Tuesday, February 3

- Reading available for every lecture
- Typos / mistakes now have to be submitted after 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. $X_1, \ldots, X_m$

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

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

Coin Example

$C_1, \dots, C_{100}$

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

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

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?

$x_1, \ldots, x_n \in U$

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

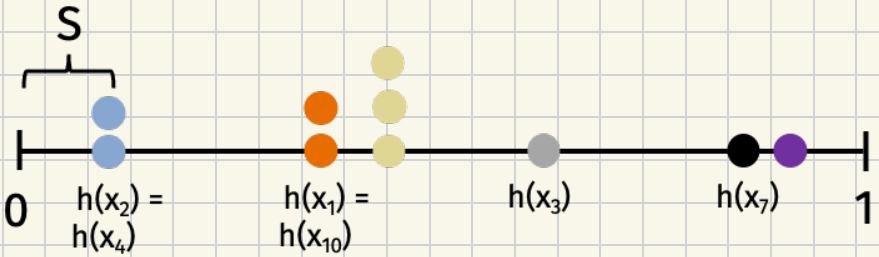
$h : U \to [0,1]$

$S \leftarrow 1$

For $x_1, \ldots, x_n$,

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

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



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

Calculus Proof:

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

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

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

Proof "from the Book":

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

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

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

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

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

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

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

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

Proof from the Book: