Machine Learning

– in 90 Minuten

Jonas Beste & Arne Bethmann

9. Oktober 2015

Überblick

- Konzepte und Terminologie
- Modellauswahl
- Naive Bayes
- k-nearest-neighbor
- Entscheidungsbäume
- Ensemble Learning
- Weiteres

Was ist Machine Learning

- Machine Learning wird u.a. auch Statistical Learning genannt
- Darunter ist eine umfangreiches Auswahl an Verfahren zur Auswertung von Daten zu verstehen
- Diese umfassen weit mehr als den Ansatz der linearen Regression
- Algorithmen die von Daten lernen können
- Ermöglicht den Umgang mit großen Datenmengen

Grundlegendes

- Verwendung vorrangig für Vorhersagen (*predictions*), aber auch für Schätzungen (*estimations*)
- Unterscheidung zwischen *supervised* und *unsupervised Learning* (Output beobachtet oder nicht)
- Unterscheidung zwischen Regressions- und Klassifikationsproblemen (quantitative und qualitative Variablen)
- Trainings- und Testdaten:
  - Lernen des Klassifikators (Konstruktion des Modells) anhand von Trainingsdaten
  - Interesse an guter Vorhersage bislang unbekannter Fälle (Testdaten)

Genauigkeit und Interpretierbarkeit

- Abwägung zwischen Genauigkeit und Interpretierbarkeit von Methoden
- Manche Methoden sind flexibles (weniger restriktiv) als andere
- Flexible Methoden können sich sehr gut den Daten anpassen und ermöglichen so sehr genaue Vorhersagen
- Restriktive Methoden hingegen sind meist deutlich besser zu interpretieren (bevorzugt bei Inferenz)
- Problem des *Overfitting*: Zu starke Anpassung an die Trainingsdaten, wodurch Testdaten ungenügend vorhergesagt werden
- Daher führen weniger flexible Methoden häufig zu besseren Ergebnissen

Bewertung von Statistical Learning Methoden

- Keine Methode ist die allgemein Beste für alle Fragen und Daten
- Daher kommt der Auswahl der geeigneten Methode eine besondere Bedeutung zu
- Maß für die Qualität: Wie gut passen die Vorhersagen tatsächlich zu den beobachteten Daten?
- Bei regressionsbasierten Vorhersagen wird meistens der *mean squared error* (MSE) herangezogen:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

- Bei Klassifikationen ergibt sich die Fehlerrate aus dem Anteil falsch zugeordneter Fälle:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

Bias-Variance Trade-Off

- Bei Erhöhung der Flexibilität sinkt die Verzerrung und steigt die Varianz
- Auswahl einer Methode die geringe Verzerrung und geringe Varianz aufweist
- Abwägung zwischen Ausmaß an Verzerrung und Varianz notwendig

Resampling Methoden

- Resampling Methoden können verwendet werden um Aussagen über die Eignung von Modellen zu machen
- Technik um zu prüfen, wie sich die Modelle bei unabhängigen Daten verhalten (Overfitting)
- Zu dem Resampling Methoden gehört die Kreuzvalidierung sowie das Bootstrapping
- Kreuzvalidierung kann verwendet werden um die Performance eines Modells zu bewerten (*model assessment*) oder den geeigneten Grad an Flexibilität zu bestimmen (*model selection*)

Kreuzvalidierung

- Validation Set Approach:
  - Unterteilt die Beobachtungen zufällig in einen Trainingsset und einen Validierungsset
  - Das Modell wird am Trainingsset konfiguriert und am Validierungsset getestet
- k-Fold Cross-Validation:
  - Unterteilt die Beobachtungen zufällig in k ungefähr gleich große Teilmengen
  - Verwendet k-1 Teilmengen zur Erstellung des Modells und die verbleibende Teilmenge zur Validierung
  - Berechnet k Modelle, wobei jede Teilmenge genau einmal als Validierungsset dient
  - Die Ergebnisse der einzelnen Modelle können dann kombiniert werden
  - Jede Beobachtung wird für Training und genau einmal für die Validierung verwendet
  - In der Regel wird k = 5 oder k = 10 gewählt

