Probability Theory

Concentration Inequalities and Applications

Dinh, Nhi

Outline

Concentration Bound

Foundational Inequalities

Concentration Inequalities for Independent Random Variables

High-Dimensional Concentration Inequalities

Section 1

Concentration Bound

Distribution of Expected Value

The expected value E[X] of a random variable X is not a guarantee that each trials of X will be near that value.

- Uniform Distribution Over a Wide Range
- Heavy-Tailed Distributions (such as the Cauchy or certain Pareto distributions)

So, Expected Value is very useless since we don't know what distribution we dealing with oftenly?

Concentration bounds

Concentration bounds quantify how "concentrated" a random variable is around its expected value.

- In complex systems, such as networks or large datasets, these inequalities allow us to ensure that the behaviors we observe are not just artifacts of randomness.
- These bounds are widely used in machine learning, statistics, combinatorics, and computer science to ensure that algorithms and statistical estimates are reliable.

Section 2

Foundational Inequalities

Foundational Inequalities

- Markov's Inequality (linear in the reciprocal of the threshold)
- Chebyshev's Inequality (based on variance, leading to a bound that polynominaly decay in the deviation)

They are really weak :(

MCQ Problem

You do an MCQ True/False. 500 questions and I have bad luck with 0.2 right.

For $X \sim \text{Bern}(500, 0.2)$, the expected value is given by

$$E[X] = 500 \times 0.2 = 100.$$

However, passed score is 130. So I need 30 over the expected value. What is the probability that I can pass?

Markov Bound

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

$$\text{If } a = \mathbb{E}[X] + t, \text{ then } \mathbb{P}\big(X - \mathbb{E}[X] \geq t\big) \leq \frac{\mathbb{E}[X]}{t + \mathbb{E}[X]}.$$

$$\text{So } \mathbb{P}\big(X - 100 \geq 30\big) \leq \frac{100}{130} = 0.769.$$

There still 0.769 chances I passes.

Section 3

Concentration Inequalities for Independent Random Variables

Concentration Inequalities for Independent Random Variables

Involving moment-generating functions, to get much stronger results.

- McDiarmid's Inequality (if satisfy a bounded difference condition,)
- Chernoff Bounds (if Sums of Bernoulli Variables)

MCQ Problem

You do an MCQ True/False. 500 questions and I have bad luck with 0.2 right.

For $X \sim \text{Bern}(500, 0.2)$, the expected value is given by

$$E[X] = 500 \times 0.2 = 100.$$

However, passed score is 130. So I need 30 over the expected value. What is the probability that I can pass?

Chernoff Bound

Let $X_1, X_2, \ldots, X_n$ be *independent Bernoulli random variables* with parameter p

$$\Pr(X_i = 1) = p, \quad \Pr(X_i = 0) = 1 - p.$$

Define

$$X = \sum_{i=1}^{n} X_i$$

then

$$\Pr\Big(X \ - \ \mathbb{E}[X] \geq \delta \mathbb{E}[X]\Big) \leq \exp\left(-\frac{\delta^2\, \mathbb{E}[X]}{3}\right).$$

Chernoff Bound

Let $X_1, X_2, \ldots, X_n$ be *independent Bernoulli random variables* with parameter p

$$\Pr(X_i = 1) = 0.2, \quad \Pr(X_i = 0) = 0.8.$$

Define

$$X = \sum_{i=1}^{n} X_i$$

then

$$\Pr\Big(X \ - \ 100 \geq 30\Big) \leq \exp\left(-\frac{0.3^2\,100}{3}\right) = 0.0498$$

This seem more right?

McDiarmid's inequality

(or Hoeffding's)

Let $T_1, \ldots, T_n$ be independent random variables taking values in some set $\mathcal{T}$. Suppose there is a function $f : \mathcal{T}^n \to \mathbb{R}$:

$$\left| f(t_1, \ldots, t_n) \;-\; f(t_1, \ldots, t_{i-1}, t_i', t_{i+1}, \ldots, t_n) \right| \;\leq\; c_i$$

for $t_j = t_j'$ for every $j \neq i$.

