Vorlesung Maschinelles Lernen

Überblick

Katharina Morik

LS 8 Künstliche Intelligenz Fakultät für Informatik
Technische Universität Dortmund

11.10.2008

Gliederung

Bekannte Anwendungen

- Google ordnet die Suchergebnisse nach der Anzahl der auf sie verweisenden Hyperlinks an.
- Amazon empfiehlt einem Kunden, der A gekauft hat, das Produkt B, weil alle (viele) Kunden, die A kauften, auch B kauften.
- Die Post sortiert handbeschriftete Briefe per Schrifterkennung.
- Firmen ordnen ihre eingehende Post automatisch der zuständigen Abteilung zu.
- Aktienkurse oder Verkaufszahlen werden vorhergesagt.

Interesse an Anwendungen

- Business Reporting soll automatisiert werden. On-line Analytical Processing beantwortet nur einfache Fragen. Zusätzlich sollen Vorhersagen getroffen werden.
- Wissenschaftliche Daten sind so umfangreich, dass Menschen sie nicht mehr analysieren können, um Gesetzmäßigkeiten zu entdecken.
- Geräte sollen besser gesteuert werden, indem aus den log-Dateien gelernt wird.
- Roboter sollen sich besser an menschliche Umgebung und Kommunikation anpassen.
- Das Internet soll nicht nur gesamte Dokumente liefern, sondern Fragen beantworten.
- Multimedia-Daten sollen personalisiert strukturiert und gezielter zugreifbar sein.

Was ist Lernen beim Menschen?

Menschen lernen durch:

- Auswendig lernen.
- Einüben. (Fertigkeiten)
- Logisch schließen:
  - Alle Menschen sind sterblich.
    Sokrates ist ein Mensch.
    Sokrates ist sterblich. (Deduktion)
  - Sokrates, Uta, Udo, Veronika, Volker, ... sind Menschen.
    Sokrates, Uta, Udo, Veronika, Volker, ... sind sterblich.
    Alle Menschen sind sterblich. (Induktion)
- Begriffe bilden.
- Grammatiken lernen.
- Gesetze entdecken.
- Theorien entwickeln. (Wissen)

Begriffsbildung

- Eins von diesen Dingen gehört nicht zu den anderen!

Clustering
Kategorisierung

- Alle Beobachtungen, die sich ähneln, werden zu einer Gruppe zusammengefasst.
- Auf diese Weise strukturiert man die vielen Beobachtungen.
- Von den vielen Merkmalen wählt man zur Ähnlichkeitsbestimmung eine möglichst kleine Anzahl aus.
- Die ausgewählten Merkmale sind immer erkennbar (operational).

Die Kategorisierung ordnet jede Beobachtung mindestens einer Gruppe zu. Die Gruppen können sich überlappen.

Menschen kategorisieren immer, ob sie wollen oder nicht! Es ist ein unbewusster kognitiver Prozess.

Clustering
Kategorisierung

- Alle Beobachtungen, die sich ähneln, werden zu einer Gruppe zusammengefasst.
- Auf diese Weise strukturiert man die vielen Beobachtungen.
- Von den vielen Merkmalen wählt man zur Ähnlichkeitsbestimmung eine möglichst kleine Anzahl aus.
- Die ausgewählten Merkmale sind immer erkennbar (operational).

Die Kategorisierung ordnet jede Beobachtung mindestens einer Gruppe zu. Die Gruppen können sich überlappen. Menschen kategorisieren immer, ob sie wollen oder nicht! Es ist ein unbewusster kognitiver Prozess.

Einige Gründe für die Kategorisierung

- Handlungen können nicht auf der Gesamtheit der Beobachtungen ausgeführt werden. Menschen haben eine beschränkte Wahrnehmungs- und Aktionskapazität.
  - Menschen können nur 5-7 kognitive Objekte gleichzeitig beachten (ansehen, hören, merken).
  - Hände können nur eine begrenzte Anzahl physikalischer Objekte fassen.
  - Deshalb muss eine große Grundgesamtheit für Menschen in kleine, wahrnehmbare, handhabbare Untermengen aufgeteilt werden.
- Es gibt schon ein Wort dafür.
  - Jemand nennt ein Objekt x *Tasse*.
  - Alle Objekte, die von jemandem als *Tasse* bezeichnet wurden, gehören in eine Gruppe mit dem Titel *Tasse*.

Positive Beispiele

- Dies sind Tassen.

