Vorlesung Wissensentdeckung
Additive Modelle

Katharina Morik, Weihs

LS 8 Informatik
Computergestützte Statistik
Technische Universität Dortmund

2.6.2015

Gliederung

Ausgangspunkt: Funktionsapproximation

- Die bisher vorgestellten Lernverfahren, sind Instanzen der Funktionsapproximation.
- Gegeben sind die Trainingsbeispiele $\mathcal{T}$, gesucht ist eine Funktion

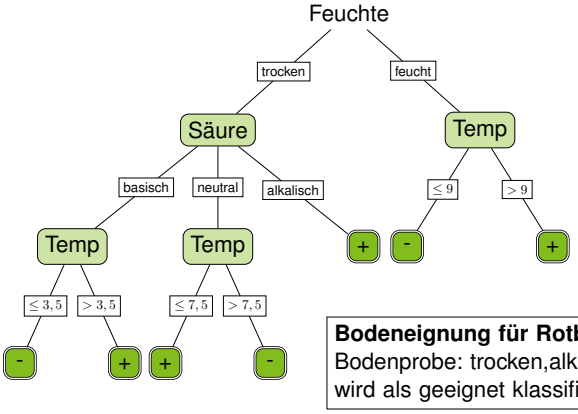
$$f_\theta(x) = \sum_{m=1}^{M} h_m(x)\theta_m$$

- Dabei gibt es Parameter θ, die abzuschätzen sind, bei den linearen Modellen ist dies $\hat{\beta}$.
- $h_m(x)$ ist eine Transformation der Beispiele.

Aufteilen der Beispiele und Modellierung jeder Region

- Lineare Modelle passen die Parameter für den gesamten Raum der Beispiele an, der evtl. durch eine implizite Transformation (Kernfunktionen) oder explizite Transformationen (Vorverarbeitung) in einen Merkmalsraum überführt wurde.
- Baumlerner teilen den Merkmalsraum in Rechtecke auf und passen in jedem ein Modell an. Dabei wird die Wahl des Merkmals in der rekursiven Aufteilung automatisch bestimmt.
- kNN teilt den Raum der Beispiele bei einer Anfrage x in die Nachbarschaft von x und den Rest auf.

Klassifizieren mit Entscheidungsbäumen



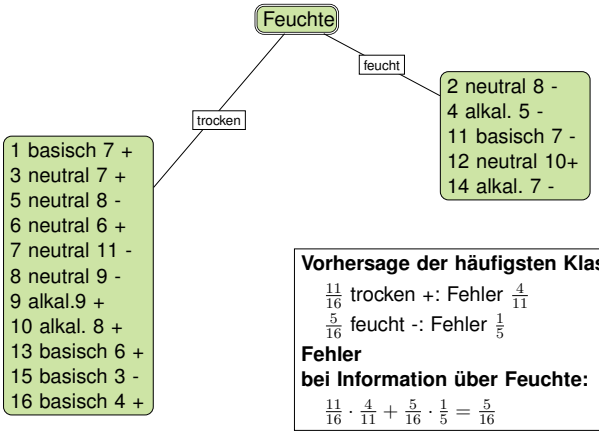
Bodeneignung für Rotbuchen:
Bodenprobe: trocken,alkalisch,7
wird als geeignet klassifiziert (+)

Lernen aus Beispielen

+				-			
ID	Feuchte	Säure	Temp	ID	Feuchte	Säure	Temp
1	trocken	basisch	7	2	feucht	neutral	8
3	trocken	neutral	7	4	feucht	alkal.	5
6	trocken	neutral	6	5	trocken	neutral	8
9	trocken	alkal.	9	7	trocken	neutral	11
10	trocken	alkal.	8	8	trocken	neutral	9
12	feucht	neutral	10	11	feucht	basisch	7
13	trocken	basisch	6	14	feucht	alkal.	7
16	trocken	basisch	4	15	trocken	basisch	3

Ohne weiteres Wissen können wir als Vorhersage immer - sagen. Der Fehler ist dann 8/16.

