Προηγούμενα φορά

Markov's Inequality

Αν X μη αρνητική τ.μ. $\forall t$

$$Pr[X \geq t] \leq \frac{E[X]}{t}$$

Chebyshev's Inequality

Αν X είναι τυχαία μεταβλητή με $Var[X] = \sigma^2$

$$Pr[|X - E[X]| \geq k\sigma] \leq \frac{1}{k^2} \quad \forall k \geq 1$$

Hoeffding Inequality

Αν $X = \sum X_i$ με X_i ανεξάρτητες τ.μ. στο $[a_i, b_i]$ τότε για κάθε $t > 0$

$$Pr[|X - E[X]| \geq t] \leq 2e^{-\frac{2t^2}{\sum_i (b_i - a_i)^2}}$$

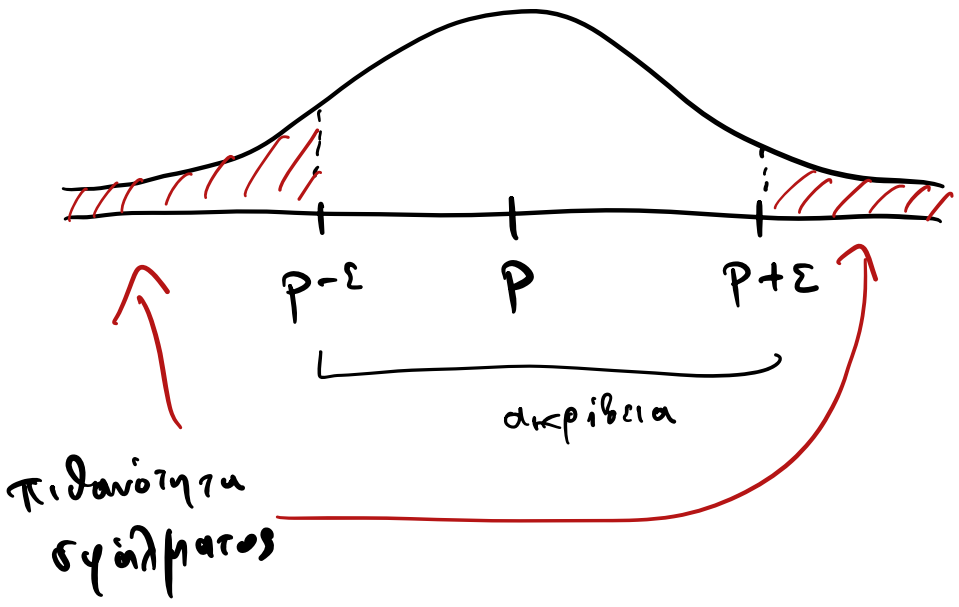
Πιθανότητα ενός κέρματος p

Αν $\hat{p} = \frac{\sum_{i=1}^{m} x_i}{m}$

$$Pr\left[|\hat{p} - p| > \varepsilon \right] \leq e^{-\Omega(\varepsilon^2 m)}$$

Με $m = \Theta\left(\frac{1}{\varepsilon^2} \log(1/\delta)\right)$ δείγματα

$|\hat{p} - p| < \varepsilon$ με πιθανότητα $1 - \delta$



Εκτίμηση Πλήθους Διαφορετικών Στοιχείων

Πρόβλημα: Μεγάλο stream από δεδομένα

$x_1, \ldots, x_N$

Πόσα είναι διαφορετικά;

Έστω τα διαφορετικά στοιχεία έχουν πλήθος n

με $O(n)$ μνήμη μπορώ να τα υπολογίσω ακριβώς.

Τεχνική: Hash κάθε στοιχείο

$x_i \to h(x_i) \sim U[0,1]$

$h(x_1) = h(x_8)$ αν το $x_1 = x_8$

$h(x) = md5(x)$

$$Y = \min_{i=1}^{N} h(x_i) = \min_{i=1}^{n} s_i$$

όπου $s_i \sim U[0,1]$

Claim: $E[Y] = \frac{1}{n+1}$

Απόδειξη

$$E[Y] = \int_0^1 Pr[Y=t] \cdot t \, dt$$

$$= \int_0^1 \left(-Pr[Y \geq t]\right)' t \, dt$$

$$= \int_0^1 Pr[Y \geq t] \, dt$$

$$= \int_0^1 (1-t)^n \, dt = \int_0^1 t^n dt$$

$$= \frac{1}{n+1}$$

□

$$n = \frac{1}{y} - 1$$

Claim : $\mathrm{Var}[Y] \le \frac{1}{(n+1)^2}$

$$\mathrm{Var}[Y] = E[Y^2] - E[Y]^2$$

Αποδείξη.