Negative Beispiele

- Dies sind keine Tassen.

Klassifikation

- Eine Funktion ordnet einer Wahrnehmung eine Klasse zu.
  - Dem Wort *Tasse* entspricht eine Erkennungsfunktion, die jeder Wahrnehmung die Klasse *Tasse* oder *Nicht-Tasse* zuordnet.
- Die einfachste Funktion ist das Aufzählen. Dies begrenzt aber die Klassifikation auf bereits gesehene Objekte.
- Als Wissenschaftler verwenden Menschen gern numerische Funktionen.
- Besonders verständlich sind logische Funktionen. Dies sind meist Definitionen.

Definitionen

Eine Definition ist eine Erkennungs- und Ergänzungsfunktion (hinreichende und notwendige Bedingungen).

Definition: Eine Tasse ist ein Behälter mit flachem Boden und einem Henkel an der Seite.

Erkennungsfunktion: Aha, konkav und undurchlässig, flacher Boden, Henkel an der Seite – eine Tasse!

$$konkav(x), opak(x), hatBoden(x, y), flach(y), hatHenkel(x, z) \rightarrow tasse(x)$$

Ergänzungsfunktion: Kann ich eine Tasse hinstellen? – Ja, denn eine Tasse hat einen flachen Boden und Objekte mit flachem Boden stehen sicher!

$$tasse(x) \rightarrow kannStehen(x)$$

Ein Begriff erleichtert oft die Definition anderer Begriffe.

- Wer nicht weiß, was ein *Boden* oder ein *Henkel* ist, hat Probleme, eine *Tasse* zu definieren.
- Die Definition für *Boden* und *Henkel*
  $\ldots \rightarrow hatBoden(x, y)$
  $\ldots \rightarrow hatHenkel(x, z)$
  erlaubt die Definition von *Tasse*:
  $konkav(x), opak(x), hatBoden(x, y), flach(y),$
  $hatHenkel(x, z) \rightarrow tasse(x)$

Menschliches Lernen

- Die kognitive Psychologie untersucht das menschliche Lernen.
- Die Entwicklungspsychologie untersucht das Lernen über die Alterstufen hinweg [4].
- Einflüsse auf das Lernen werden untersucht:
  - Reihenfolge der Beobachtungen oder Lernschritte [3]
  - Umgebung beim Lernen [1]
  - Soziale Zusammenarbeit (kollaboratives Lernen) [2]
  - ...

Literatur zu menschlichem Lernen


J. Bliss, R. Saljo, and P. Light, editors.
Learning Sites – Social and technological Resources for Learning.


P. Dillenbourg, editor.
Collaborative Learning – Cognitive and Computational Approaches. Pergamon Press, 1998.


Frank E. Ritter, Erno Lehtinen, Josef Nerb, and Timothy O'Shea, editors.
In Order to Learn – How the Sequence of Topics Influences Learning. Oxford University Press, 2007.


R.S. Siegler.
Children's Thinking. Prentice-Hall, 2nd edition, 1991.

Maschinelles Lernen – generische Aufgabe

Population: Eine Menge von Objekten, um die es geht.

Merkmale: Eine Menge von Merkmalen (quantitativ oder qualitativ) beschreibt die Objekte.

Ausgabe: Ein quantitativer Wert (Messwert) oder ein qualitativer (label, z.B. *Tasse*) gehört zu jeder Beobachtung.

Ein Lernverfahren findet eine Funktion, die Objekten einen Ausgabewert zuordnet. Oft minimiert die Funktion einen Fehler.

Modell: Das Lernergebnis (die gelernte Funktion) wird auch als *Modell* bezeichnet.

Notation

ExampleSet

Meta Data View | Data View | Plot View

ExampleSet (14 examples, 1 special attribute, 4 regular attributes)

row no.	Play	Outlook	Temperat...	Humidity	Wind
1	no	sunny	85	85	false
2	no	sunny	80	90	true
3	yes	overcast	83	78	false
4	yes	rain	70	96	false
5	yes	rain	68	80	false
6	no	rain	65	70	true
7	yes	overcast	64	65	true
8	no	sunny	72	95	false
9	yes	sunny	69	70	false
10	yes	rain	75	80	false
11	yes	sunny	75	70	true
12	yes	overcast	72	90	true
13	yes	overcast	81	75	false
14	no	rain	71	80	true

- Der Raum möglicher Beobachtungen wird als p-dimensionale Zufallsvariable X geschrieben.
- Jede Dimension der Beobachtungen wird als X_i notiert (Merkmal).
- Die einzelnen Beobachtungen werden als $\vec{x_1}, ..., \vec{x_N}$ notiert.
- Die Zufallsvariable Y ist die Ausgabe (label).
- N Beobachtungen von Vektoren mit p Komponenten ergeben also eine $N \times p$-Matrix.