Aufteilen nach Bodenfeuchte



Vorhersage der häufigsten Klasse:
$\frac{11}{16}$ trocken +: Fehler $\frac{4}{11}$
$\frac{5}{16}$ feucht -: Fehler $\frac{1}{5}$
Fehler
bei Information über Feuchte:
$\frac{11}{16} \cdot \frac{4}{11} + \frac{5}{16} \cdot \frac{1}{5} = \frac{5}{16}$

Bedingte Wahrscheinlichkeit

- Wahrscheinlichkeit, dass ein Beispiel zu einer Klasse gehört, gegeben der Merkmalswert

$$P(Y|X_j) = P(Y \cap X_j)/P(X_j)$$

- Annäherung der Wahrscheinlichkeit über die Häufigkeit
- Gewichtung bezüglich der Oberklasse
- Beispiel: $Y = \{+, -\}, X_j = \{feucht, trocken\}$

$P(+|feucht) = 1/5,\ P(-|feucht) = 4/5$ gewichtet mit $5/16$

$P(+|trocken) = 7/11,\ P(-|trocken) = 4/11$ gewichtet mit $11/16$

Wahl des Merkmals mit dem höchsten Wert (kleinsten Fehler)

Information eines Merkmals

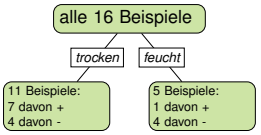
- Wir betrachten ein Merkmal als Information.
- Wahrscheinlichkeit p_+, dass das Beispiel der Klasse + entstammt. $I(p_+, p_-) = (-p_+ \log p_+) + (-p_- \log p_-)$ Entropie
- Ein Merkmal X_j mit k Werten teilt eine Menge von Beispielen $\mathbf{X}$ in k Untermengen $\mathbf{X}_1, ..., \mathbf{X}_k$ auf. Für jede dieser Mengen berechnen wir die Entropie.

$$Information(X_j, \mathbf{X}) := -\sum_{i=1}^{k} \frac{|\mathbf{X}_i|}{|\mathbf{X}|} I(p_+, p_-)$$

- Der Informationsgewinn ist die Differenz zwischen der Entropie der Beispiele mit und ohne die Aufteilung durch X_j.

Feuchte

Güte des Attributs Feuchte mit den 2 Werten *trocken* und *feucht*:

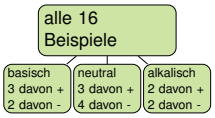


$$-\left[\underbrace{\frac{11}{16} \cdot I(+,-)}_{trocken} + \underbrace{\frac{5}{16} \cdot I(+,-)}_{feucht}\right]$$

$$= -\left[\underbrace{\frac{11}{16} \cdot \left(-\frac{7}{11} \cdot \log\left(\frac{7}{11}\right) - \frac{4}{11} \cdot \log\left(\frac{4}{11}\right)\right)}_{trocken} + \underbrace{\frac{5}{16}\left(-\frac{1}{5} \cdot \log\left(\frac{1}{5}\right) - \frac{4}{5} \cdot \log\left(\frac{4}{5}\right)\right)}_{feucht}\right] = -0,27$$

Säure

Güte des Attributs Säure mit den 3 Werten basisch, neutral und alkalisch:



$$-\left(\underbrace{\frac{5}{16} \cdot I(+,-)}_{basisch} + \underbrace{\frac{7}{16} \cdot I(+,-)}_{neutral} + \underbrace{\frac{4}{16} \cdot I(+,-)}_{alkalisch}\right)$$

$$= -0,3$$

basisch $-\frac{3}{5} \cdot \log\left(\frac{3}{5}\right) + -\frac{2}{5} \cdot \log\left(\frac{2}{5}\right)$

neutral $-\frac{3}{7} \cdot \log\left(\frac{3}{7}\right) + -\frac{4}{7} \cdot \log\left(\frac{4}{7}\right)$

alkalisch $-\frac{2}{4} \cdot \log\left(\frac{2}{4}\right) + -\frac{2}{4} \cdot \log\left(\frac{2}{4}\right)$

