Data Mining Zusammenfassung

Maximilian Ortwein

8. Februar 2013

Inhaltsverzeichnis

1 Data Understanding **3**
- 1.1 Data Vizualization . . . 4
  - 1.1.1 Types . . . 4
  - 1.1.2 Correlation Analysis . . . 4
  - 1.1.3 Outlier Detection . . . 4
  - 1.1.4 Missing Values . . . 4
  - 1.1.5 Checklist MUST DO . . . 5
- 1.2 Visualizing multidimensional data . . . 5
  - 1.2.1 Types . . . 5
  - 1.2.2 Outlier Detection . . . 5
  - 1.2.3 MDS Multi Dimensional Scaling . . . 5

2 Version Space **5**
- 2.1 Candidate Elimination . . . 5

3 Data Preparation **6**
- 3.1 Feature Selection . . . 6
- 3.2 Record Selection . . . 6
- 3.3 Data Cleansing . . . 6
- 3.4 improve data quality . . . 6
- 3.5 Missing Values . . . 7
- 3.6 Transformation of Data . . . 7
  - 3.6.1 categorical to numerical . . . 7
  - 3.6.2 numerical to categorical . . . 7
- 3.7 Normalisierung . . . 7
- 3.8 Principal Component Analysis PCA . . . 7

4 Modeling **8**
- 4.1 Errorfunctions . . . 8
- 4.2 Cost Matrix . . . 8

4.3 Cross Validation . . . 8
4.4 MDL (Minimum Description Length Principle) . . . 8

5 Clustering **8**
5.1 Hierarchical Clustering . . . 9
5.1.1 Dissimilarity (Abstände) . . . 9
5.2 Dendograms . . . 9
5.3 k-Means . . . 10
5.4 DBSCAN . . . 10

6 Association Rule **10**
6.1 frequent itemset mining . . . 10
6.2 Searching for frequent itemsets . . . 10
6.2.1 a priori . . . 10
6.2.2 Hassediagramm/Baum . . . 10
6.2.3 kanonische Form . . . 11
6.2.4 Kanonische Präfix Regel . . . 11
6.2.5 Frequent, Closed, Maximal . . . 11

7 Bayes + Regression **11**
7.1 Bayes Theorem . . . 11
7.2 Regression . . . 12
7.2.1 Regressions Linie . . . 12
7.2.2 Overfitting . . . 12

8 Decision Tree **12**
8.1 ID3 Algorythmus . . . 12
8.2 entropy . . . 12
8.3 prunning . . . 13
8.4 Regression Tree . . . 13

1 Data Understanding

Project Understanding

- Problem Formulation → Mapping to Datamining Task →
  Understanding the Situation
- 80-20 Rule: 20% Wird in Data und Project Understanding verwendet, ist aber zu 80% ausschlaggebend für den erfolg.

Data Understanding

- Questions:
  - Welche Arten von Attributen haben wir?
  - Wie ist die Qualität der Daten?
  - Hilft eine Visualisierung?
  - Sind die Attribute Correlated?
  - Gibt es Ausreißer?
  - Wie sollen missing values behandelt werden?

- Reihen sind Instanzen, Objekte oder Records, Spalten sind Attribute, features oder values.

- Datentypen:
  - Nominal (Klassen oder Kategorien; meist String)
  - Ordinal(Lineare Ordnung; Schulnoten oder Temperaturen)
  - Numeric (Zahlen)
  - discrete (Numeric oder Ordinal als Teilmenge von Integern)
  - continous (Reelle Zahlen)

- Data Quality:
  - Garbage in, garbage out
  - Accuracy = Nähe des Wertes aus den Daten am realen wert. - Geringe Accuracy durch: Noise, fehlerhafte eingaben, Tippfehler...
  - **Syntaktische Accuracy:** Eintrag ist nicht in der Domain, z.B. Text in Numerischen Daten. Einfach überprüfbar
  - **Semantische Accuracy:** Eintrag ist in der Domain aber fehlerhaft z.B. John Smith Female. Überprüfung aufwändiger
  - **Completeness:** Ist verletzt wenn die Daten nicht Vollständig sind und dadurch verzerrt (biased)
  - **Unbalanced Data:** Wenn ein eine Art von Einträgen extrem verrauscht ist.
  - **Timeliness:** Sind die Daten aktuell?

1.1 Data Vizualization

So bekommt man einen schnellen Überblick über die Daten und erkennt zum Beispiel verzerrungen oder Fehlende Werte.