Lernaufgabe Clustering

Gegeben

- eine Menge $\mathcal{T} = \{\vec{x_1}, ..., \vec{x_N}\} \subset X$ von Beobachtungen,
- eine Anzahl K zu findender Gruppen $C_1, ..., C_K$,
- eine Abstandsfunktion $d(\vec{x}, \vec{x'})$ und
- eine Qualitätsfunktion.

Finde

- Gruppen $C_1, ..., C_K$, so dass
- alle $\vec{x} \in X$ einer Gruppe zugeordnet sind und
- die Qualitätsfunktion optimiert wird: Der Abstand zwischen Beobachtungen der selben Gruppe soll minimal sein; der Abstand zwischen den Gruppen soll maximal sein.

Lernaufgabe Klassifikation

Gegeben

- Klassen Y, oft $y \in \{+1, -1\}$,
- eine Menge $\mathcal{T} = \{(\vec{x_1}, y_1), ..., (\vec{x_N}, y_N)\} \subset X \times Y$ von Beispielen,
- eine Qualitätsfunktion.

Finde

- eine Funktion $f : X \to Y$, die die Qualitätsfunktion optimiert.

Lernaufgabe Regression

Gegeben

- Zielwerte Y mit Werten $y \in \mathcal{R}$,
- eine Menge $\mathcal{T} = \{(\vec{x_1}, y_1), ..., (\vec{x_N}, y_N)\} \subset X \times Y$ von Beispielen,
- eine Qualitätsfunktion.

Finde

- eine Funktion $f : X \to Y$, die die Qualitätsfunktion optimiert.

Funktionsapproximation

Wir schätzen die wahre, den Beispielen unterliegende Funktion. Gegeben

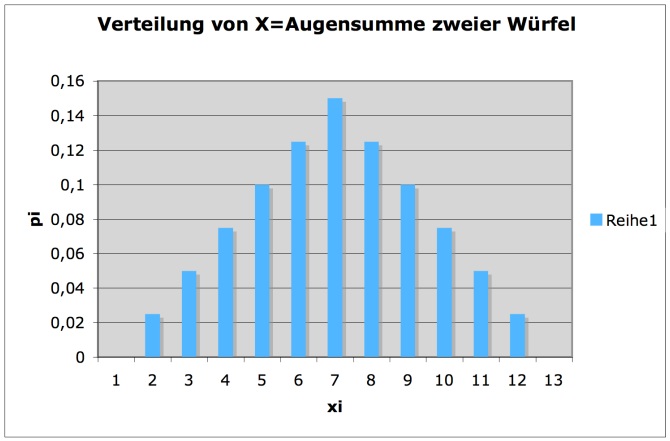
- eine Menge von Beispielen $\mathcal{T} = \{(\vec{x_1}, y_1), ..., (\vec{x_N}, y_N)\} \subset X \times Y$,
- eine Klasse zulässiger Funktionen f_θ (Hypothesensprache),
- eine Qualitätsfunktion,
- eine feste, unbekannte Wahrscheinlichkeitsverteilung $P(X)$.

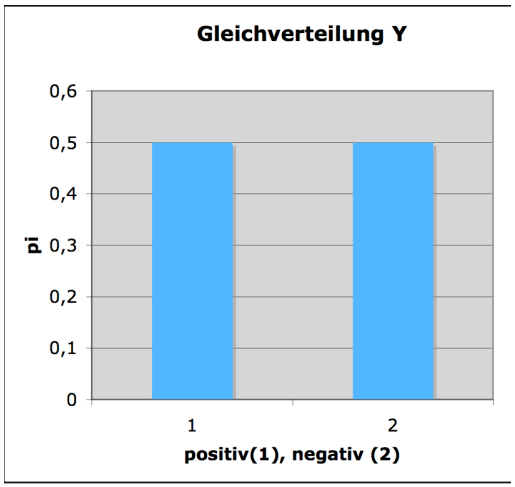
Finde

- eine Funktion $f_\theta : X \to Y$, die die Qualitätsfunktion optimiert.