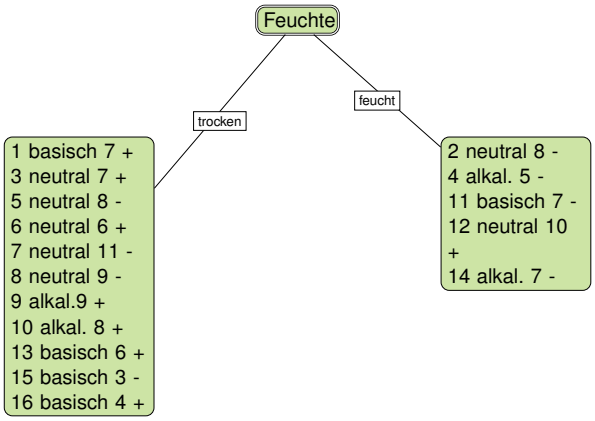
Temperatur

- Numerische Merkmalswerte werden nach Schwellwerten eingeteilt.
  - 9 verschiedene Werte in der Beispielmenge, also 8 Möglichkeiten zu trennen.
  - Wert mit der kleinsten Fehlerrate bei Vorhersage der Mehrheitsklasse liegt bei 7.
  - 5 Beispiele mit Temp < 7, davon 3 in +,
    11 Beispiele Temp $\gtrsim$ 7, davon 6 in -.
- Die Güte der Temperatur als Merkmal ist $-0,29$.

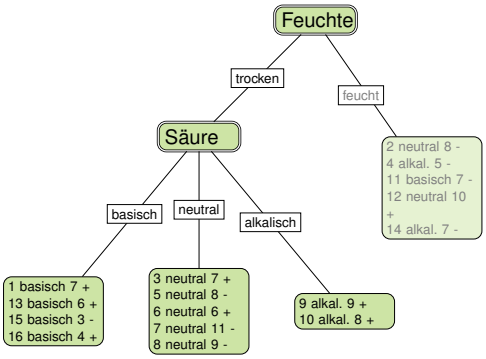
Merkmalsauswahl

- Gewählt wird das Merkmal X_j, dessen Werte am besten in (Unter-)mengen $\mathbf{X}_i$ aufteilen, die geordnet sind.
- Das Gütekriterium Information (Entropie) bestimmt die Ordnung der Mengen.
- Im Beispiel hat *Feuchte* den höchsten Gütewert.

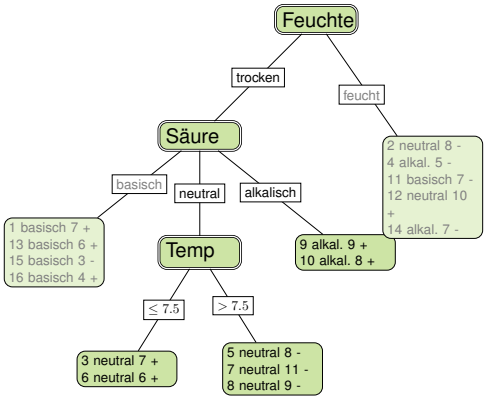
Algorithmus Top Down Induction of Decision Trees (TDIDT, hier: ID3) am Beispiel



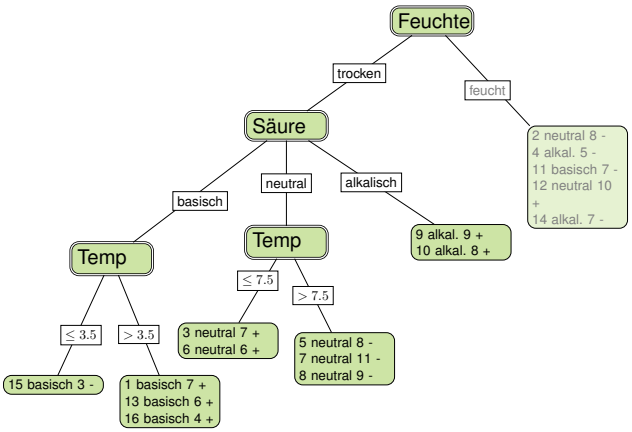
Algorithmus TDIDT (ID3) am Beispiel



Algorithmus TDIDT (ID3) am Beispiel



Algorithmus TDIDT (ID3) am Beispiel



Algorithmus ID3 (TDIDT)