Naive Bayes

- Simpler probabilistischer Klassifikator basierend auf der Anwendung des Bayes Theorem
- Starke (naive) Annahme der Unabhängigkeit zwischen den erklärenden Merkmalen
- Dabei wird jede Beobachtung der Klasse zugeteilt, die bei gegebenen Prediktoren am wahrscheinlichsten ist
- So gesehen ist naive Bayes ein Modell bedingter Wahrscheinlichkeit und kann dargestellt werden als:

$$p(C_k|x) = \frac{p(C_k) \times p(x|C_k)}{p(x)}$$

Naive Bayes – Beispiel

- Zeuge hat eine Person gesehen, konnte aber das Geschlecht der Person nicht erkennen
- Bekannt sind nur Informationen zu Größe und Farbe der Klamotten
- Die Person war kleiner als 170 cm und trug helle Klamotten
- Handelt es sich dabei um eine Frau oder um einen Mann?
- Um die Frage zu beantworten, kann auch eine Reihe von Beobachtungen zurückgegriffen werden

Naive Bayes – Example

Beobachtungen

Fall	Geschlecht	über 170 cm	dunkle Klamotten
1	Frau	nein	ja
2	Frau	nein	nein
3	Mann	ja	ja
4	Frau	ja	nein
5	Mann	nein	nein
6	Mann	ja	ja

Naive Bayes – Beispiel

- Bestimmung der bedingten Wahrscheinlichkeit:

$$p(\text{Frau}|\text{<170cm, hell}) = \frac{p(\text{<170cm}|\text{Frau}) \times p(\text{hell}|\text{Frau}) \times p(Frau)}{p(<170cm) \times p(hell)}$$

$$p(\text{Mann}|\text{<170cm, hell}) = \frac{p(\text{<170cm}|\text{Mann}) \times p(\text{hell}|\text{Mann}) \times p(Mann)}{p(<170cm) \times p(hell)}$$

$$p(\text{Frau}|\text{<170cm, hell}) > p(\text{Mann}|\text{<170cm, hell})$$

- Die Person wird als Frau klassifiziert

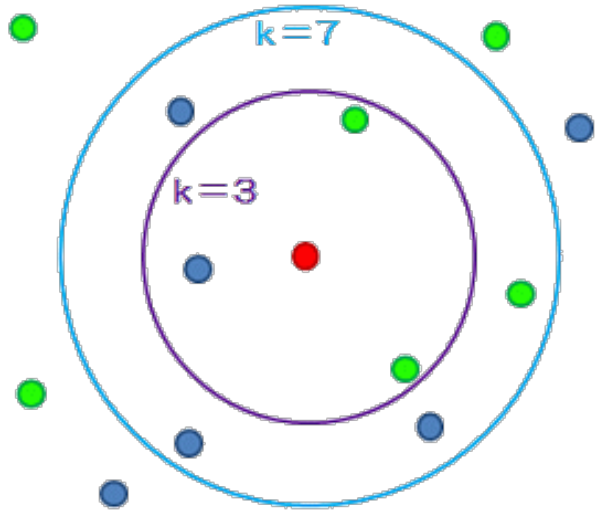
k-nearest-neighbor

- Eine einfache Methode die bei vielen Problem sehr gut funktioniert
- Ein Objekt wird der Klasse zugeteilt, die bei einer bestimmten Anzahl k der nächsten Nachbarn im Merkmalsraum am häufigsten vorkommt
- k ist dabei eine natürliche in der Regel niedrige Zahl
- Die bedingte Wahrscheinlichkeit des Objektes x_0 für die Klasse j ist der Anteil der Nachbar N_0 die dieser Klasse angehören:

$$Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$$

- Ein Schwachpunkt dieses Algorithmus ist seine Sensibilität auf lokale Strukturen in den Daten (Overfitting)

k-nearest-neighbor – Beispiel



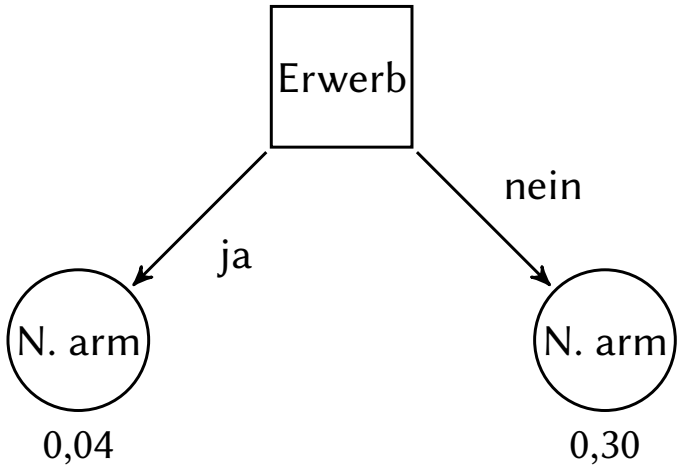
Quelle: http://www.nag-j.co.jp/nagdmc/img/knn.gif