Zur Erinnerung: Verteilung

Eine Zufallsvariable X heißt *diskret*, wenn sie nur endlich oder abzählbar unendlich viele Werte $x_1, ..., x_m$ annehmen kann. Zu jedem Wert gehört ein Ereignis, das mit der Wahrscheinlichkeit $P(X = x_i)$ eintreten kann. Die Realisationen x_i gemeinsam mit den zugehörigen Wahrscheinlichkeiten heißen (Wahrscheinlichkeits-)Verteilung von X.





Verteilungsfunktion

Sei X eine diskrete oder stetige Zufallsvariable. Die Funktion
$D(x) = P(X \leq x), x \in \mathcal{R}$
heißt Verteilungsfunktion von X.
Bei diskreten Zufallsvariablen gilt: $D(x) = \sum_{i:x_i \leq x} p_i$
Eine Zufallsvariable heißt stetige Zufallsvariable, wenn ihre Verteilungsfunktion stetig ist.

Dichtefunktion

Die Ableitung $D'(x)$ wird Dichtefunktion genannt. Umgekehrt erhält man die Verteilungsfunktion durch Integration der Dichtefunktion: $D(x) = \int_{-\infty}^{x} h(t)dt$
Funktionen, die eine Dichte haben, sind absolut stetig.
Die Gesamtfläche unter dem Graphen von h ist gleich 1.

Wenn wir die Verteilung kennen, können wir eine gute Prognose machen!

- Wenn wir wissen, dass $p_i = 0,01$ ist, dann ist es nicht so schlimm, wenn wir uns bei x_i irren – wir irren uns dann selten.
- Wenn wir wissen, dass $P(Y = +1) = 0,99$ ist, dann sagen wir immer +1 voraus und sind in 99% der Fälle richtig. Wir haben nur ein Risiko von 1%, uns zu irren.

Qualitätsfunktion – Fehlerfunktion

Fehlerrisiko:

$$R(Y, f(X)) = \sum_{i=1}^{N} Q(y_i, \vec{x_i}) p(\vec{x_i}) \quad (1)$$

wobei $p(\vec{x_i})$ die Wahrscheinlichkeit ist, dass das Beispiel $\vec{x_i}$ aus X gezogen wird.

Mittlerer Quadratischer Fehler:

$$MSE(Y, f(X)) = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(\vec{x_i}))^2 \quad (2)$$

Mittlerer 0-1-Verlust: $Q(Y, f(X)) = \frac{1}{N} \sum_{i=1}^{N} Q(\vec{x_i}, f)$, wobei

$$Q(y_i, f(\vec{x_i})) = \begin{cases} 0, falls & f(\vec{x_i}) = y \\ 1, falls & f(\vec{x_i}) \neq y \end{cases}$$

Problem

- Wir haben nur eine endliche Menge von Beispielen. Alle Funktionen, deren Werte durch die Beispiele verlaufen, haben einen kleinen Fehler.
- Wir wollen aber für alle Beobachtungen das richtige y voraussagen. Dann sind nicht mehr alle Funktionen, die auf die Beispiele gepasst haben, gut.
- Wir kennen nicht die wahre Verteilung der Beispiele.
- Wie beurteilen wir da die Qualität unseres Lernergebnisses?

Lern- und Testmenge

Wir teilen die Daten, die wir haben, auf:

Lernmenge: Einen Teil der Daten übergeben wir unserem Lernalgorithmus. Daraus lernt er seine Funktion $f(x) = \hat{y}$.

Testmenge: Bei den restlichen Daten vergleichen wir $\hat{y}$ mit y.

Aufteilung in Lern- und Testmenge

- Vielleicht haben wir zufällig aus lauter Ausnahmen gelernt und testen dann an den normalen Fällen. Um das zu vermeiden, verändern wir die Aufteilung mehrfach.

  leave-one-out: Der Algorithmus lernt aus $N-1$ Beispielen und testet auf dem ausgelassenen. Dies wird N mal gemacht, die Fehler addiert.

- Aus Zeitgründen wollen wir den Algorithmus nicht zu oft anwenden.

  Kreuzvalidierung: Die Lernmenge wird zufällig in n Mengen aufgeteilt. Der Algorithmus lernt aus $n-1$ Mengen und testet auf der ausgelassenen Menge. Dies wird n mal gemacht.

