) .... (3) .... (4) .... (5) .... **Summe:**

- **Aufgaben:** Diese Klausur enthält auf den folgenden Seiten 5 Aufgaben zu insgesamt 100 Punkten. Jede Aufgabe steht auf einem eigenen Blatt. Kontrollieren Sie *sofort*, ob Sie alle Blätter erhalten haben!
  Beachten Sie, daß sich oft auch auf den Rückseiten der Blätter Teilaufgaben finden!
- **Zeiteinteilung:** Die Zeit ist knapp bemessen. Wir empfehlen Ihnen, daß Sie sich zuerst einen kurzen Überblick über die Aufgabenstellungen verschaffen, und dann mit den Aufgaben beginnen, die Ihnen am meisten liegen.
- **Papier:** Verwenden Sie nur Papier, das Sie von uns ausgeteilt bekommen. Sie können Ihre Lösungen beliebig auf alle Blätter verteilen, solange klar ersichtlich ist, welche Lösung zu welcher Aufgabe gehört. Sollten sich allerdings mehrere Lösungen zu derselben Aufgabe finden, suchen wir uns eine aus.
  Insbesondere können Sie auch auf den Rückseiten schreiben!
  Brauchen Sie zusätzlich Papier (auch Schmierpaper), bitte melden.
- **Fragen:** Sollten Sie Teile der Aufgabenstellung nicht verstehen, fragen Sie bitte!
- **Abschreiben:** Sollte es sich (wie in den letzten Jahren leider immer wieder) herausstellen, daß Ihre Lösung und die eines Kommilitonen über das zu erwartende Maß hinaus übereinstimmen, werden beide Arbeiten negativ beurteilt (ganz egal wer von wem in welchem Umfang abgeschrieben hat).
- **Ausweis:** Legen Sie Ihren *Studentenausweis* sichtbar auf Ihren Platz.
- **Hilfsmittel:** Zur Lösung der Aufgaben ist ein von Ihnen selbst handschriftlich beschriebenes DIN-A4-Blatt erlaubt. Gedruckte Wörterbücher sind für ausländische Studenten erlaubt, elektronische Hilfsmittel (Taschenrechner, elektronische Wörterbücher, Handy, etc.) sind verboten! Sollten Sie etwas verwenden wollen, was nicht in diese Kategorien fällt, klären Sie das bitte *bevor* Sie zu arbeiten beginnen.
- **Aufräumen:** Sonst darf außer Schreibgerät, Essbarem, von uns ausgeteiltem Papier und eventuell Wörterbüchern nichts auf Ihrem Platz liegen. Taschen bitte unter den Tisch!

Gutes Gelingen!

Aufgabe 1 Regellernen und ROC-Analyse (6/3/8/3/4 = 24 Punkte)

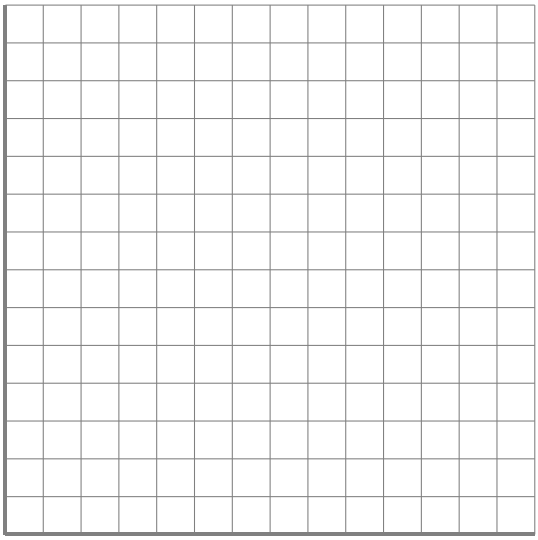
Gegeben sei folgender Datensatz, der die Ergebnisse einer Umfrage beinhaltet.

Beispiel	sex	marital	children	education	answer
1	male	single	true	primary	+
2	male	married	false	secondary	+
3	male	single	false	university	+
4	female	divorced	true	university	+
5	male	divorced	false	secondary	+
6	male	married	false	secondary	+
7	male	single	true	primary	−
8	male	married	false	university	−
9	female	single	false	primary	−
10	female	divorced	false	secondary	−
11	female	married	true	primary	−
12	male	divorced	true	secondary	−

Aus diesen Daten hat ein Separate-and-Conquer Algorithmus bisher folgende Regel gelernt, wobei die Bedingungen in der angegebenen Reihenfolge (von links nach rechts) gelernt wurden.

