Gliederung

- Motivation für Evaluation
- Schätzen des Klassifikationsfehlers
  - Trainieren, Validieren und Testen
  - Fehler eingrenzen, Vertrauensintervalle
  - Aufteilung in Trainings- und Testmenge
    - Wiederholtes Aufteilen
    - Kreuz-Validierung
    - Leave-One-Out
    - Bootstrap
- Performanzvergleich von DM Methoden
  - Einbeziehen der Varianz der Performanz: Signifikanz Tests
  - Gepaarter und ungepaarter t-Test
- Performanz bei der Vorhersage von Wahrscheinlichkeiten
  - Quadratische- und Informationsverlust-Funktion
- Performanzvergleich bei verschiedenen Kosten der Fehler
  - Lift Charts
  - ROC Kurve
  - Precision und Recall
- Evaluierung nummerischer Vorhersagen
  - verschiedene Maße
- Das MDL Prinzip
  - Modellauswahl
  - Bayes Theorem, Log-Likelyhood und MDL
  - Epikurus Prinzip und Modellmittelung
  - MDL und Clusteranalyse

Die Kosten mitzählen

- Verschiedene Typen von Klassifikationsfehlern haben oft auch verschiedene Kosten
- Beispiel:
  - Entdecken von Terroristen
    - "Kein Terrorist" 99.99% der Fälle korrekt
    - Ein nicht entdeckter Terrorist verursacht sehr viel mehr Kosten, als ein als Terrorist beschuldigter Tourist
  - Öl-Teppiche finden
  - Fehlerdiagnose
  - Postwurfsendungen

Die Kosten mitzählen

- *Confusion Matrix*:

		Predicted class	
		Yes	No
Actual class	Yes	True positive	False negative
	No	False positive	True negative

- Es gibt noch viele andere Arten von Kosten!
  - z.B.: Kosten um Trainingsdaten zu sammeln

Lift Charts

- In der Praxis sind die Kosten oft nicht bekannt
- Entscheidungen werden durch Vergleich von Szenarien getroffen
- Beispiel: Postwurfsendungen zu 1,000,000 Haushalten
  - sende an alle; 0.1% Antworten (1000)
  - DM identifiziert Teilmenge von 100,000, 0.4% von diesen antworten (400)
    40% der Antworten für 10% der Kosten
  - oder DM identifiziert Teilmenge von 400,000, 0.2% antworten (800)
- *lift chart* gibt einen visuellen Vergleich

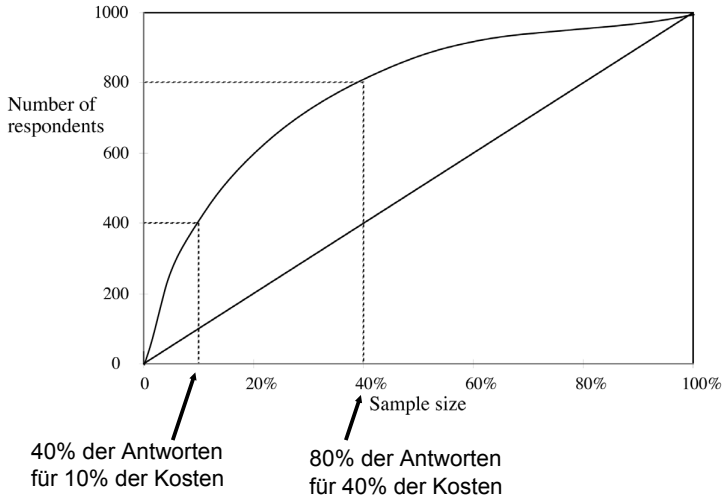
Erzeugung des Lift charts

- Aufgabe
  - Finde Teilmenge der Testinstanzen, mit überproportialer Anzahl an Positiven.
- Idee
  - Klassifikator sagt Antwort-Whr. vorher
  - Sortiere Instanzen absteigend nach der vorhergesagten Whr. positiv zu sein
  - Top-i sind beste Wahl, wenn Klasse unbekannt ist