Rekursive Aufteilung der Beispielmenge nach Merkmalsauswahl:

1. $TDIDT(\mathbf{X}, \{X_1, ...X_p\})$
2. $\mathbf{X}$ enthält nur Beispiele einer Klasse $\rightarrow$ fertig
3. $\mathbf{X}$ enthält Beispiele verschiedener Klassen:
   - $Güte(X_1, .., X_p, \mathbf{X})$
   - Wahl des besten Merkmals X_j mit k Werten
     - Aufteilung von $\mathbf{X}$ in $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_k$
     - für $i = 1, ..., k$: $TDIDT(\mathbf{X}_i, \{X_1, ..., X_p\} \backslash X_j)$
   - Resultat ist aktueller Knoten mit den Teilbäumen $T_1, ..., T_k$

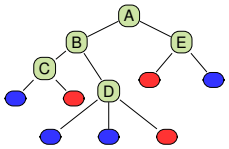
Komplexität TDIDT ohne Pruning

Rekursive Aufteilung der Beispielmenge nach Merkmalsauswahl:

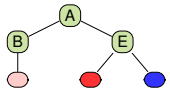
- Bei p (nicht-numerischen) Merkmalen und N Beispielen ist die Komplexität $\mathcal{O}(pN \log N)$
  - Die Tiefe des Baums sei in $\mathcal{O}(\log N)$.
  - $\mathcal{O}(N \log N)$ alle Beispiele müssen "in die Tiefe verteilt" werden, also: $\mathcal{O}(N \log N)$ für ein Merkmal.
  - p mal bei p Merkmalen!

Stutzen

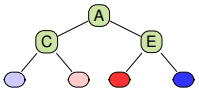
- Überanpassung des Baums an die Trainingsdaten verringern!
- Verständlichkeit erhöhen!
- Stutzen (Pruning):
  a) Knoten an Stelle eines Teilbaums setzen
  b) Einen Teilbaum eine Ebene höher ziehen
- Schätzen, wie sich der wahre Fehler beim Stutzen entwickelt.



a) Knoten an Stelle eines Teilbaums setzen



b) Einen Teilbaum eine Ebene höher ziehen



Stutzen durch Fehlerschätzen

- Wenn der Fehler eines Knotens kleiner ist als die Summe der Fehler seiner Unterknoten, können die Unterknoten weggestutzt werden.
- Dazu müssen wir (bottom-up) die Fehler an allen Knoten schätzen.
- Obendrein sollten wir berücksichtigen, wie genau unsere Schätzung ist. Dazu bestimmen wir ein Konfidenzintervall.
- Wenn die obere Schranke der Konfidenz in den Fehler beim oberen Knoten kleiner ist als bei allen Unterknoten zusammen, werden die Unterknoten gestutzt.

Was ist ein Konfidenzintervall?

Konfidenzintervall

Vorgegeben eine tolerierte Irrtumswahrscheinlichkeit α, gibt das Konfidenzintervall

$$P(u \leq X \leq o) = 1 - \alpha$$

an, dass X mit der Wahrscheinlichkeit $1 - \alpha$ im Intervall [u, o] liegt und mit der Wahrscheinlichkeit α nicht in [u, o] liegt.