1.1.1 Types

- Bar Charts/Histgrams: numbers of bins Sturges rule $k = \lceil log_2(n) + 1 \rceil$

- Boxplots: Boxgröße = Interquartilsabstand, Median einzeichnen als linie, Antennen: jeweils 1.5 facher Interquartilsabstand

- Scatterplots

1.1.2 Correlation Analysis

- Pearsons R: $r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$
  - Pearsons erkennt lineare Korrelation
  - Auch für monotone nicht lineare Korrelationen ist er nicht -1 oder 1
  - Er kann auch fast null sein Trotz einer monotonen Korrelation - Lösung: Rang Koeffizient

- Spearmans Rho: $\rho = 1 - 6 \cdot \frac{\sum_{i=1}^{n}(r(x_i) - r(y_i))^2}{n(n^2-1)}$

1.1.3 Outlier Detection

Ausreißer sind Datenpunkte die weit entfernt vom Rest liegen. Verursacht werden sie Durch schlechte Quality oder ungewöhnliche Extremwerte.
Handling: Meist ist es Sinnvoll diese Daten aus der Analyse auszuschließen. Wenn sie durch Fehler verursacht wurden auf jeden Fall.
Manchmal können Ausreißer auch das Ziel einer Analyse sein.
Für die Erkennung kann man Grubbs Test oder Boxplots verwenden.

1.1.4 Missing Values

Gründe für Missing Values sind fehlerhafte Sensoren, Verweigerung einer Antwort, oder Unsinnige Antworten unter bestimmten Bedingungen.
Behandeln kann man sie in dem man null Werte oder Standartwerte (Median, warscheinlichster Wert usw.) einsetzt.

1.1.5 Checklist MUST DO

- Die Verteilung eines Attributes überprüfen
- Korrelationen und Zusammenhänge aufdecken

1.2 Visualizing multidimensional data

1.2.1 Types

- 3D-Scatterplot
- Parallel Coordinates
- Radarplot
- Star Plot

1.2.2 Outlier Detection

- Scatterplots
- Visualize PCA or MDS
- Clustering punkte die keinem Cluster angehören sind Outlier

1.2.3 MDS Multi Dimensional Scaling

Positioniert die Datenpunkte im 2D-Raum und nutzt dabei die Abstände im n-D Raum um die Struktur zu erhalten.
Braucht extrem viele Parameter für das Iris Set z.B. 300 da es für jeden datenpunkt die Position x und y bestimmen muss. Zum minimieren der Funktion wird das gradient descent Verfahren verwendet.

2 Version Space

- Syntaktisch unterschiedlicher Hypothesenraum wird mit (k+2) multipliziert.
- Semantisch unterschiedlicher Hypothesenraum wird durch $(k_1+1)\cdot(k_2+2)\cdots(k_n+n)$ berechnet

2.1 Candidate Elimination

Anfang:
$S\{<\emptyset,\emptyset,\emptyset>\}$
$G\{<?,?,?>\}$
Dann werden für jedes Positive oder Negative Beispiel S und G so angepasst das es akzeptiert wird oder nicht. S wird verallgemeinert, G wird spezialisiert. Die Reihenfolge der

Beispiele hat dabei keinen Einfluss auf das Ergebnis.

HIER EVENTUELL ERGÄNZEN!!!!!!!!!!

3 Data Preparation

3.1 Feature Selection

Irrelevante Features entfernen und Redundante Features Entfernen

- Wähle die Features mit der Besten Bewertung wenn einzelne Features evaluiert werden.
- Wähle das bestes Subset aus. Dies ist sehr Kostenintensiv und nur für sehr kleine Mengen möglich.
- Forward Selection: Starte mit einer Leeren Menge und füge immer das Feature hinzu das die Accuracy am meisten erhöht.
- Backward Elimination: Starte mit allen Features und entferne alle ungeeigneten.

3.2 Record Selection

Gründe für Subsamples:

- Schnellere Berechnung
- Crossvalidation mit einen Trainings und einem Testset
- Timeliness Veraltete Daten können entfernt werden
- Represenetativeness: Finde ein Repräsentatives Subsample
- Rare Events müssen insbesondere mit einbezogen werden

3.3 Data Cleansing

Finde und Korrigiere oder entferne Inaccurate, Incorrecte oder Incomplete Einträge aus dem Datensatz.

3.4 improve data quality

Alle Buchstaben zu Großbuchstaben ändern. Spaces und unsichtbare Zeichen entfernen. Format von Zahlen anpassen. Aufteilen von Spalten mit mehreren Informationen. Abkürzungen entfernen und Rechtschreibung korrigieren. Schreibweise von addressen vereinheitlichen. Zahlen in standard Format bringen. Überprüfe durch Wörterbücher ob die Werte in die Domäne passen.