	Predicted probability	Actual class
1	0.95	Yes
2	0.93	Yes
3	0.93	No
4	0.88	Yes
…	…	…

- *x* Achse ist Stichprobengröße
  y Achse ist Anzahl der wahren Positiven

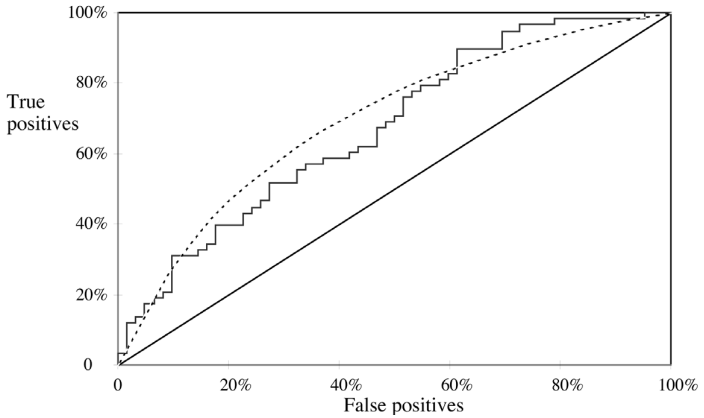
Ein hypothetischer Lift Chart



ROC Kurven

- *ROC Kurve* sind ähnlich zu Lift Charts
  - ROC … "receiver operating characteristic"
  - Wird in der Signalverarbeitung genutzt, um den Kompromiss zwischen Treffer- und Fehlalarmrate zu zeigen
- Unterschiede zu Lift Charts:
  - *y* Achse zeigt Prozente der wahren Positiven der Stichprobe *im Vergleich zu einer absoluten Anzahl*
  - *x* Achse zeigt Prozente der falschen Positiven der Stichprobe *im Vergleich zur Stichprobengröße*

Beispiel für ROC Kurve

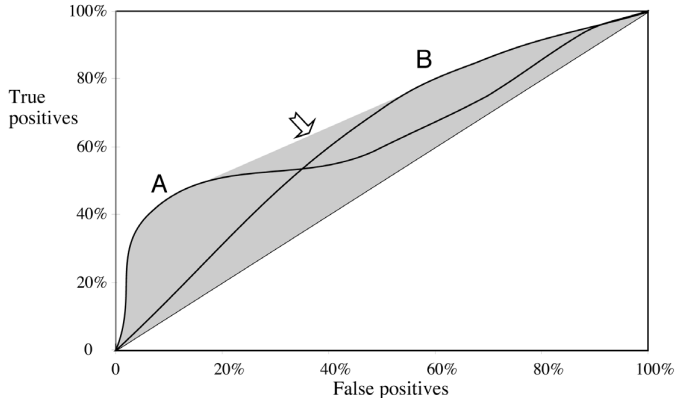


- Treppenstufen Kurve — eine Testdatenmenge
- Glatte Kurve — mit Kreuzvalidierung

Kreuzvalidierung und ROC Kurven

- Einfache Methode um eine ROC Kurve aus der Kreuzvalidation zu erhalten:
  - Berechne Whr. für Instanzen in Testmenge
  - Sammle Whr.s für alle Testmengen
  - Sortiere Instanzen aller Testmengen nach Whr.
- Diese Methode ist implementiert in WEKA
- Andere Möglichkeit
  - Mittele die ROC Kurven für jede Testmenge

ROC Kurven für zwei Schemata



- Für eine kleine, fokusierte Menge, nutze Methode A
- Für größere Mengen, nutze Methode B
- Für mittlere Mengen, nutze Methoden A und B mit passenden Whr.