Meist wird das Konfidenzintervall für den Erwartungswert gebildet. Beispiel $\alpha = 0,1$: Mit 90% iger Wahrscheinlichkeit liegt der Mittelwert $\bar{X}$ im Intervall [u, o], nur 10% der Beobachtungen liefern einen Wert außerhalb des Intervalls.

z-Transformation in eine standard-normalverteilte Zufallsvariable

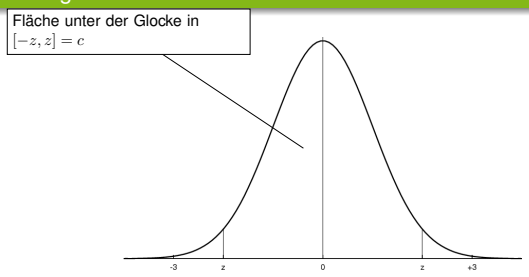
Die Zufallsvariable X wird bezüglich ihres Mittelwerts $\bar{X}$ standardisiert unter der Annahme einer Normalverteilung:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} \sim \mathcal{N}(0; 1)$$

Die Wahrscheinlichkeit dafür, dass der Mittelwert im Intervall liegt, ist nun:

$$P\left(-z\left(1 - \frac{\alpha}{2}\right) \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} \leq z\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha$$

Verteilung mit z-Werten



- $P(-z \leq X \leq z) = 1 - \alpha$ Konfidenzniveau Wahrscheinlichkeit, dass X mit Mittelwert 0 im Intervall der Breite $2z$ liegt ist $1 - \alpha$.
- z kann nachgeschlagen werden (z.B. Bronstein), wobei wegen Symmetrie nur angegeben ist: $P(X \geq z)$

Rechnung für reellwertige Beobachtungen und Mittelwert

Wir wollen ein bestimmtes Konfidenzniveau erreichen, z.B. 0,8.

- $P(X \geq -z)$ $P(X \leq z)$ ist dann $(1 - 0,8)/2 = 0,1$.
- Der z-Wert, für den die Fläche der Glockenkurve zwischen $-z$ und z genau $1 - \alpha = 0,8$ beträgt, ist das $(1 - \frac{\alpha}{2})$-Quantil der Standardnormalverteilung, hier: $1,28$ (nachschlagen).
- Das standardisierte Stichprobenmittel liegt mit der Wahrscheinlichkeit 0,8 zwischen -1,28 und +1,28.

$$\begin{aligned} 0,8 &= P(-1,28 \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}} \leq 1,28) \\ &= P(-1,28\frac{\sigma}{\sqrt{N}} \leq \bar{X} - \mu \leq 1,28\frac{\sigma}{\sqrt{N}}) \\ &= P(\bar{X} - 1,28\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} - 1,28\frac{\sigma}{\sqrt{N}}) \end{aligned}$$

Das Intervall ist $[\bar{X} - 1,28\frac{\sigma}{\sqrt{N}}; \bar{X} + 1,28\frac{\sigma}{\sqrt{N}}]$.

Fehler oder Erfolg schätzen

- Bei den Entscheidungsbäumen beobachten wir nur zwei Werte $Y \in \{+, -\}$.
- Wir haben eine Binomialverteilung mit wahrer Wahrscheinlichkeit p_+ für $y = +$ (Erfolg).
- Beobachtung der Häufigkeit f_+ bei N Versuchen.
  Varianz:
  $$\sigma^2 = \frac{f_+(1 - f_+)}{N}$$
  Erwartungswert:
  $$E(p_+) = f_+/N$$
- In das allgemeine Konfidenzintervall $[\bar{X} - z(1 - \alpha/2)\frac{\sigma}{\sqrt{N}}; \bar{X} + 1,28\frac{\sigma}{\sqrt{N}}]$ setzen wir diese Varianz ein und erhalten:
  $$\left[f_+ - z(1 - \alpha/2)\frac{\sqrt{f_+(1 - f_+)}}{N}; f_+ z(1 - \alpha/2)\frac{\sqrt{f_+(1 - f_+)}}{N}\right]$$

Konfidenz bei Binomialverteilung

Allgemein berechnet man die obere und untere Schranke der Konfidenz bei einer Binomialverteilung für ein Bernoulli-Experiment:

$$p_+ = \frac{f_+ + \frac{z^2}{2N} \pm z\sqrt{\frac{f_+}{N} - \frac{f_+^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}}$$

Hierzu muss lediglich die Häufigkeit f_+ gezählt werden, N, z bekannt sein.
Diese Abschätzung für den Erfolg können wir symmetrisch für den Fehler (p_-) durchführen.