R_1: **if** sex = male **and** children = false **and** education = secondary **then** +

1–a Zeichnen Sie das Lernen der Regel R_1 im Coverage Space. Beschriften Sie die Dimensionen des Raums, sowie alle gezeichneten Punkte mit den jeweiligen Koordinaten (konkrete Zahlen). Nehmen Sie an, daß die Bedingungen in der angegebenen Reihenfolge (von links nach rechts) gefunden wurden.



Hinweis: Der karierte Teil des Papiers ist als Zeichenhilfe gedacht. Er gibt keinerlei Aufschluss über die tatsächlichen Dimensionen des Raumes.

1–b Was ändert sich, wenn Sie dieses Problem im ROC-Raum statt im Coverage-Raum darstellen sollen?

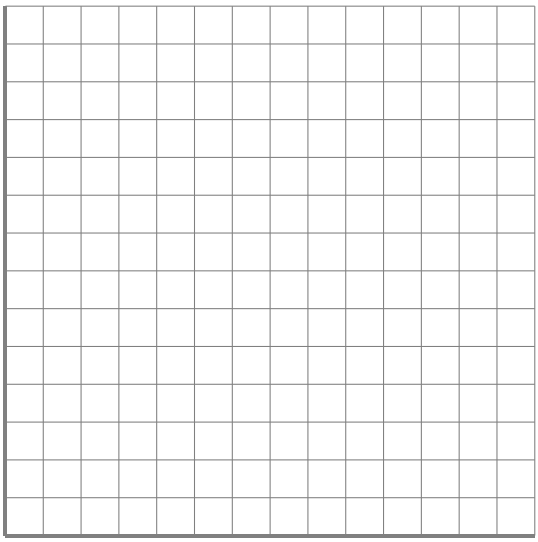
1–c Geben Sie für jede der beim Lernen von R_1 evaluierten Teilregeln (inklusive der universellen Regel, die alle Beispiele abdeckt, und der leeren Regel, die keine Beispiele abdeckt) an, für welches Kostenverhältnis von positiven und negativen Beispielen $\frac{c(+|-)}{c(-|+)}$ diese Teilregel optimal wäre.

1–d Mit welcher der in der Vorlesung besprochenen Regel-Lern-Heuristiken ließe sich eine Auswahl nach einem vorgegebenen Kostenverhältnis realisieren?

1–e Nachdem Regel R_1 gelernt wurde, setzt der Separate-and-Conquer Regel-Lerner fort und beginnt die nächste Regel wie folgt:

R_2: **if** marital = divorced **then** +

Zeichnen und beschriften Sie den Coverage Space, in dem diese Regel gelernt wird, und markieren Sie die Lage der Regel in diesem Raum.

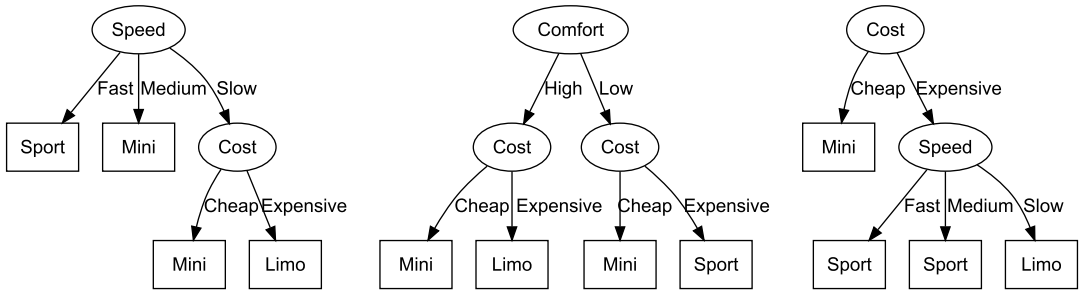


Aufgabe 2 Ensemble Methoden (4/4/3/6/3 = 20 Punkte)

Gegeben sei folgende Datenmenge

Speed	Comfort	Cost	Class
Fast	High	Expensive	Limo
Slow	High	Expensive	Limo
Slow	Low	Expensive	Limo
Fast	High	Cheap	Mini
Medium	Low	Cheap	Mini
Slow	High	Cheap	Mini
Slow	Low	Cheap	Mini
Fast	Low	Cheap	Sport
Fast	Low	Expensive	Sport
Medium	High	Expensive	Sport

Auf dieser Datenmenge wurden durch ein Ensemble-Verfahren die folgenden Bäume gelernt:



2–a Wandeln Sie den linken Teilbaum in eine äquivalente Regelmenge um.