- Roter Punkt ist zu klassifizieren
- k = 3: Rot wird grün klassifiziert
- k = 7: Rot wird blau klassifiziert
- Die Distanz zu den Nachbarn kann als Gewichtung mit aufgenommen werden

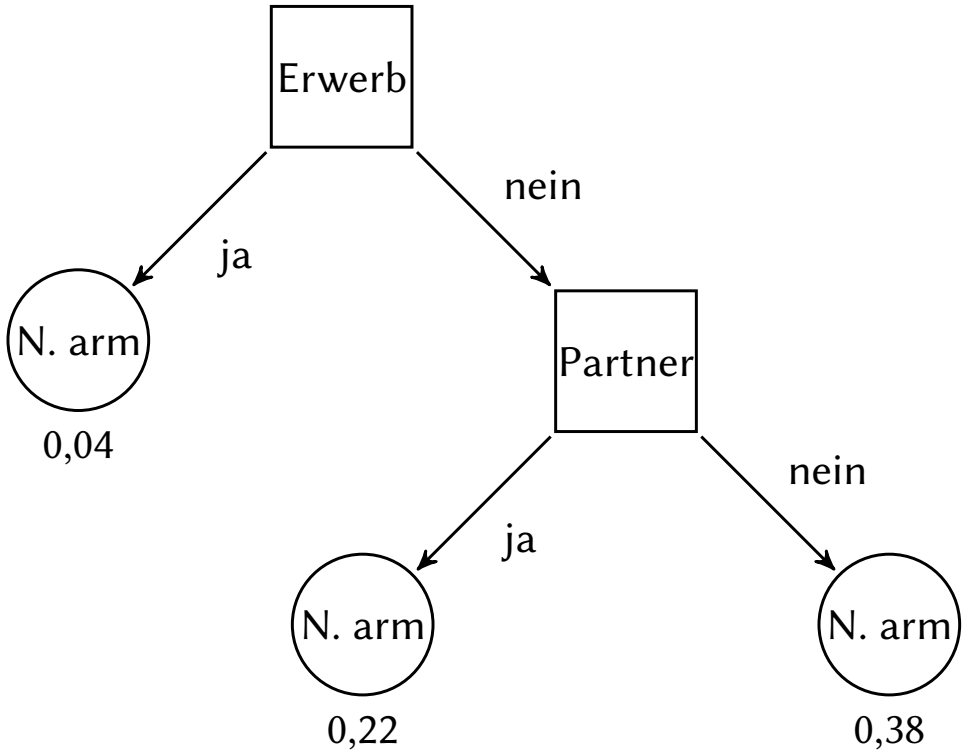
Entscheidungsbäume

- Caret: Classification and regression tree
- Abfolge von binären Entscheidungen, die als Baum mit Blätter, Zweige und Knoten darstellbar sind
- Top-down Induktion: Beginnt an der Wurzel (Startknoten) des Baumes und verzweigt sich mit jedem Schritt
- Die Entscheidungen werden so getroffen, dass die Knoten hinsichtlich der Zielvariable möglichst homogene Klasse darstellen
- In den Blättern (Endknoten) wird eine Zuordnung zu der Klasse getroffen, die dort am häufigsten vorkommt (oder Mittelwert bei Regressionsbäumen)
- *Pruning* eines großen Entscheidungsbaums, um Overfitting zu vermeiden
- Entscheidungsbäume sind gut interpretierbar, aber nicht sehr akkurat