Anwendung zum Stutzen

- Für jeden Knoten nehmen wir die obere Schranke (pessimistisch):
  $$p_- = \frac{f_- + \frac{z^2}{2N} + z\sqrt{\frac{f_-}{N} - \frac{f_-^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}}$$
- Wenn der Schätzfehler eines Knotens kleiner ist als die Kombination der Schätzfehler seiner Unterknoten, werden die Unterknoten weggestutzt. Die Kombination wird gewichtet mit der Anzahl der subsumierten Beispiele.

Gütemaße

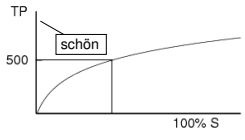
- Konfusionsmatrix:

tatsächlich	Vorhergesagt $+$	Vorhergesagt $-$	
$+$	True positives TP	False negatives FN	Recall: $TP/(TP + FN)$
$-$	False positives FP	True negatives TN	
	Precision: $TP/(TP+FP)$		

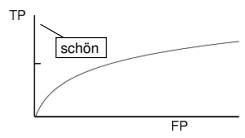
- Accuracy: $P(\hat{f}(x) = y)$ geschätzt als $(TP + TN)/total$

Balance von FP und FN

- F-measure: $\frac{\beta \cdot recall \cdot precision}{recall + precision} = \frac{\beta TP}{\beta TP + FP + FN}$
- Verlaufsformen:
  - Lift: TP für verschiedene Stichprobengrößen S

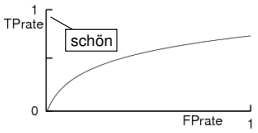


  - Receiver Operating Characteristic (ROC): für verschiedene TP jeweils die FP anzeigen



ROC genauer

- Statt der absoluten Anzahl TP nimm die Raten von true oder false positives – ergibt eine glatte Kurve.
  - Für jeden Prozentsatz von falschen Positiven nimm eine Hypothese h, deren Extension diese Anzahl von FP hat und zähle die TP.
  - $TP_{rate} := TP/P \sim recall$ bezogen auf eine Untermenge
  - $FP_{rate} := FP/N \sim FP/FP + TN$ bezogen auf Untermenge



Kosten von Fehlern

- Nicht immer sind FP so schlimm wie FN
  - medizinische Anwendungen: lieber ein Alarm zu viel als einen zu wenig!
- Gewichtung der Beispiele:
  - Wenn FN 3x so schlimm ist wie FP, dann gewichte negative Beispiele 3x höher als positive.
  - Wenn FP 10x so schlimm ist wie FN, dann gewichte positive Beispiele 10x höher als negative.
- Lerne den Klassifikator mit den gewichteten Beispielen wie üblich. So kann jeder Lerner Kosten berücksichtigen!

Was wissen Sie jetzt?

- Sie kennen den Algorithmus ID3 als Beispiel für TDIDT.
- Für das Lernen verwendet ID3 das Gütemaß des Informationsgewinns auf Basis der Entropie.
- Man kann abschätzen, wie nah das Lernergebnis der unbekannten Wahrheit kommt $\rightarrow$ Konfidenz
- Man kann abschätzen, wie groß der Fehler sein wird und dies zum Stutzen des gelernten Baums nutzen.
- Lernergebnisse werden evaluiert:
  - Einzelwerte: accuracy, precision, recall, F-measure
  - Verläufe: Lift, ROC

Diese Evaluationsmethoden gelten nicht nur für Entscheidungsbäume!