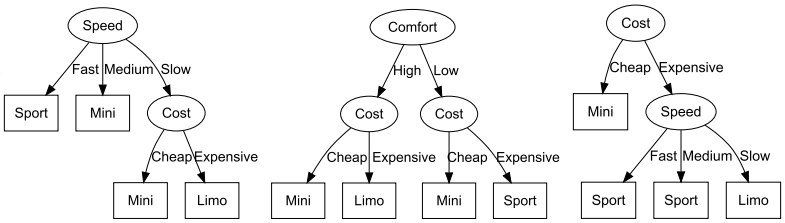
2–b Bestimmen Sie den Average Gini-Index für den Wurzelknoten des linken Baums (Attribut Speed) auf der gesamten Datenmenge.

2–c Wie würde folgendes Beispiel klassifiziert werden, wenn die Vorhersagen der Bäume mit ungewichtetem Voting kombiniert werden?

Speed	Comfort	Cost	Class
Medium	Low	Expensive	?

Hier nochmals die Beispiele und Bäume

Speed	Comfort	Cost	Class
Fast	High	Expensive	Limo
Slow	High	Expensive	Limo
Slow	Low	Expensive	Limo
Fast	High	Cheap	Mini
Medium	Low	Cheap	Mini
Slow	High	Cheap	Mini
Slow	Low	Cheap	Mini
Fast	Low	Cheap	Sports
Fast	Low	Expensive	Sports
Medium	High	Expensive	Sports



2–d Verwenden Sie diese Bäume zur Erstellung eines Meta-Datensatzes für Stacking.
Hinweis: Sie können die folgende leere Tabelle zum Eintragen der Werte verwenden. Dabei müssen nicht unbedingt alle Felder gefüllt werden.

2–e Verwenden Sie den Meta-Datensatz um das folgende Beispiel:

Speed	Comfort	Cost	Class
Medium	Low	Expensive	?

mit 1-Nearest-Neighbour mit Manhattan-Distanz zu klassifizieren. Sollte mehr als ein nächster Nachbar existieren, sagen Sie die Majority-Class innerhalb aller nächsten Nachbarn vorher. Geben Sie den/die nächsten Nachbarn an.

Aufgabe 3 Assoziationsregeln (4/3/8/4 = 19 Punkte)

Sie möchten in einer Anwendung alle Frequent Itemsets mit einem Minimum Support von 0.2 erhalten. Sie erhalten als Ergebnis folgende Liste von Itemsets mit dem jeweiligen Support.

$$
\begin{aligned}
I \quad = \quad & \{\{A\}:0.70, \{B\}:0.52, \{C\}:0.60, \{D\}:0.65, \\
& \{A,B\}:??, \{A,C\}:0.50, \{A,D\}:0.48, \{B,C\}:0.38, \{B,D\}:0.40, \{B,E\}:0.46, \\
& \{C,D\}:0.52, \{C,E\}:0.56, \{D,E\}:0.61, \\
& \{A,B,C\}:0.30, \{A,B,D\}:0.35, \{B,C,D\}:.32, \{C,D,E\}:0.50, \{B,C,E\}:0.35, \{B,D,E\}:0.35, \\
& \{A,B,C,D\}:0.25\}
\end{aligned}
$$

3–a In der Menge I fehlen einige Itemsets, die ebenfalls frequent sein *müssen*. Welche sind das?

Hinweis: Alle oben genannten Itemsets sind richtig. Nennen Sie nur die fehlenden frequent Itemsets. Zusätzlich genannte Itemsets zählen als Fehler.

3–b Für das Itemset $\{A, B\}$ ist kein Support angegeben. Geben Sie eine möglichste große untere und eine möglichst kleine obere Schranke an.