3.5 Missing Values

- Entferne den Eintrag
- Ersetze den Wert (Median, Mittelwert...)
- Ersetze den Wert durch einen speziellen Wert

3.6 Transformation of Data

3.6.1 categorical to numerical

- Binary: 1 und 0
- Ordinal: In ihrer Ordnung nummerieren z.B. 1 - k
- Categroical Sollten nicht in eine Einzelne Zahl konvertiert werden.

3.6.2 numerical to categorical

Einen Numerischen Bereich in Bins Aufteilen.

- Equi-Width discretization: Teile die Intervalle in gleichlange Bins auf.
- Equi-frequency discretization: Teile die Daten in Bins mit der gleichen Anzahl an Elementen auf.
- V-optimal discretization.
- Minimal entropy discretization. Minimizes the entropy. (Only applicable in the case of classification problems.)

3.7 Normalisierung

Für Manche Datenanalysen (PCA, MDS Clustering) ist die Skala der Daten entscheidend. Deshalb ist es Notwendig die Daten auf eine Standard Skala zu mappen. Normalerweise werden die Daten auf einen Bereich zwischen 0 und 1 gemappt.

3.8 Principal Component Analysis PCA

Erzeugt eine Projektion aus einem Hochdimensionalen Raum in einen Niederdimensionalen Raum. Es nutzt die Varianz der Daten zur Strukturerhaltung. Es versucht möglichst viel der originalvarianz zu erhalten.

4 Modeling

1. Select the model class
2. select the score function
3. apply the algorithm
4. validate the result

Wähle das Einfachste Model was die Daten noch erklärt!
Globale Modelle (regression) zeigen das Komplette Dataset.
Lokale Modelle (Association Rules) nur ein Subset.

4.1 Errorfunctions

Falschklassifieziertenrate: $\frac{\#Falschklassifiziert}{\#alleDaten}$
Sagt nichts über die Qualität des Classifyers aus.
Klassen können nicht Balanciert sein.

4.2 Cost Matrix

Ein generellerer Ansatz als die Errorfunction.
Die Kosten einer falschen Klassifikation können für jede Klasse anders sein.
Deshalb werden sie für jede Klasse in eine Matrix geschrieben.

4.3 Cross Validation

Teile die Daten in k Teile. Dann nutze immer einen Teil als Testdatensatz den Rest als Trainingsdatensatz. Der die durchschnittliche fehlerrate ist die Fehlerrate des Models.

4.4 MDL (Minimum Description Length Principle)

Ein zu komplexes Modell führt oft zu Overfitting (Das Modell ist zu stark auf die Trainingsdaten angepasst). Die MDL besagt, wähle das Modell das bei gleicher Vorhersagequalität das Modell mit den Wenigsten Trainingsdaten zu nehmen ist.

5 Clustering

Objekte eines Clusters sollten möglichst gleich sein, und Objekte unterschiedlicher Cluster möglichst verschieden.
Cluster können unterschiedliche Formen, Größen und Dichten haben

Können eine Hirarchie bilden
können sich überschneiden

Data Understanding, Class Identification, Reduction (Repräsentanten finden), Outlier Detection, Nois Detection

Typen:
- Linkage Based
- by Partitions
- Desity based
- grid based

Vor dem Clustern sollten die Daten normalisiert werden.
Cluster sollten nicht von Maßeinheiten abhängen!

5.1 Hierarchical Clustering

Cluster werden Schritt für Schritt aufgebaut normalerweise Bottom up. Das heißt jeder Datenpunkt ist ein Cluster und wird dann in jedem Schritt verschmolzen. agglomerative

Das Gegenteil ist die Top Down Strategie, alle Daten sind in einem Cluster und werden Zerlegt. divisive

Für die Entscheidung welcher Datenpunkt zu welchem Cluster gehört, muss man die Similarity messen.

5.1.1 Dissimilarity (Abstände)

- Centroid: Der Abstand zwischen den Mittelpunkten der Cluster.
- Average: Der durchschnittliche Abstand aller Punkte der Cluster
- Single Linkage: der Abstand der zwei am engsten beieinander liegenden Punkte der Cluster. (Kann ketten verursachen!)
- Complete Linkage: Abstand der am weitesten entfernten Punkte der Cluster

5.2 Dendograms

Sind binäre Bäume. Unten oder Links kommen die Datenpunkte hin. Die Cluster werden miteinander verbunden und es wird bei der Verbindung der Abstand zwischen den Clustern eingezeichnet.