Konvexe Hülle

- Für zwei gegebene Lernschemata kann jeder Punkt auf der konvexen Hülle erreicht werden!
- TP und FP Raten für Schema 1: t_1 und f_1
- TP und FP Raten für Schema 2: t_2 und f_2
- Falls Schema 1 die ersten $100 \times q$ % Fälle vorhersagen soll und Schema 2 den Rest, dann
  - TP Rate für kombiniertes Schema: $q \times t_1 + (1-q) \times t_2$
  - FP Rate für kombiniertes Schema: $q \times f_2 + (1-q) \times f_2$

Kosten-orientiertes Lernen

- Meisten Lernschemata sind nicht Kosten-orientiert
  - Generieren den gleichen Klassifikator unabhängig von den Kosten für verschiedene Klassen
  - Beispiel: Standard Entscheidungsbaum Algorithmus
- Einfache Methode für kosten-orientiertes Lernen:
  - Vervielfältige Instanzen im Verhältnis zu den Kosten
  - Gewichte Instanzen im Verhältnis zu den Kosten
- Einige Schemata können Kosten über Parameter berücksichtigen, z.B. naïve Bayes

Maße in Information Retrieval

- Prozentsatz der zurückgegebenen Dokumente die relevant sind: *precision*=TP/(TP+FP), (Präzision)
- Prozentsatz der relevanten Dokumente, die zurückgegeben werden: *recall* =TP/(TP+FN), (Ausbeute)
- Zusammengefaßte Maße: Durchschnittliche Precision für 20%, 50% und 80% Recall (Drei-Punkt Recall-Durchschnitt)
- *F-Maß*=(2×recall×precision)/(recall+precision)

Zusammenfassung der Maße

	Domäne	Kurve	Erklärung
Lift chart	**Marketing**	**TP** **Teilmengen-größe**	**TP** **(TP+FP)/(TP+FP+TN+FN)**
ROC Kurve	**Kommunikation**	**TP rate** **FP rate**	**TP/(TP+FN)** **FP/(FP+TN)**
Recall-Precision Kurve	**Information Retrieval, Suche**	**Recall** **Precision**	**TP/(TP+FN)** **TP/(TP+FP)**

Gliederung

- Motivation für Evaluation
- Schätzen des Klassifikationsfehlers
  - Trainieren Validieren und Testen
  - Fehler eingrenzen, Vertrauensintervalle
  - Aufteilung in Trainings und Testmenge
    - Wiederholtes Aufteilen
    - Kreuz-Validierung
    - Leave-One-Out
    - Bootstrap
- Performanzvergleich von DM Methoden
  - Einbeziehen der Varianz der Performanz: Signifikanz Tests
  - Gepaarter und Ungepaarter t-Test
- Performanz bei der Vorhersage von Wahrscheinlichkeiten
  - Quadratische und Informationsverlust Funktion
- Performanzvergleich bei verschiedenen Kosten der Fehler
  - Lift Charts
  - ROC Kurve
  - Precision und Recall
- Evaluierung nummerischer Vorhersagen
  - verschiedene Maße
- Das MDL Prinzip
  - Modellauswahl
  - Bayes Theorem, Log-Likelyhood und MDL
  - Epikurus Prinzip und Modellmittelung
  - MDL und Clusteranalyse

Evaluierung nummerischer Vorhersagen

- Gleichen Strategien: unabhängige Testmenge, Kreuz-Validierung, Signifikanz Tests, usw..
- Unterschiede: Fehlermaße
- Zielwerte: $a_1\ a_2\ \ldots a_n$
- Vorhergesagte Werte: $p_1\ p_2\ \ldots\ p_n$
- Oft gnutztes Maß: *gemittelter, quadrierter Fehler* (mean-squared error)

$$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}$$

Andere Maße

- *Root Mean-Squared Error (RMSE)*:

$$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}}$$

- Der *Mean Absolute Error* ist weniger empfindlich gegenüber Ausreißern als der mean-squared error:

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{n}$$

- In manchen Situatitionen ist der *relative* Fehler passender, z.B. 10% als Fehler, wenn 50 statt 500 vorhergesagt wurde