Consider the random variable

$$X \;=\; f(t_1, \ldots, t_n)$$

Then, for any $t > 0$, McDiarmid's inequality states that

$$\Pr\big(X \;-\; \mathbb{E}[X] \;\geq\; t\big) \;\leq\; \exp\!\Big(-\frac{2\,t^2}{\sum_{i=1}^{n} c_i^2}\Big)\,.$$

McDiarmid Bound

Let $T_1, \ldots, T_n$ be independent random variables taking values in some set $\mathcal{T}$. Suppose there is a function $f : \mathcal{T}^n \to \mathbb{R}$:

$$\left| f(t_1, \ldots, t_n) \; - \; f(t_1, \ldots, t_{i-1}, t_i', t_{i+1}, \ldots, t_n) \right| \; \leq \; c_i = 1$$

for $t_j = t_j'$ for every $j \neq i$.

Consider the random variable

$$X \;=\; T_1 + \cdots + T_n$$

Then, for any $t > 0$, McDiarmid's inequality states that

$$\Pr\left(\left| X \; - \; 100 \right| \; \geq \; 30 \right) \; \leq \; 2 \exp\left(- \frac{2\,(30)^2}{\sum_{i=1}^{500} 1^2} \right) \; = 0.0546$$

Probably more right

Section 4

High-Dimensional Concentration Inequalities

What if they are all exponents?

In high-dimensional settings, many functions of independent random variables rarely deviate significantly from a central value. So we need an even better bound.

Talagrand's Inequality

Instead of simply counting individual changes, Talagrand's approach looks for the combined effect of changes across all variables required to shift a sample point X into a specific target set.

Talagrand's Inequality

(multiple version) Let X be a non-negative random variable (not identically zero) determined by n independent trials $T_1, T_2, \ldots, T_n$, and suppose there exist constants $c, r > 0$ such that:

1. Changing the outcome of any one trial can affect X by at most c.
2. For every s, if $X \geq s$, then there exists a set of at most rs trials whose outcomes certify that $X \geq s$.

Then, for any $0 \leq t \leq \mathbb{E}[X]$,

$$\Pr\Big(|X - \mathbb{E}[X]| > t + 60\, c\, \sqrt{r\, \mathbb{E}[X]}\Big) \;\leq\; 4 \exp\Big(-\frac{t^2}{8\, c^2\, r\, \mathbb{E}[X]}\Big).$$

Talagrand Inequality

Talagrand's inequalities have indeed found significant application in the study of graphs, particularly within the realm of probabilistic combinatorics and random graph theory.

Random Subgraph Problem

Random subgraphs are a powerful tool in graph theory and related fields because they enable researchers and practitioners to study complex networks and algorithms through probabilistic methods.

Random Subgraph Problem

- Let G be a graph with v vertices.
- Construct a random subgraph H by including each edge independently with probability p.
- Define the random variable X as the number of vertices that appear as endpoints of at least one edge in H.
- Goal: show that X is sharply (strongly) concentrated around its expected value.

Why cant Chernoff or Hoeffding?

- **Chernoff Bound:**
  - Typically applies to sums of independent Bernoulli random variables.
  - Here X is not a direct sum but a derived count (vertices activated by at least one edge).
- **Hoeffding's Inequality:**
  - Relies on the number of trials—here, there are roughly $O(v^2)$ independent edge decisions.
  - X is bounded by v, so the large number of underlying trials makes the Hoeffding bound too weak.

$$\Pr\left(X - \mathbb{E}[X] \geq t\right) \leq \exp\left(-\frac{2\,t^2}{4v^2}\right).$$

Talagrand's Bound

- **Lipschitz Condition:**
  - Flipping any one edge change X by at most 2 (already includes both vertices or have not included both)
  - Thus, c = 2
- **Certifiability (Witness) Condition:**
  - If $X \geq s$, then there exist at least s vertices which are activated.
  - For each such vertex, one can select a single incident edge that is present in H.
  - These s edges certify that $X \geq s$.
  - So r = 1

.

Talagrand's Bound

Consequently, one obtains a tail bound of the form

$$\Pr\left(|X - \mathbb{E}[X]| > t + 60 \cdot 2\sqrt{\mathbb{E}[X]}\right) \;\leq\; 4\exp\left(-\frac{t^2}{32\mathbb{E}[X]}\right)$$

This concentration is strong even though there are $O(v^2)$ independent edge trials, because the structure of X limits the impact of each individual trial.