DataMining CheatSheet

Julian Schubert

13. Juli 2021

1 Gütemaße

1.1 Davies-Bouldin Index (DB)

Güte innerhalb des Clusters C_i	$S_i \sqrt[q]{\frac{1}{\lvert C_i \rvert} \sum_{x \in C_i} \text{dist}(x, \mu_i)^q}$
Güte Trennung C_i und C_j	$M_{i,j} = \text{dist}(\mu_i, \mu_j)$
$R_{i,j}$ für $i \neq j$	$R_{i,j} = \frac{S_i + S_j}{M_{i,j}}$
Davis-Bouldin Index	$DB = \frac{1}{k} \sum_{i=1}^{k} D_i$ mit $D_i = \max_{i \neq j} R_{i,j}$

2 Distanzfunktionen

2.1 Distanzfunktionen für Cluster

Single Link	dist - sl(X, Y) = $\min_{x \in X, y \in Y}$ dist(x, y)
Complete Link	dist - cl(X, Y) = $\max_{x \in X, y \in Y}$ dist(x, y)
Average Link	dist - al(X, Y) = $\frac{1}{\lvert X \rvert \cdot \lvert Y \rvert} \cdot \sum_{x \in X, y \in Y}$ dist(x, y)

3 Dichtebasiertes Clustern

- **Kernobjekt**: Mehr als MinPts in ϵ-Umgebung
- **direkt dichte-erreichbar**: $p \in N_\epsilon(q)$ und q ist Kernobjekt
- **dichte-erreichbar**: Kette von dichte-erreichbaren Objekten zwischen q und p
- **dichte-verbunden** Beide von einem dritten Objekt dichte-erreichbar

4 DBSCAN

Beschreibung in Worten:

1. Wählt zufälligen noch nicht klassifizierten Punkt
2. Führt ExpandiereCluster für diesen Punkt aus

3. **ExpandiereCluster**:
   - Punkt ist Noise -> FALSE zurück geben
   - Sonst: Füge alle dichte-erreichbaren Punkte vom gegebenen Punkt zum Cluster hinzu

5 OPTICS

Beschreibung in Worten:

1. Über alle Punkte iterieren
2. Wenn Punkte im Umkreis vom aktuellen Punkt liegen Distanzen updaten
3. Alle Nachbarn vom Punkt abarbeiten
4. Sortiert in die Liste einfügen

6 Assoziationsregeln

- **Support**: $supp_D(X) = \frac{|\{T \in D | X \subseteq T\}|}{|D|}$
- **Frequency:** $supp_X(D) \cdot |D|$
- **Confidence:** $conf_D(X \rightarrow Y) = \frac{supp_D(X \cup Y)}{supp_D(X)}$

7 Auswahl von Assoziationsregeln

7.1 Added Value

$$\frac{sup(A \wedge B)}{sup(A)} - sup(B) = conf(A \rightarrow B) - sup(B)$$

Um wie viel steigt die Wahrscheinlichkeit von B, wenn die Bedingung A Hinzugefügt wird?

7.2 Kriterien für Interessantheitsmaße

1. Conciseness:
   Kürzere Regeln sind besser (weniger Items)
2. Generality:
   Generelle Regeln sind besser (mehr Fälle abgedeckt)
3. Reliability:
   Hohe confidence / accuracy ist besser
4. Diversity:
   Regeln sollten untereinander unähnlich sein

5. Novelty:
   Vorher unbekannt, nicht aus anderen Regeln ableitbar
6. Surprisingness / Unexpectedness:
   Gute Regeln widersprechen Vorwissen / Erwartungen
7. Applicability:
   Kann praktisch (in der Anwendung) umgesetzt werden

7.3 Monotonie

- **Monotonie**: If a set S violates C, its supersets **might not** violate C, while its subsets **must** violate C
- **Anti-Monotonie**: If a set S violates C, its supersets **must** violate C, while its subsets **might not** violate C

8 Naive Bayes

Entscheidungsregel des naiven Bayes Klassifikators:

$$\text{argmax}_{c_j \in C} P(c_j) \cdot \prod_{i=1}^{d} P(o_i|c_j)$$

9 Hierarchische Assoziationsregeln

- **Definition** Hierarchische Assoziationsregel:
  $X \Rightarrow Y$, mit $X \subseteq I, Y \subseteq I, X \cap Y = \varnothing$
  Kein Item in Y ist sVorfahre eines Items in X (bezüglich H)
- **Support s** einer hierarchischen Assoziationsregel $X \Rightarrow Y$ in D:
  Support der Menge $X \cup Y$
- **Konfidenz c** einer hierarchischen Assoziationsregel $X \Rightarrow Y$ in D:
  Prozentsatz der Transaktionen, die auch die Menge Y unterstützen, in der Teilmenge aller Transaktionen, welche die Menge X unterstützen

10 Gütemaße für Klassifikation

K: Klassifikator, **TR** Trainingsmenge, **TE** Testmenge

- **Klassifikationsgenauigkeit** G_{TE}:
  Alles was aus dem Testset richig klassifiziert wurde
- **Tatsächlicher Klassifikationsfehler** F_{TE}:
  Alles was aus dem Testset falsch klassifiziert wurde
- **Beobachteter Klassifikationsfehler** F_{TR}:
  Alles was aus dem Trainingsset falsch klassifiziert wurde

11 Precision und Recall

Precision: $\frac{\text{True positive}}{\text{True Positive+false Positive}}$
Wenn die klasse vorhergesagt wird, wie sicher ist die Vorhersage

Recall: $\frac{\text{True positive}}{\text{True Positive+False Negative}}$
Wie oft wird die Klasse c wieder gefunden

12 Gütemaße für Splits

12.1 Informationsgewinn

Entropie:
$entropie(T) = -\sum_{i=1}^{k} p_i \log(p_i)$

Informationsgewinn:
$informationsgewinn(T, A) = entropie(T) - \sum_{i=1}^{m} \frac{|T_i|}{|T|} \cdot entropie(T_i)$

12.2 Gini-Index

$$gini(T) = 1 - \sum_{j=1}^{k} p_j^2$$

Kleiner Gini-Index $\Leftrightarrow$ geringe Unreinheit
Großer Gini-Index $\Leftrightarrow$ hohe Unreinheit
Gini-Index des Attributs A in Bezug auf T ist definiert als

$$gini_A(T) = \sum_{i=1}^{m} \frac{|T_i|}{|T|} \cdot gini(T_i)$$

13 Delta-Rule

$$w = w + \eta \cdot x \cdot (t - o)$$

Vereinfacht:

- Berechne $o = \Theta(w^T x)$
- Falls $o > 0$ und $t = 0$
  - Setze $w = w - x$
- Falls $o \leqslant 0$ und $t = 1$
  - Setze $w = w + x$

14 Backpropagation

Fehler Output Layer:

$$-(t_d - o_d) \cdot \Theta'(Wy) \cdot y_j$$