Humboldt-Universität zu Berlin
Institut für Informatik
Lehrstuhl Logik in der Informatik
Prof. Dr. Nicole Schweikardt

12. Juli 2016

Big Data Analytics in Theorie und Praxis

Sommersemester 2016

Theorie-Übungsblatt 5

Zu bearbeiten bis 19. Juli 2016

Aufgabe 1: **(35 Punkte)**

Arbeiten Sie die Details dazu aus, wie man die Datenstruktur *Count-Min-Sketch* dazu verwenden kann, *„approximative heavy hitters“* zu berechnen.

Aufgabe 2: **(25 Punkte)**

Seien $m, M \in \mathbb{N}_{\geqslant 1}$, sei $U := \{0, \ldots, m-1\}$ und $V := \{0, \ldots, M-1\}$. Sei H eine streng 2-universelle Familie von Hashfunktionen $h : U \to V$.

Sei $t \in V$, sei $s \geqslant 2$ und seien $u_1, \ldots, u_s$ paarweise verschiedene Elemente aus U.

Wir wählen zufällig und gleichverteilt ein $h \in H$ und betrachten für jedes $i \in \{1, \ldots, s\}$ die Zufallsvariable $Y_i := \begin{cases} 0 & \text{falls } h(u_i) > t \,, \\ 1 & \text{sonst} \,. \end{cases}$

Beweisen Sie, dass für alle $i, j \in \{1, \ldots, s\}$ mit $i \neq j$ die beiden Zufallsvariablen Y_i und Y_j unabhängig sind.

Zur Erinnerung:

(1) Zwei Zufallsvariablen X und Y heißen *unabhängig*, wenn für alle Werte a und b gilt, dass $\Pr(X = a \text{ und } Y = b) \;=\; \Pr(X = a) \cdot \Pr(Y = b)$.

(2) Eine Familie H von Funktionen von U nach V heißt *streng 2-universell*, falls für alle $x_1, x_2 \in U$ mit $x_1 \neq x_2$ und alle $y_1, y_2 \in V$ gilt: Für ein zufällig und gleichverteilt aus H gewähltes h ist

$$\Pr\Big(\, h(x_1) = y_1 \;\text{ und }\; h(x_2) = y_2 \,\Big) \;\;=\;\; \frac{1}{M^2} \,.$$

Anmerkung: Zusammen mit der Tatsache, dass für paarweise unabhängige Zufallsvariablen die Varianz der Summe gleich der Summe der Varianzen der Zufallsvariablen ist, liefert diese Aufgabe die Rechtfertigung dafür, dass wir im Beweis des Gütekriteriums des in der Vorlesung behandelten Schätzers F_0^* für F_0 verwendet haben, dass

$$\mathrm{Var}\left(\sum_{i=1}^{A} Y_i\right) \;=\; \sum_{i=1}^{A} \mathrm{Var}(Y_i)$$

für die dort betrachteten Variablen Y_i ist.

Auf der Rückseite finden Sie eine weitere Aufgabe.

Aufgabe 3: **(40 Punkte)**

Sei $m \in \mathbb{N}_{\geqslant 1}$ und $U := \{0, \ldots, m-1\}$. Für einen Datenstrom $x_1, \ldots, x_n$ mit $x_i \in U$, $m > n$ und $x_i \neq x_j$ für $i, j \in \{1, \ldots, n\}$ mit $i \neq j$ soll eine Näherung für den Median ermittelt werden. Diese Näherung des Medians soll durch eine zufällig, gleichverteilte Stichprobe S der Größe s aus dem Datenstrom bestimmt werden. Die Größe s der Stichprobe soll so gewählt sein, dass für vorgegebene Parameter ε und δ der Median der Stichprobe eine *ε-δ-Approximation* des Medians des Datenstroms liefert. Der Begriff „ε-δ-Approximation“ bedeutet hier, dass (1) gilt, wobei folgende Notationen verwendet werden.

Sei $x_{\pi_1}, \ldots, x_{\pi_n}$ der Datenstrom, wenn man ihn nach den Werten aufsteigend sortieren würde. Dies wird natürlich nicht durchgeführt, ist aber für die nun folgende Definition hilfreich.

Sei $\mathrm{Pos}(x_{\pi_i}) := i$ für $i \in \{1, \ldots, n\}$ die *Position* von x_{π_i} im sortierten Datenstrom. Für ungerade n ist offensichtlich der Median *Med* des Datenstroms das Element $x_{\pi_{\frac{n+1}{2}}}$, d.h. das Element mit Position $\frac{n+1}{2}$ im sortierten Datenstrom. Für gerade n wäre der Median *Med* das Element $x_{\pi_{\frac{n}{2}}}$ oder $x_{\pi_{\frac{n}{2}+1}}$. Zur Vereinfachung nehmen wir für all diese Fälle $\mathrm{Pos}(\mathit{Med}) = \lceil \frac{n}{2} \rceil$ an, und in Rechnungen können Sie für $\mathrm{Pos}(\mathit{Med})$ einfach $\frac{n}{2}$ einsetzen.

Bestimmen Sie eine möglichst kleine Zahl s für die Größe der zu ziehenden Stichprobe S, so dass s nur von ε und δ (für gegebene ε, δ mit $0 < \varepsilon < 1$ und $0 < \delta < 1$) und einer Konstanten c abhängt. Hierbei hängt c nicht von $\varepsilon, \delta, x_1, \ldots, x_n, n, m$ oder s ab. Außerdem soll für den Median *med* der Stichprobe S Folgendes gelten:

$$\Pr\Big(\mathrm{Pos}(\mathit{med}) \in (\tfrac{1}{2} \pm \varepsilon) \cdot n\Big) \geqslant 1-\delta\,. \tag{1}$$

D.h., die Position vom Median *med* der Stichprobe im sortierten Datenstrom weicht von der Position des Medians des gesamten Datenstroms mit einer Wahrscheinlichkeit von mindestens $1-\delta$ höchstens um $\varepsilon \cdot n$ ab.

Hinweis: Überlegen Sie, in welchen Fällen die Position des Medians der Stichprobe im sortierten Datenstrom um mehr als $\varepsilon \cdot n$ von $\frac{n}{2}$ abweicht und benutzen Sie dann die Chernoff-Schranke für die einzelnen Fälle. Beachten Sie, dass das ε in der Chernoff-Schranke nicht notwendigerweise dasselbe ε wie in dieser Aufgabenstellung ist.