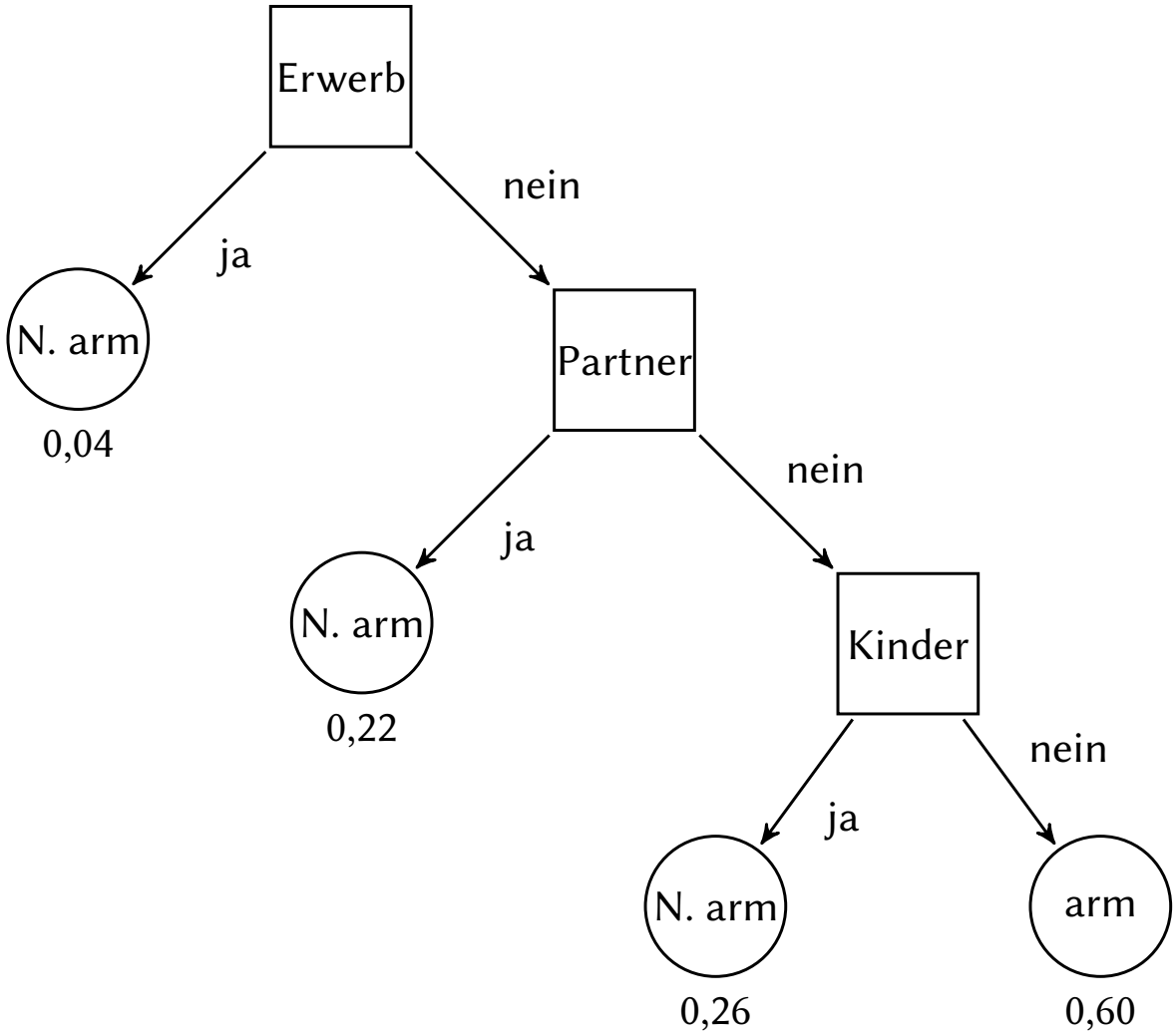
Entscheidungsbäume – Beispiel



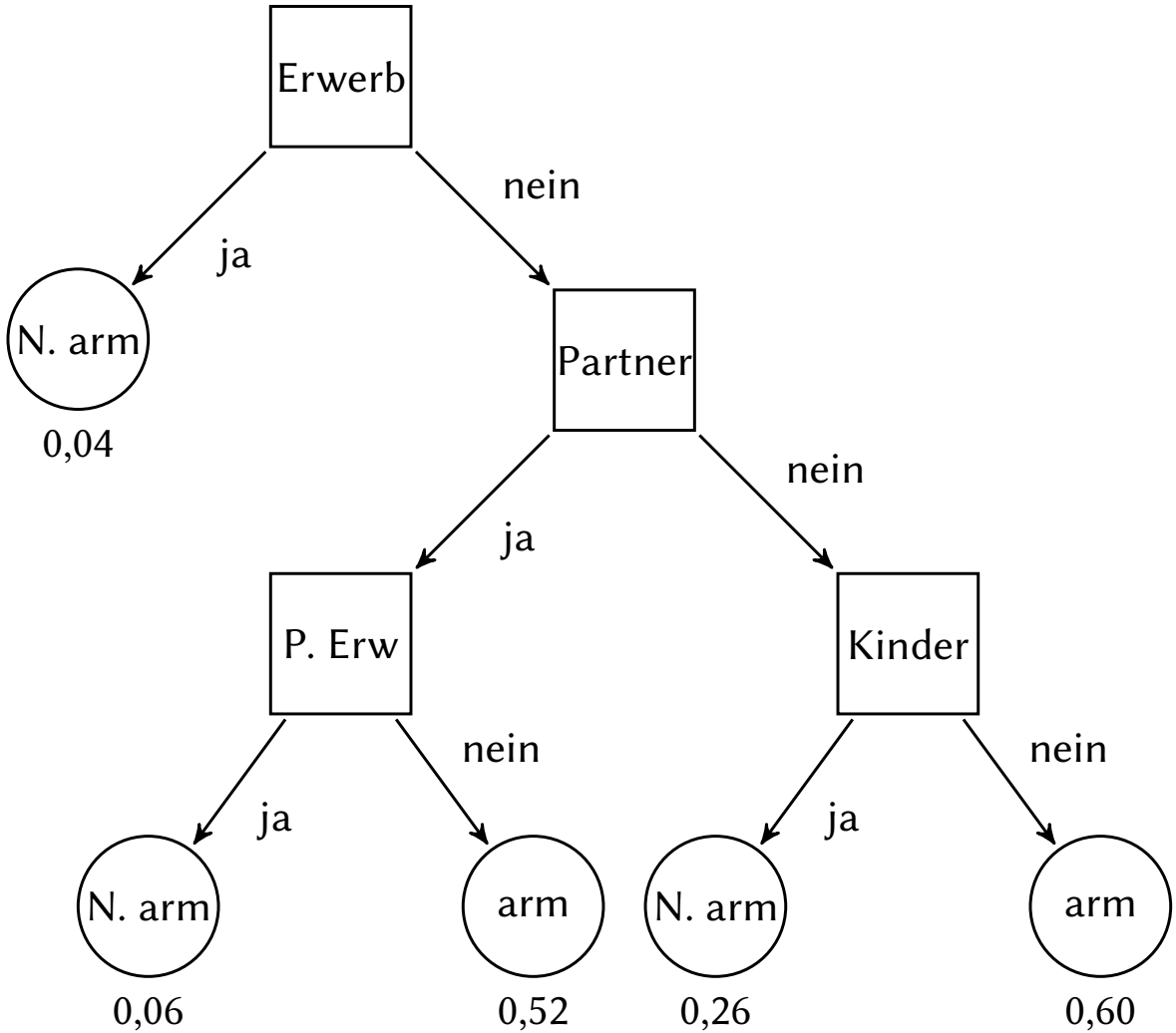
Entscheidungsbäume – Beispiel



Entscheidungsbäume – Beispiel



Entscheidungsbäume – Beispiel



Bagging

- Bootstrap aggregation (Bagging) ist ein Verfahren um die Varianz von statistischen Modellen zu verringern und so aussagekräftigere Vorhersagemodelle zu erhöhen
- Hintergrund: Reduktion der Varianz durch Mittelwertbildung über mehrere Beobachtungen
- Ziehung von B Trainingsdatensets aus dem ursprünglichen Trainingsdatenset (bootstrap)
- Bildung separater Modelle $\hat{f}_b^*(x)$ mit den einzelnen Trainingssets
- Mittelwertbildung der resultierenden Vorhersagen:

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} (\hat{f}_b^*(x))$$

- Besonders effizient bei Verfahren mit hoher Varianz und geringem Bias (wie z.B. Entscheidungsbäumen)

Random Forest

- Ähnliches Verfahren wie Bagging für Entscheidungsbäume, nur dass bei jedem einzelnen Baum nur eine zufällige Auswahl der erklärenden Variablen berücksichtigt wird
- In der Regel wird Wurzel aller Predictoren
- Verringert die Korrelation der Vorhersagen zwischen den einzelnen Bäumen
- Hohe Korrelation führt zu geringerer Effizienz der Aggregation

Boosting

- Weiteres Verfahren um die aussagekräftigere von Vorhersagemodellen zu erhöhen
- Kombination einzelnen für sich genommen schwachen Klassifikatoren
- Reduktion von Bias und Varianz
- Während beim Bagging die einzelnen Modelle unabhängig voneinander geschätzt werden, bauen diese beim Boosting aufeinander auf
- Erfolgt über Gewichtung der Daten
- Sehr effizient für Entscheidungsbäume

Weitere Techniken

- Weitere Klassifikationsmethoden (Logit, Probit und Diskriminanzanalyse)
- Modellauswahl (Ridge Regression und Lasso)
- Nichtlineare Modelle (Splines und additive Modelle)
- Support Vector Machines
- Neuronale Netze
- Unsupervised Learning (Hauptkomponentenanalyse und Clustern)
- ROC-Kurven, Sensitivität und Spezifität
- Weitere Ensemble Learning Techniken (Stacking)