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

2. Juli 2015

Big Data Analytics in Theorie und Praxis

Prof. Johann-Christoph Freytag, Ph.D. und Prof. Dr. Nicole Schweikardt
Sommersemester 2015

Theorie-Übungsblatt 4

Zu bearbeiten bis 9. Juli 2015

Aufgabe 1: **(40 Punkte)**

Für einen Datenstrom $x_1, \ldots, x_n$ mit $x_i \in \{1, \ldots, m\}$, $m > n$ 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.

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

Sei $\text{Pos}(x_{\pi_i}) := i$ 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 $\text{Pos}(\mathit{Med}) = \lceil \frac{n}{2} \rceil$ an, und in Rechnungen können Sie für $\text{Pos}(\mathit{Med})$ 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:

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

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 aus der Vorlesung für die einzelnen Fälle. Beachten Sie, dass die Chernoff-Schranke auch gilt, wenn die Betragsstriche weggelassen werden. Außerdem ist das ε in der Chernoff-Schranke nicht notwendigerweise dasselbe ε wie in dieser Aufgabenstellung.

Auf der Rückseite finden Sie weitere Aufgaben.

Aufgabe 2: **(30 Punkte)**

Sei $m \in \mathbb{N}$ und sei M eine Primzahl mit $M > m$. Für alle $a, b \in \{0, 1, \ldots, M-1\}$ sei $h_{a,b}$ die Funktion $h_{a,b} : \{1, \ldots, m\} \to \{0, \ldots, M-1\}$ mit

$$h_{a,b}(x) \;=\; ax + b \pmod{M}, \qquad \text{für alle } x \in \{1, \ldots, m\}.$$

Sei H die Familie all dieser Funktionen $h_{a,b}$, d.h.

$$H \;:=\; \Big\{\, h_{a,b} \;:\; a, b \in \{0, \ldots, M-1\} \,\Big\}$$

Beweisen Sie, dass H paarweise unabhängig ist.

Zur Erinnerung: Eine Familie H von Funktionen von $\{1, \ldots, m\}$ nach $\{0, \ldots, M-1\}$ heißt *paarweise unabhängig* (oder *2-universell*), falls für alle $x_1, x_2 \in \{1, \ldots, m\}$ mit $x_1 \neq x_2$ und alle $y_1, y_2 \in \{0, \ldots, M-1\}$ gilt: Für ein zufällig, gleichverteilt aus H gewähltes h ist

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

Aufgabe 3: **(20 Punkte)**

Beweisen Sie, dass für $A \geq 2$ paarweise unabhängige Zufallsvariablen $Y_1, \ldots, Y_A$ gilt, dass

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

Erläutern Sie Ihre Zwischenschritte.

Hinweis: Seien X und Y zwei unabhängige Zufallsvariablen und sei a eine beliebige Zahl. Dann gelten folgende Regeln:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}(\,(X - \mathrm{E}(X))^2\,)\,, \\
\mathrm{Var}(X) &= \mathrm{E}(X^2) - \mathrm{E}(X)^2\,, \\
\mathrm{E}(X + Y) &= \mathrm{E}(X) + \mathrm{E}(Y)\,, \\
\mathrm{E}(X \cdot Y) &= \mathrm{E}(X) \cdot \mathrm{E}(Y)\,, \\
\mathrm{E}(a \cdot X) &= a \cdot \mathrm{E}(X)\,.
\end{aligned}$$

Aufgabe 4: **(10 Punkte)**

Beweisen Sie, dass für jede Zufallsvariable Y und jede Zahl a gilt:

$$\mathrm{Var}(a \cdot Y) \;=\; a^2 \cdot \mathrm{Var}(Y)\,.$$