Verbesserung gegenüber dem Durchschnitt

- Um wieviel ist das Schema besser als wenn einfach immer der Durchschnitt vorhergesagt werden würde?
- Der *relative quadrierte Fehler* ist ($\bar{a}$ ist Durchschnitt):

$$\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{(\bar{a} - a_1)^2 + \ldots + (\bar{a} - a_n)^2}$$

- Der *relative absolute Fehler* ist:

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{|\bar{a} - a_1| + \ldots + |\bar{a} - a_n|}$$

Korrelationskoeffizient

- Mißt die *statistische Korrelation* zwischen den vorhergesagten und tatsächlichen Werten

$$\frac{S_{PA}}{\sqrt{S_P S_A}}$$

$$S_{PA} = \frac{\sum_i (p_i - \bar{p})(a_i - \bar{a})}{n-1} \qquad S_P = \frac{\sum_i (p_i - \bar{p})^2}{n-1} \qquad S_A = \frac{\sum_i (a_i - \bar{a})^2}{n-1}$$

- Skalenunabhängig, zwischen –1 und +1
- Gute Performanz führen zu großen Werte!

Welches Maß?

- Am besten alle betrachten
- Oft ergibt sich kein Unterschied
- Beispiel:

	A	B	C	D
Root mean-squared error	67.8	91.7	63.3	57.4
Mean absolute error	41.3	38.5	33.4	29.2
Root rel squared error	42.2%	57.2%	39.4%	35.8%
Relative absolute error	43.1%	40.1%	34.8%	30.4%
Correlation coefficient	0.88	0.88	0.89	0.91

- D bester
- C zweit-bester
- A, B vergleichbar

Gliederung

- Motivation für Evaluation
- Schätzen des Klassifikationsfehlers
  - Trainieren Validieren und Testen
  - Fehler eingrenzen, Vertrauensintervalle
  - Aufteilung in Trainings und Testmenge
    - Wiederholtes Aufteilen
    - Kreuz-Validierung
    - Leave-One-Out
    - Bootstrap
- Performanzvergleich von DM Methoden
  - Einbeziehen der Varianz der Performanz: Signifikanz Tests
  - Gepaarter und Ungepaarter t-Test
- Performanz bei der Vorhersage von Wahrscheinlichkeiten
  - Quadratische und Informationsverlust Funktion
- Performanzvergleich bei verschiedenen Kosten der Fehler
  - Lift Charts
  - ROC Kurve
  - Precision und Recall
- Evaluierung nummerischer Vorhersagen
  - verschiedene Maße
- Das MDL Prinzip
  - Modellauswahl
  - Bayes Theorem, Log-Likelyhood und MDL
  - Epikurus Prinzip und Modellmittelung
  - MDL und Clusteranalyse

Das MDL-Prinzip

- MDL steht für *minimum description length*
- Die Beschreibungslänge ist definiert als:

  Platz um die Theorie zu beschreiben

  +

  Platz um die Ausnahmen der Theorie zu beschreiben
- In unserem Fall ist die Theorie der Klassifikator und die Ausnahmen sind die Fehler aus der Trainingsmenge
- Ziel: suche Klassifikator mit minimaler Beschreibungslänge
- MDL-Prinzip ist ein Modellauswahlkriterium

Modellauswahlkriterium

- Modellauswahlkriterium sucht einen guten Kompromiß zwischen:
  - Der Komplexiät des Modells und
  - der Vorhersagegenauigkeit auf den Trainingsdaten
- Idee: ein gutes Modell ist ein einfaches Modell, das eine hohe Genauigkeit auf den gegebenen Daten erreicht
- Auch bekannt als *Ockham's Rasiermesser* : die beste Theorie ist die Kleinste, die alle Fakten beschreibt

William of Ockham, born in the village of Ockham in Surrey (England) about 1285, was the most influential philosopher of the 14th century and a controversial theologian.