$$E[Y^2] = \int_0^1 \Pr[Y = t]\, t^2\, dt$$

$$= \int_0^1 n(1-t)^{n-1} t^2\, dt$$

$$= \frac{2}{(n+1)(n+2)} \le \frac{2}{(n+1)^2}$$

Από Chebyshev

$$\Pr\left[\left|Y - \tfrac{1}{n+1}\right| \ge \frac{\kappa}{n+1}\right] \le \frac{1}{\kappa^2}$$

Μείωση του Variance

Επαναλαμβάνω το πείραμα t φορές.

$$Y_1 , Y_2 , \dots , Y_t$$

$$\bar{Y} = \frac{\sum_{i=1}^{t} Y_i}{t} \rightarrow \begin{cases} E[\bar{Y}] = \frac{1}{t}\sum_{i=1}^{t} E[Y_i] \\ \quad = \frac{1}{n+1} \\ Var[\bar{Y}]_t = \frac{1}{t^2} Var[\sum Y_i] \\ \quad = \frac{\sum_{i=1} Var[Y_i]}{t^2} \\ \quad \leq \frac{1}{t(n+1)^2} \end{cases}$$

Από Chebyshev

$$Pr\left[\left| \bar{Y} - \frac{1}{n+1} \right| \geq \frac{\kappa}{\sqrt{t}(n+1)} \right] \leq \frac{1}{\kappa^2}$$

Αν $\kappa = 2$ $\frac{1}{\sqrt{\delta}}$ και $t = \frac{4}{\varepsilon^2}$ $\frac{1}{\delta \varepsilon^2}$

$$Pr\left[\left| \bar{Y} - \frac{1}{n+1} \right| \geq \varepsilon \frac{1}{(n+1)} \right] \leq \frac{1}{4}$$

Άρα με πιθανότητα $3/4$

$$\frac{1}{n+1}(1-\varepsilon) \leq \bar{Y} \leq (1+\varepsilon)\frac{1}{n+1}$$

$$\hat{n} = \frac{1}{\bar{Y}} - 1 \qquad \frac{n+1}{1+\varepsilon} \leq \frac{1}{\bar{Y}} \leq \frac{n+1}{1-\varepsilon}$$

[Flajolet - Martin algorithm]

Πιθανότητα $1-\delta$

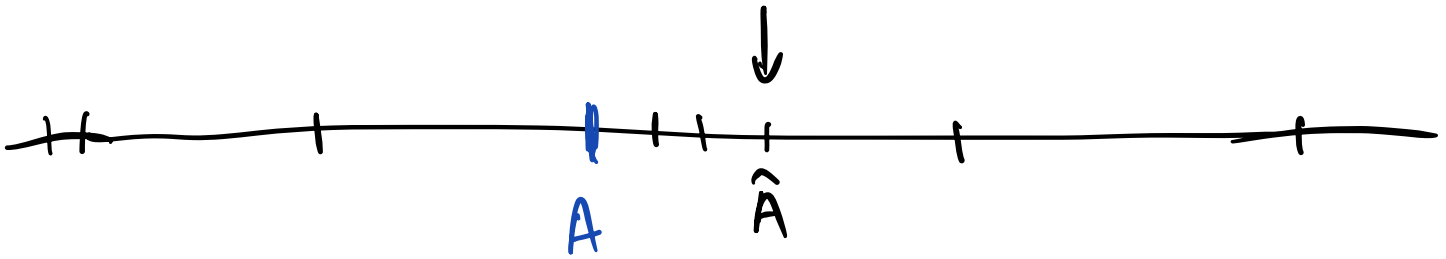
Γενική τεχνική για boosting της πιθανότητας ορθής εκτίμησης.