5.3 k-Means

- Bestimme die Anzahl von Clustern k (Ist eine Benutzereingabe)
- wähle zufällig k Datenpunkte als Anfangsmittelpunkte
- Weise jedem Datenpunkt sein Zentrum zu (das mit dem Kleinsten Abstand)
- Berechne die Zentren neu als Durchschnitt aller punkte eines Clusters
- Wiederhole ab schritt 3 bis keine Änderungen mehr auftreten

5.4 DBSCAN

Dichtenbasierte Clusteringalgorithmen haben oft die besten Ergebnisse auf numerische Daten. Sie weisen die regionen mit Hohen dichten einem Cluster zu.

- Finde einen Datenpunkt mit einer hohen Dichte um sich
- Alle e Nachbarn gehören zu dem Cluster
- So lange die dichte hoch ist erweitere das Cluster
- entferne das cluster aus dem datensatz und fang wieder von vorne an

6 Association Rule

6.1 frequent itemset mining

- Support: $\frac{\#\text{Transaktionen in denen das Itemset vor kommt}}{\#\text{alle Transaktionen}}$
- Confidence: $\frac{support(X \cup Y)}{support(X)}$

6.2 Searching for frequent itemsets

6.2.1 a priori

Nimmt den MinSupport entgegen.
Es werden nach und nach alle Itemsets ausgeschlossen, die den MinSupport nicht erfüllen. Man fängt bei einzelnen Items an und Schließt nach und nach noch verbleibende kombinationen aus. Alle übrigbleibenden Sets sind dann Regeln.

6.2.2 Hassediagramm/Baum

Wenn man jedem Knoten in einem Hassediagramm nur einen Elternknoten zuweist, zum Beispiel das erste Item, erhält man den Baum, dieser hat den Vorteil das bei einer Traversierung jeder Knoten genau einmal besucht werden muss.

6.2.3 kanonische Form

Eine kanonische Form ist dann gegeben wenn die Items lexikografisch geordnet sind: aus einer Menge {A,B,C,D} wären ABC kanonisch ACB aber nicht.

6.2.4 Kanonische Präfix Regel

Jeder Präfix einer kanonischen Form ist selbst kanonisch!
Mit dieser Regel können Itemsets einfach rekursiv durchsucht werden. Es können einfach durch hinzufügen eines Items an das ende eines kanonischen Wortes alle daraus resultierenden kanonischen Gefunden werden. Dazu muss nur noch überprüft werden ob das neue Wort kanonisch ist.

6.2.5 Frequent, Closed, Maximal

Frequent: Itemsets, die den Min Support erfüllen.
Closed: Itemsets, die Keine Supersets mit gleichem oder größerem Support haben.
Maximal: Itemsets die Keine Supersets mehr haben die Frequent sind.

7 Bayes + Regression

7.1 Bayes Theorem

$P(h|E) = \frac{P(h|E) \cdot P(h)}{P(E)}$ Es wird immer die Höchste warscheinlichkeit genommen.
Naive Bayes heisst, es wird davon ausgegangen, dass die Ereignisse voneinander unabhängig sind.
LaPlace Correction kann divisionen durch 0 Entfernen.

Pros:
– Gold standard for comparison with other classifiers
– High classification accuracy in many applications
– Classifier can easily be adapted to new training objects
– Integration of domain knowledge

Cons:
– The conditional probabilities my not be available
– Independence assumptions might not hold for data set

7.2 Regression

7.2.1 Regressions Linie

Eine Lineare Funktion die sich den Daten möglichst gut annähert.
$y = a \cdot x + b$
$b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$
$a = \overline{y} - b\overline{x}$

7.2.2 Overfitting

Komplexe Funktionen neigen zum Overfitting. Das heißt die Funktion ist zu stark an die Trainingsdaten angepasst und daher dann schlechtere Vorhersagen liefert.

Pros:
– Strong mathematical foundation
– Simple to calculate and to understand (for a moderate number of dimensions)
– High predictive accuracy

Cons:
– Many dependencies are non-linear
– Global model does not adapt to locally different data distributions
=> Locally weighted regression

8 Decision Tree

8.1 ID3 Algorythmus

- Berechne die Entropy aller Attribute
- Das Attribut mit der höchsten Entropy wird die Wurzel
- Führe das für alle Teilbäume aus
- den Baum dadurch rekursiv aufbauen

8.2 entropy

Entropy: $H = -\sum_{i=1}^{n} p_i \log p_i$
Information Gain = H(Class) - H(Class|Atrribut)
$H(C|A) = -\sum_{j=1}^{An} p_j - (\sum_{i=1}^{Cn} p_{i|j} \log p_{i|j})$

Gini Index: $1 - \sum_{j=1}^{An} p_j$

8.3 prunning

Schlechte zweige abschneiden und durch ein Blatt austauschen.
Schützt vor overfitting.

8.4 Regression Tree

Damit werden numerische Werte vorhergesagt statt Klassen. Ähnlich aufgebaut wie Entscheidungsbäume.