3–c Berechnen Sie für die Assoziationsregel $A, C \rightarrow B, D$

1. confidence
2. lift
3. leverage

3–d Für eine andere Datenmenge wissen Sie, daß das Itemset $I_1 = \{X, Y, Z\}$ ein Element der positiven Border ist, und daß das Itemset $I_2 = \{W, X, Y\}$ ein Element der negativen Border ist.

- Geben Sie die Menge aller Frequent Itemsets an.
- Gibt es noch Elemente, die in der positiven oder der negativen Border sein *müssen*?

Aufgabe 4 Version Space (8/3/3/3 = 17 Punkte)

Gegeben seien die folgenden Daten, die das Konzept "japanischer Kleinwagen" repräsentieren. Die Repräsentationssprache läßt nur Konjunktionen einzelner Attribut/Wert-Tests zu, Disjunktionen (mehrere Werte) sind nicht erlaubt.

Origin	Manufacturer	Color	Decade	Type	Class
Japan	Honda	Blue	1980	Economy	+
Japan	Toyota	Green	1970	Sports	−
Japan	Toyota	Blue	1990	Economy	+
USA	Chrysler	Red	1980	Economy	−
Japan	Honda	White	1980	Economy	+

4–a Wenden Sie den Candidate Elimination Algorithm an und geben Sie nach jedem Schritt das S-Set und das G-Set an.

4–b Geben Sie nach jedem Schritt des Candidate Algorithmus in der vorherigen Teilaufgabe an, wie der jeweilige Version Space das folgende Beispiel klassifizieren würde (mit kurzer Begründung):

USA	Chrysler	White	1980	Sports	?

4–c Nehmen Sie an, Sie erhalten nun als zusätzliches, sechstes Lern-Beispiel:

Japan	Toyota	Red	1990	Sports	+

Was würde passieren? Welche Schlußfolgerungen können Sie daraus ziehen?

4–d Angenommen, Sie wissen, daß Sie für ein gegebenes Datenset mit Hilfe des Candidate Elimination Algorithmus eine vollständige und konsistente Regel für die positive Klasse finden können. Können Sie dann mit diesem Algorithmus auch notwendigerweise eine vollständige und konsistente Regel für die negative Klasse finden? Warum (nicht)?

Aufgabe 5 Allgemeines (4/4/4/4/4 = 20 Punkte)

5–a Was versteht man unter dem Begriff Pre-Pruning? In der Vorlesung haben Sie 3 verschiedene Möglichkeiten kennengelernt wie man Pre-Pruning in einen Regel-Lern Algorithmus integrieren kann. Nennen und erklären Sie eine davon.

5–b Was versteht man unter der konvexen Hülle im ROC-Raum? Was kennzeichnet Klassifizierer unterhalb der konvexen Hülle?

5–c Der Algorithmus RELIEF wird zur Erstellung von Gewichten für die Attribute verwendet. Was versteht man bei diesem Algorithmus unter den beiden Mengen *hits* und *misses*? Erklären Sie außerdem kurz die Grundidee des Algorithmus anhand der beiden Mengen.

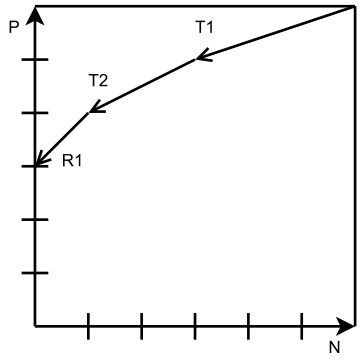
5–d Die Klassifizierungsphase des NEAREST NEIGHBOUR Algorithmus ist sehr ineffizient. Mit welcher Datenstruktur können wir diese Phase beschleunigen? Erklären Sie ganz kurz die Grundidee.

5–e Was ist der Unterschied zwischen Divide-and-conquer und Separate-and-conquer Algorithmen? Nennen Sie jeweils einen Algorithmus pro Strategie.

Lösung 1

1–a Coverage Space: 6x6

- T1: **if** sex = male **then** +: (p,n)=(5,3)
- T2: **if** sex = male **and** children = false **then** +: (p,n)=(4,1)
- R1: **if** sex = male **and** children = false **and** education = secondary **then** +: (p,n)=(3,0)



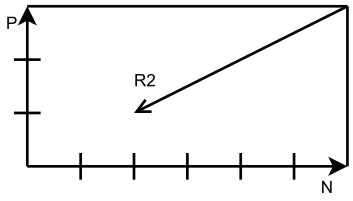
1–b Die Dimensionen bleiben erhalten, da eine Gleichverteilung vorliegt. Es werden nun anstatt p und n (absolut) tpr und fpr (relativ) an den Achsen aufgetragen.