Eleganz vs. Fehler

- Theorie 1: sehr einfache, elegante Theorie, welche die Daten fast perfekt erklärt
- Theorie 2: signifikant komplexere Theorie, welche die Daten ohne Fehler erklärt
- Theorie 1 wird whr. bevorzugt
- Klassisches Beispiel: Kepler's drei Gesetze über Planetenbahnen
  - Weniger genau als Copernicus letzte Verbesserung der Ptolemäischen Theorie der Epizyklen

MDL und Kompression

- MDL Prinzip steht in Beziehung zur Datenkompression
  - Die beste Theorie ist jene, die die Daten am meisten komprimiert
  - I. Allg. wird eine Datenmenge komprimiert, indem ein Modell der Daten erstellt wird und die Ausnahmen zusätzlich gespeichert werden
- Folgendes muß berechnet werden
  - (a) Größe des Modells und
  - (b) Platzbedarf für die Ausnahmen
- (b) Leicht: nutze Entropie
- (a) das Modell muß kodiert werden

MDL und Bayes's Theorem

- L[T]="Länge" der Theorie
- L[E|T]=Trainingsmenge mittels Theorie kodiert
- Beschreibungslänge= L[T] + L[E|T]
- Bayes's Theorem ergibt *Posterior*-Whr. einer Theorie bei gegebenen Daten:

$$\Pr[T \mid E] = \frac{\Pr[E \mid T]\Pr[T]}{\Pr[E]}$$

- Äquivalent zu:

$$-\log \Pr[T \mid E] = -\log \Pr[E \mid T] - \log \Pr[T] + \underbrace{\log \Pr[E]}_{\textbf{\textit{konstant}}}$$

MDL und MAP

- MAP steht für *maximum a posteriori probability*
- Finden der MAP Theorie entspricht dem Finden der MDL Theorie
- Schwieriger Schritt beim Anwenden des MAP Prinzips: Bestimmen der Prior Whr. Pr[T] der Theorie
- Entspricht dem schwierigen Teil beim Anwenden des MDL Prinzips: finden des passenden Kodierungsschemas für die Theorie
- I.Allg. wenn bekannt ist, daß eine bestimmte Theorie wahrscheinlicher ist als andere, braucht man weniger Bits um sie zu kodieren

Diskussion des MDL Prinzips

- Vorteil: nutzt die Trainingsdaten voll aus, um das Modell zu bestimmen
- Nachteil 1: passendes Kodierungsschema bzw. Prior-Whr. für die Theorien sind entscheidend
- Nachteil 2: keine Garantie, daß die MDL Theorie den erwarteten Klassifikationsfehler minimiert
- Bemerkung: Ockham's Rasiermesser ist ein Axiom!
- Epikurus *Prinzip der mehrfachen Erklärungen*: nutze alle Theorien, die im Einklang mit den Daten sind

Bayesche Modellmittelung, BMA

- Reflektiert Epikurus Prinzip: alle Theorien werden mittels P[T|E] gewichtet und zur Vorhersage genutzt
- Sei *I* eine neue Instanz, deren Klasse vorhergesagt werden soll
- Sei *C* die Zufallsvariable, welche die Klasse angibt
- Dann ergibt BMA die Whr. für *C* bei gegebenem
  - *I*
  - Trainingsdaten *E*
  - möglichen Theorien T_j

$$\Pr[C \mid I, E] = \sum_j \Pr[C \mid I, T_j] \Pr[T_j \mid E]$$

MDL und Clusteranalyse

- Beschreibungslänge der Theorie:
  Anzahl der Bits um die Cluster zu kodieren
  - z.B. Cluster Repräsentanten
- Beschreibungslänge der Daten bezüglich der Theorie:
  kodiere Clustermitgliedschaft und Position relative zum Cluster
  - z.B. Distanz zum Clusterrepräsentanten
- Funktioniert falls das Kodierungsschema weniger Platz für kleine Zahlen als für große Zahlen verbraucht
- Bei nominalen Attributen muß die Klassenwahrscheinlichkeitsverteilung für jeden Cluster kodiert werden
- MDL kann genutzt werden, um den Parameter k bei k-Means zu bestimmen