Ludwig-Maximilians-Universität München
Institut für Informatik
PD Dr. Matthias Schubert, Dr. Eirini Ntoutsi
Erich Schubert

Knowledge Discovery in Databases II

WS 2012/2013

Übungsblatt 5: Stream data clustering

Aufgabe 5-1 *Berechnung der Varianz*

Betrachten Sie die 1-dimensionalen Datensätze:

$$A = \{0, +1, -1, +2, -2, +3, -3, \ldots, +100, -100\} \quad B = \left\{a_i + 10^{10}\right\} \quad C = \left\{a_i \cdot 10^{-10} + 1\right\}$$

Berechnen Sie die Varianz der drei Datensätze mit:

- Dem naiven Ansatz in zwei Durchläufen: berechnen Sie zunächst den Mittelwert, dann die Varianz über die mittlere quadratische Abweichung.

$$\mu := \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{Var}(X) := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$$

- Mit Hilfe der Steiner-Verschiebung in nur einem Durchlauf:

$$\text{Var}(X) := \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i^2) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$$

- Mit Hilfe des von Knuth und Welfort diskutierten Algorithmus:

```
n = 0
mean = 0.0
sum2 = 0.0
for x in data:
        n = n + 1
        delta = x - mean
        mean = mean + delta/n
        sum2 = sum2 + delta*(x - mean)
var = sum2 / (n-1)
```

- Probieren Sie es in ihrer Lieblings-Anwendung aus (e.g. NumPy, R, ELKI).
  Liefert das Programm den korrekten Wert? Verwenden Sie $n - 1$ Freiheitsgrade!
- Zeigen Sie, dass alle drei Methoden mathematisch äquivalent sind.
  (Tipps: kürzen Sie den Term $\frac{1}{n-1}$ frühzeitig. Für die dritte Methode: verwenden Sie vollständige Induktion und bestimmen Sie geeignete Invarianten!)

Verwenden Sie `double` Genauigkeit für ihre Brechnungen.
Was beobachten Sie? Wie können Sie dieses Problem erklären?

Aufgabe 5-2 *k-means auf Streams*

Von k-means sind zwei Varianten geläufig:

- Lloyd-Forgy-Variante: 2 Phasen, Objekte zuweisen und Mittelwerte aktualisieren
- MacQueen-Variante: immer je 1 Objekt zuweisen, Mittelwert aktualisieren

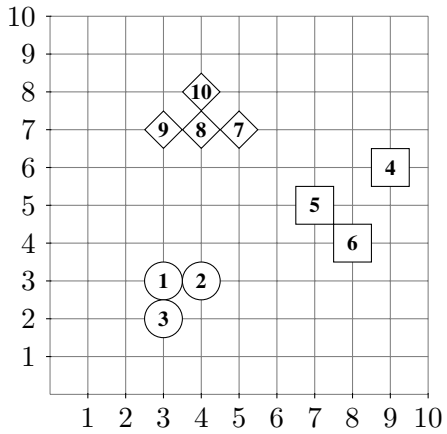
Rufen Sie sich beide Varianten in Erinnerung. Auf Stream-Daten kann MacQueen unverändert verwendet werden (kann aber mit "concept drift" nicht gut umgehen). In verteilten Systemen ist Lloyd einfacher umzusetzen.

Überlegen Sie sich, woran man festmachen/erkennen kann, wieso die eine Variante einfacher für Streams einfacher zu adaptieren ist, die andere einfacher für parallele Systeme verwendet werden kann.

Aufgabe 5-3 *Cluster Features*

Gegeben folgender Datensatz:

ObjID	Cluster	X	Y	t
1	A	3	3	1.7
2	A	4	3	3.5
3	A	3	2	1.2
4	B	9	6	4.1
5	B	7	5	5.0
6	B	8	4	1.2
7	C	5	7	4.7
8	C	4	7	2.3
9	C	3	7	2.2
10	C	4	8	2.2



Berechnen sie die CluStream Cluster Features CFT für jeden der drei Cluster.

Eine neue Beobachtung im Stream sei $p = (X = 8, Y = 5, t = 6.1)$.

Führen Sie die "online micro-cluster maintainance" von CluStream für diesen Punkt p durch.