Έστω ότι έχω έναν εκτιμητή $\hat{A}$ ενός μεγέθους A ώστε $|\hat{A}-A| \leq \varepsilon$ με πιθανότητα $3/4$

Μπορώ να αυξήσω την πιθανότητα με ανεξάρτητα πειράματα του εκτιμητή

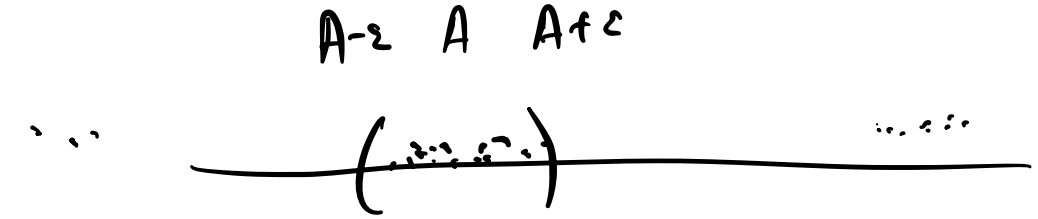
$$\hat{A}_1, \hat{A}_2, \dots, \hat{A}_m$$

$$\text{median}(\hat{A}_1, \dots, \hat{A}_m)$$



Claim: Αν πάνω από $\frac{m}{2}$ των $\hat{A}_i$ ικανοποιούν $|\hat{A}_i - A| \le \varepsilon$

τότε $|\text{median}(\hat{A}_1, \dots \hat{A}_m) - A| \le \varepsilon$



Έστω Z_i είναι η δείκτρια τμ. του γεγονότος ότι το $\hat{A}_i$ είναι καλό. Δηλαδή:

$$Z_i = \begin{cases} 1 & \text{αν } |\hat{A}_i - A| > \varepsilon \\ 0 & \text{αλλιώς} \end{cases}$$

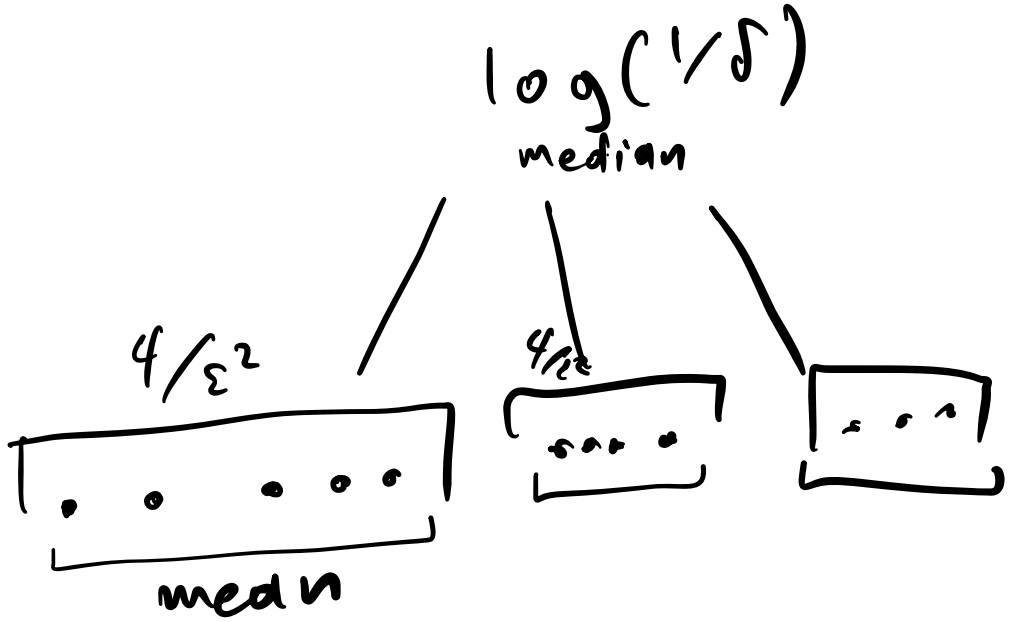
$$E[Z_i] \leq \frac{1}{4}$$

$$Pr\left[\sum_{i=1}^{m} Z_i \geq \frac{m}{2}\right]$$

$$= Pr\left[\left|\sum_{i=1}^{n} Z_i - \frac{m}{4}\right| \geq \frac{m}{4}\right]$$

$$\leq e^{-\Theta(m)} \leq \delta$$

$$m = \Theta(\log 1/\delta)$$



$O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$

Median - of - Means