1–c $c := \frac{c(+|-)}{c(-|+)}$

- Universelle Regel: $c \leq \frac{1}{3}$
- T1: $\frac{1}{3} \leq c \leq \frac{1}{2}$
- T2: $\frac{1}{2} \leq c \leq 1$
- R1: $c \geq 1$
- Leere Regel: $c(-|+) = 0$, $c(-|+)$ beliebig.

1–d Linear Cost Metric, Relative Cost Metric

1–e R2: **if** marital = divorced **then** +: (p,n)=(1,2)



Lösung 2

2–a Regelmenge des linken Teilbaums:

- Speed=Fast → Sport
- Speed=Medium → Mini
- Speed=Slow And Cost=Cheap → Mini
- Speed=Slow And Cost=Expensive → Limo

2–b Berechnung des Gini-Index:

- Fast (1/1/2): $1 - \left(\frac{1}{4}\right)^2 - \left(\frac{1}{4}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{5}{8}$
- Medium (0/1/1): $1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0,5$

- Slow (2/2/0): $1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0,5$
- Total: Gini(S, Speed)=$\frac{4}{10} \cdot \frac{5}{8} + \frac{2}{10} \cdot \frac{1}{2} + \frac{4}{10} \cdot \frac{1}{2} = \frac{11}{20}$

2–c Der erste Baum sagt für das Beispiel Mini voraus, die beiden anderen jedoch Sport. Das Beispiel wird als Sport klassifiziert.

2–d Meta-Datensatz:

B1	B2	B3	Type
Sport	Limo	Sport	Limo
Limo	Limo	Limo	Limo
Limo	Sport	Limo	Limo
Sport	Mini	Mini	Mini
Mini	Mini	Mini	Mini
Mini	Mini	Mini	Mini
Mini	Mini	Mini	Mini
Sport	Mini	Mini	Sport
Sport	Sport	Sport	Sport
Mini	Limo	Sport	Sport

2–e Konvertieren

Speed	Comfort	Cost	Class
Medium	Low	Expensive	?

zu

B1	B2	B3	Class
Mini	Sport	Sport	?

Distanzen zwischen dem Beispiel und den Beispielen des Meta-Datensatzes

B1	B2	B3	Type	Distanz
Sport	Limo	Sport	Limo	2
Limo	Limo	Limo	Limo	3
Limo	Sport	Limo	Limo	2
Sport	Mini	Mini	Mini	3
Mini	Mini	Mini	Mini	2
Mini	Mini	Mini	Mini	2
Mini	Mini	Mini	Mini	2
Sport	Mini	Mini	Sport	3
Sport	Sport	Sport	Sport	1
Mini	Limo	Sport	Sport	1

Minimale Distanz zu den beiden letzten Beispiele (2xSport), Vorhersage: Sport

Lösung 3

3–a $\{E\}$ (da z.B. $\{B, E\}$ frequent ist) und $\{A, C, D\}$ (da $\{A, B, C, D\}$ frequent ist) sind auch frequent.

3–b
- $support(A, B) \leq \min(support(A), support(B)) = 0.52$
- $support(A, B) \geq \max(support(A, B, C), support(A, B, D)) = 0.35$

3–c
- $confidence(A, C \rightarrow B, D) = \frac{support(A,B,C,D)}{support(A,C)} = 0.5$
- $lift(A, C \rightarrow B, D) = \frac{support(A,B,C,D)}{support(A,C)support(B,D)} = 1.25$
- $leverage(A, C \rightarrow B, D) = support(A, B, C, D) - support(A, C)support(B, D) = 0.05$

3–d
- $I = \{\{W\}, \{X\}, \{Y\}, \{Z\}, \{W, X\}, \{W, Y\}, \{X, Y\}, \{X, Z\}, \{Y, Z\}, \{X, Y, Z\}\}$
- Das Grundproblem ist, daß man nicht weiß, ob $\{W, Z\}$ frequent ist oder nicht, bzw. (wenn es frequent ist), ob seine Obermengen $\{W, X, Z\}$ bzw. $\{W, Y, Z\}$ frequent sind oder nicht. Davon hängt z.B. ab, ob $\{W, X\}, \{W, Y\}$ zur positiven Border gehören oder nicht.
  Die Antwort $\{W, X\}$ und $\{W, Y\}$ wurde dennoch als richtig gewertet.