Kreuzvalidierung

- Man teile alle verfügbaren Beispiele in n Mengen auf. z.B. $n = 10$.
- Für i=1 bis i=n:
  - Wähle die i-te Menge als Testmenge,
  - die restlichen $n - 1$ Mengen als Lernmenge.
  - Messe die Qualität auf der Testmenge.
- Bilde das Mittel der gemessenen Qualität über allen n Lernläufen. Das Ergebnis gibt die Qualität des Lernergebnisses an.

Fragestellungen des maschinellen Lernens

- Welche Zusicherungen kann ich meinen Kunden geben? (Fehlerschranken)
- Wieviele Beispiele brauche ich?
- Welche Eigenschaften sollen die Beispiele haben, um gut vorherzusagen und wie finde (erzeuge) ich sie?
- Welche Modellklasse soll ich wählen?
- Welcher Algorithmus wird mit vielen Beispielen und vielen Dimensionen in kurzer Zeit fertig?

Was wissen Sie jetzt?

- Sie haben Clustering (Kategorisierung) und Klassifikation als menschliches Lernen gesehen.
- Die Lernaufgaben Clustering, Klassifikation, Regression haben Sie auch als Aufgaben des maschinellen Lernens gesehen.
- Sie wissen, was die Kreuzvalidierung ist.

Was wissen Sie noch nicht?

- Es gibt viele verschiedene Modellklassen. Damit werden die Lernaufgaben spezialisiert.
- Es gibt unterschiedliche Qualitätsfunktionen. Damit werden die Lernaufgaben als Optimierungsaufgaben definiert.
- Die Algorithmen zur Lösung der Lernaufgaben werden Sie in der Vorlesung kennenlernen und ihre Kernmethoden in den Übungen selbst implementieren.

Themen

- lineare Modelle und k neirest Neighbor und das Problem von *bias* und *variance*
- Stützvektormethode (SVM) und strukturelle Risikominimierung mit verschiedenen Algorithmen zur Lösung des Optimierungsproblems
  - Klassifikation von Texten
- Entscheidungsbäume
- Merkmalsselektion
- Graphische Modelle
  - Informationsextraktion aus Texten
- K-Means Clustering
- Tag Clustering
  - Clustering von Texten nach ihren Annotationen

Grundidee der Vorlesung

Die Vorlesung behandelt die Themen unter drei Aspekten:

- Theorie: abstrakte Darstellung der Lernaufgabe, ihrer Annahmen, Eigenschaften. Dies gründet sich auf die statistische Lerntheorie [2]. Als Mathe-Buch kann man dazu verwenden [3] und [1].
- Algorithmik: wie löst man nun also die Lernaufgabe? Verschiedene Algorithmen am Beispiel der SVM
- Praxis: Realistische Anwendungen werden bearbeitet, vom Datensatz zum Report.

Übungen

Wir verwenden das System RapidMiner und können damit

- (fast) alle Lernverfahren und Transformationen der Daten durchführen
- den Kern bestimmter Lernverfahren selbst implementieren und in der RapidMiner-Umgebung ablaufen lassen.

Wir gehen die Fragestellungen bei Anwendungen durch und Sie erwerben *know-how*, das nicht in Büchern steht.

Wofür bekommen Sie einen Schein?

- Kommen Sie in jede Vorlesung – dann können Sie auch das Tempo bestimmen und Fragen stellen.
- Gehen Sie in die Übungsgruppe!
- Lösen Sie jede Übungsaufgabe: Werden 80% der Punkte erreicht, bekommt man einen Schein.
- Nutzen Sie die Vorlesung/Übung zur Vorbereitung auf eine Fachprüfung!

Wir sehen uns...

In der ersten Übung wird RapidMiner vorgestellt. Sie findet statt:
Am Donnerstag 13.10.2011
um 14 Uhr
in GB IV Raum 113
Wenn Sie einen Laptop haben, bringen Sie ihn mit. Sie können auch schon bei
http://rapid-i.com/
kostenlos RapidMiner herunterladen.

Literatur


Gerald Farin and Dianne Hansford.
Lineare Algebra – Ein geometrischer Zugang.
Springer, 2003.


Trevor Hastie, Robert Tibshirani, and Jerome Friedman.
The Elements of Statistical Learning: Data Mining, Inference, and Prediction.
Springer series in statistics. Springer, New York, USA, 2001.


Gerald Teschl and Susanne Teschl.
Mathematik für Informatiker.
Springer, 2006.