Tuesday, February 3

- Reading available for every lecture
- Typos/mistakes now have to be submitted <u>after</u> class
- Research talk Friday 11-12:15 in Davidson Lecture Hall

Plan

- Use variance for stronger concentration inequalities
- Estimate number of distinct elements

Markov's Inequality

Let X be non-negative rv, $\epsilon > 0$

$$\Pr(X > \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon}$$

Comparison:

- Chebyshev's applies to any rv with bounded variance
- Chebyshev's is two-sided

But bounding variance is harder than bounding expectation

Chebyshev's Inequality

Let X be rv with variance $\sigma^2 = Var(X)$, $k > 0$

$$\Pr(|X - \mathbb{E}[X]| \geq k\sigma) \leq \frac{1}{k^2}$$

Proof: $S = (X - \mathbb{E}[X])^2$

By Markov's,

$$\Pr(S \geq t) \leq \frac{\mathbb{E}[S]}{t}$$

Note: $\mathbb{E}[S] = \mathbb{E}[(X - \mathbb{E}[X])^2] = Var(X) = \sigma^2$

Set $t = k^2\sigma^2$,

$$\Pr((X - \mathbb{E}[X])^2 \geq k^2\sigma^2) \leq \frac{\sigma^2}{k^2\sigma^2}$$

$\Leftrightarrow$

$$\Pr(|X - \mathbb{E}[X]| \geq k\sigma) \leq \frac{1}{k^2}$$

Independence

tool for analyzing variance

Consider $X_1, \ldots, X_m$

Pairwise indep if, for $i \neq j$,

$$\Pr(X_i = v_i, X_j = v_j) = \Pr(X_i = v_i) \Pr(X_j = v_j)$$

k-wise indep if, for all $1, \ldots, k$,

$$\Pr(X_1 = v_1, \ldots, X_k = v_k) = \Pr(X_1 = v_1) \cdot \ldots \cdot \Pr(X_k = v_k)$$

Linearity of Variance

For pairwise indep. $X_1, \ldots, X_m$

$$\mathrm{Var}\left(\sum_{i=1}^{m} X_i\right) = \sum_{i=1}^{m} \mathrm{Var}(X_i)$$

Q: Can you think of three variables that are 2-wise indep but not 3-wise?

Coin Example

$C_1, \ldots, C_{100}$

$$C_i = \begin{cases} 1 & \text{wp } 1/2 \\ 0 & \text{else} \end{cases}$$

$$C = \sum_{i=1}^{100} C_i$$

What's the probability that $C \geq 70$?

- Using Markov's
- Chebyshev's
- Exact distribution

Distinct Elements

Data arrives in a stream,
how many unique elements?

$x_1, \ldots, x_n \in \mathcal{U}$

$D =$ # distinct elements

For example,

Input: 1, 10, 2, 4, 9, 2, 10, 4

Output: $D = 5$

Applications:

- webpage visitors
- queries to search engine
- motifs in DNA

HyperLogLog used at all the big tech companies

Naive Attempt: Hash map, $O(d)$ space

Our Goal: Return $\hat{D}$ s.t.

$$(1-\epsilon) D \leq \hat{D} \leq (1+\epsilon) D$$

with $O(1/\epsilon^2)$ space, basically* independent of D

Algorithm

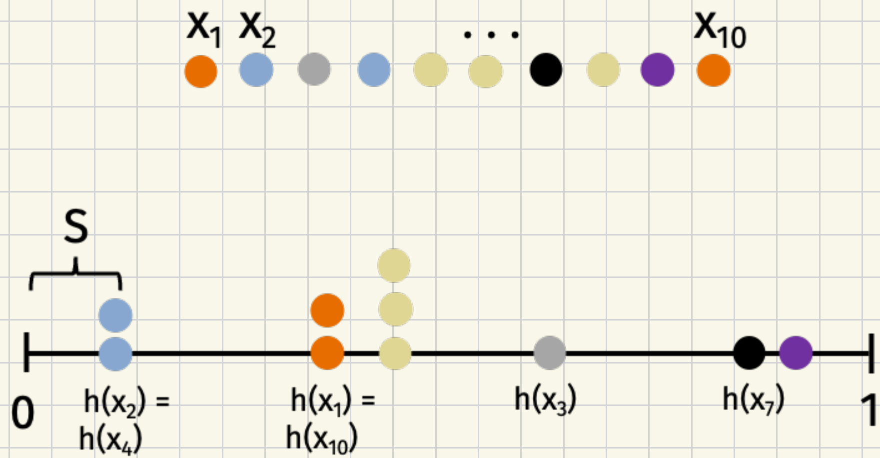
$h: \mathcal{U} \to [0,1]$

$S \leftarrow 1$

For $x_1, \dots, x_n$,

$S \leftarrow \min(S, h(x_i))$

$\hat{D} \leftarrow \frac{1}{S} - 1$



Intuition: S gets smaller as D gets larger

Why $\hat{D} = \frac{1}{S} - 1$?