Lösung 4

4–a Bsp1: $G_1 = \{(?,?,?,?,?)\}$, $S_1 = \{(Japan, Honda, Blue, 1980, Economy)\}$

Bsp2: $G_2 = \{(?, Honda, ?, ?, ?), (?, ?, Blue, ?, ?), (?, ?, ?, 1980, ?), (?, ?, ?, ?, Economy)\}$, $S_2 = S_1$

Bsp3: $G_3 = \{(?, ?, Blue, ?, ?), (?, ?, ?, ?, Economy)\}$, $S_3 = \{(Japan, ?, Blue, ?, Economy)\}$

Bsp4: $G_4 = \{(?, ?, Blue, ?, ?), (Japan, ?, ?, ?, Economy)\}$, $S4 = S_3$

Bsp5: $G_5 = \{(Japan, ?, ?, ?, Economy)\}$, $S_5 = \{(Japan, ?, ?, ?, Economy)\}$

4–b Nach 1: ?
Nach 2: ?
Nach 3: −
Nach 4: −
Nach 5: −

4–c Der Version Space kollabiert. S müßte generalisiert werden, sodaß es allgemeiner als G wird, was nicht sein darf.

Mögliche Schlußfolgerungen:

- Eines der Beispiele ist fehlerhaft (nicht notwendigerweise dieses) oder
- Das Konzept ist nicht mit einer Regel darstellbar

4–d Nein, da die Negation einer Konjunktion eine Disjunktion ist, also aus mehrere Regeln besteht. Es ist natürlich möglich, daß es dennoch eine alternative Beschreibung mit einer Regel gibt, aber zwingend ist es nicht.

Lösung 5

5–a Pre-Pruning bezeichnet den Mechanismus bereits während dem Lernen zu prunen, also das Modell gegen Overfitting zu schützen.

Möglichkeit 1: Threshold, es werden nur die Kandidatenregeln zugelassen, die einen heuristischen Wert oberhalb des Thresholds haben

Möglichkeit 2: MDL (Minimum Description Length), es werden nur die Kandidatenregeln zugelassen, deren minimale Beschreibungslänge (Informationsgehalt) geringer ist als die Beschreibungslänge der von der Regel abgedeckten Beispiele.

Möglichkeit 3: Signifikanztests, hier wird jede Kandidatenregel darauf geprüft, ob sich die Verteilung der von ihr abgedeckten Beispiele signifikant von der Verteilung aller Beispiele unterscheidet.

5–b Die konvexe Hülle einer Menge von Punkten im ROC-Raum verbindet alle Punkte, die für irgendein Kostenmodell optimal sind. Klassifizierer die unterhalb der konvexen Hülle liegen sind für kein Kostenverhältnis optimal und können demnach ignoriert werden.

5–c Die Menge der *hits* beinhaltet für die Nachbarschaft der Instanz die Instanzen, die den gleichen Klassenwert wie das Beispiel x haben und die Menge der *misses* beinhaltet genau die Instanzen, die unterschiedliche Klassenwerte haben. Ein gutes Attribut sollte daher mit möglichst vielen *hits* den gleichen Attributwert teilen, und sich von möglichst vielen *misses* anhand des Attributwertes unterscheiden.

5–d Datenstrukturen: kd-Trees (oder ball trees).

Idee: es müssen nicht alle Instanzen betrachtet werden, da man durch die Baumstruktur abschätzen kann, dass es in bestimmten Zweigen keinen näheren Nachbarn geben kann.

5–e Bei Divide-and-conquer Algorithmen wird die Datenmenge mit jedem Split weiter aufgeteilt, bis man "reine" Knoten (Blätter) hat. Bei Separate-and-conquer Algorithmen werden abgedeckte Instanzen sukzessive aus der Datenmenge entfernt.

Vertreter für Divide-and-conquer: Baum-Lerner, Vertreter für Separate-and-conquer: Regel-Lerner