Implies: $S = \frac{1}{\hat{D}+1}$

Note: We pretend h maps to real numbers. In practice, use discrete values but continuous case is easier to analyze.

Lemma: $\mathbb{E}[S] = \frac{1}{D+1}$

Calculus Proof:

$$X = \int_{x=0}^{X} dx = \int_{x=0}^{\infty} \mathbb{1}[X \geq x]\, dx \overset{\substack{\text{linearity of}\\ \text{expectation}}}{\Rightarrow} \mathbb{E}[X] = \int_{x=0}^{\infty} \Pr(X \geq x)\, dx$$

$$\mathbb{E}[S] = \int_{s=0}^{1} \Pr(S \geq s)\, ds = \int_{s=0}^{1} (1-s)^{D}\, ds = \left. \frac{-(1-s)^{D+1}}{D+1} \right|_{s=0}^{1} = \frac{1}{D+1}$$

Thursday, February 5

- Shapley values for causal inference
  - ↳ tomorrow 11am @ Davidson Lecture Hall
- GEMS Mentors!
  - ↳ Saturday 9:30am @ Shanahan
    (Please <u>email</u> me)

Plan

Apply Chebyshev's to distinct elements

Algorithm

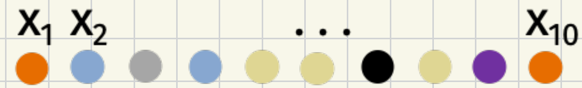
$h : U \to [0,1]$

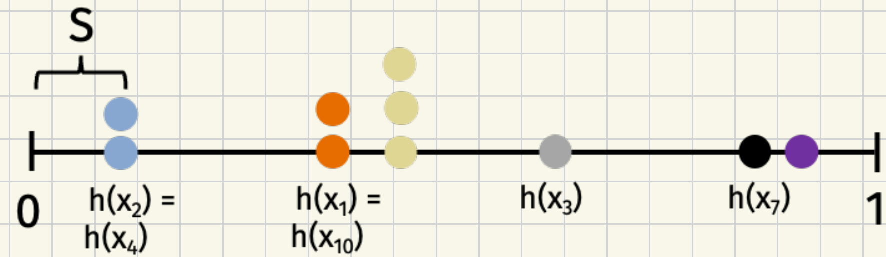
$S \leftarrow 1$

For $x_1, \ldots, x_n$,

$\quad S \leftarrow \min(S, h(x_i))$

$\hat{D} \leftarrow \frac{1}{S} - 1$





Plan: Bound S

☑ $\mathbb{E}\, S = \frac{1}{D+1}$

☐ $\mathbb{E}[S^2] = \frac{2}{(D+1)(D+2)}$

Chebyshev's

RV X with $\sigma^2 = \mathrm{Var}(X)$, $\epsilon > 0$

$$\Pr(|X - \mathbb{E} X| \geq \sigma \epsilon) \leq \frac{1}{\epsilon^2}$$

Lemma: $\mathbb{E}[S] = \frac{1}{D+1}$

Proof "from the Book":

$$\mathbb{E}_x\left[\Pr(A|x)\right] = \mathbb{E}_x\left[\,\mathbb{E}\left[\mathbb{1}[A]\,|\,x\right]\right] = \mathbb{E}_x\left[\mathbb{1}[A]\right] = \Pr(A)$$

$$S = \Pr\left(h_{D+1} \le \min_{i\in[D]} h_i \,\middle|\, h_1,\dots,h_D\right)$$

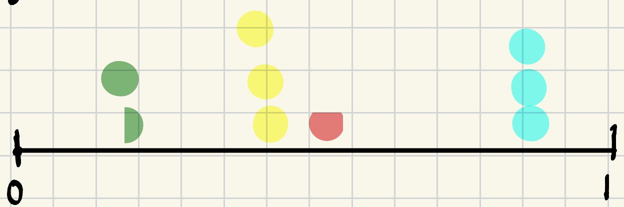
$$\mathbb{E}_{h_1,\dots,h_D}[S] = \mathbb{E}_{h_1,\dots,h_D}\left[\Pr\left(h_{D+1} \le \min_{i\in[D]} h_i \,\middle|\, h_1,\dots,h_D\right)\right]$$

$$= \Pr_{h_1,\dots,h_{D+1}}\left(h_{D+1} \le \min_{i\in[D]} h_i\right)$$

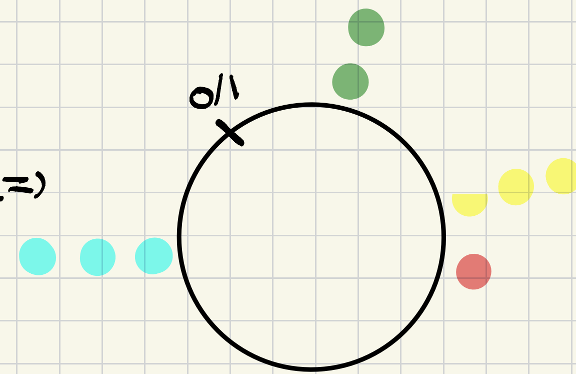
$$= \frac{1}{D+1}$$

Lemma: $\mathbb{E}[S] = \frac{1}{D+1}$

Larry's Proof:



$\Leftrightarrow$



D+1 items distributed uniformly* on circle,

By symmetry, $\frac{1}{D+1}$ distance between them in expectation

* 0/1 fixed but hashing is rotationally invariant so appears uniformly random

Know $\mathbb{E}[S]$, we also need $\mathrm{Var}(S)$ for Chebyshev's

Lemma: $\mathbb{E}[S^2] = \dfrac{2}{(D+1)(D+2)}$

Calculus Proof: Hint $\displaystyle\int_{s=0}^{1} (1-\sqrt{s})^D \, ds = \frac{2}{(D+1)(D+2)}$

Proof from the Book:

$$\mathbb{E}[S] = \frac{1}{D+1} \stackrel{\Delta}{=} \mu \qquad \mathbb{E}[S^2] = \frac{2}{(D+1)(D+2)}$$

$$\begin{aligned} \mathrm{Var}(S) &= \mathbb{E}[S^2] - \mathbb{E}[S]^2 \\ &= \frac{2}{(D+1)(D+2)} - \frac{1}{(D+1)^2} \leq \frac{1}{(D+1)^2} = \mu^2 \end{aligned}$$

By Chebyshev,

$$\Pr(|S-\mu| \geq \epsilon\mu) \leq \frac{1}{\epsilon^2}$$

vacuous!

Motif: Lower failure probability by repeating subroutine!

Repeat k times!

$h_1, \dots, h_k : \mathcal{U} \to [0,1]$

$S_1, \dots, S_k \leftarrow 1$

For $i = 1, \dots, n$:

For $j = 1, \dots, k$:

$S_j \leftarrow \min(S_j, h_j(x_i))$

$\bar{S} \leftarrow \frac{1}{k} \sum_{j=1}^{k} S_j$

$\hat{D} \leftarrow \frac{1}{\bar{S}} - 1$

$$\mathbb{E}[\bar{S}] = \mathbb{E}\left[\frac{1}{k}\sum_{j=1}^{k} S_j\right] = \frac{1}{k}\sum_{j=1}^{k} \mathbb{E}[S_j] = \mu$$

$$\mathrm{Var}(\bar{S}) = \mathrm{Var}\left(\frac{1}{k}\sum_{j=1}^{k} S_j\right) = \frac{1}{k^2}\sum_{j=1}^{k} \mathrm{Var}(S_j) = \frac{1}{k}\sigma^2$$

Now, $\mathbb{E}[\bar{S}] = \mu$, $\mathrm{Var}(\bar{S}) \leq \frac{\mu^2}{k}$

By Chebyshev, $\Pr\left(|\bar{S} - \mu| \geq c\frac{\mu}{\sqrt{k}}\right) \leq \frac{1}{c^2}$

Choose $c = \frac{1}{\sqrt{\delta}}$, $k = \frac{1}{\epsilon^2 \delta}$

$$\Pr(|\bar{S} - \mu| \geq \epsilon\mu) \leq \delta$$

$\Leftrightarrow$

$$\mu - \epsilon\mu \leq \bar{S} \leq \mu + \epsilon\mu \quad \text{wp } 1-\delta$$

$$(1-\epsilon)\mu \leq \bar{S} \leq (1+\epsilon)\mu$$

$$\frac{1}{(1+\epsilon)\mu} \leq \frac{1}{S} \leq \frac{1}{(1-\epsilon)\mu}$$

$1-2\epsilon \leq \frac{1}{1+\epsilon}$; $\frac{1}{1-\epsilon} \leq 1+2\epsilon$ by Desmos

$$(1-2\epsilon)\frac{1}{\mu} - 1 \leq \frac{1}{S} - 1 \leq (1+2\epsilon)\frac{1}{\mu} - 1$$

$$(1-4\epsilon)\left(\frac{1}{\mu} - 1\right) \leq \hat{D} \leq (1+4\epsilon)\left(\frac{1}{\mu} - 1\right)$$

$$(1-4\epsilon)D \leq \hat{D} \leq (1+4\epsilon)D$$

$(1-\epsilon)D \leq \hat{D} \leq (1+\epsilon)D$ wp $1-\delta$ in $O\left(\frac{\log D}{\epsilon^2 \delta}\right)$ space

Advantage: Easy to implement in distributed setting!

Hyper Log Log:

- Discrete hashing
- Harmonic mean to reduce outliers
- other tricks :)