Convergence

钟宛君　（18110955），黄羽盼（18110943）
Nov 1

Content

Part A: Sampling and Estimation

Part B: Probably Approximately Correct (PAC)

Part C: Concentration of Measure: Markov, Chebyshev Inequality and Chernoff-Hoeffding Inequality

Content

Part A: Sampling and Estimation

Part B: Probably Approximately Correct (PAC)

Part C: Concentration of Measure: Markov, Chebyshev Inequality and Chernoff-Hoeffding Inequality

Sampling and Estimation

- Data: Set of n data points $P = \{p_1, p_2, \ldots, p_n\}$
- Powerful Assumption: Data comes iid (Identically and Independent Distributed) from fixed, unknown PDF (Probability density function).
- Our goal is to estimate the underlying data distribution.

Sampling and Estimation

- What's the mean of PDF?

  Considering random variable X : X~f:

  mean of f is the expected value of X: $\mathrm{E}[\mathrm{X}]$

- How we estimate the mean of f? - By sample mean

$$\overline{P} = \frac{1}{n} \sum_{i=1}^{n} p_i$$

We can randomly sample n data points to estimate the mean of f by sample mean.

$$\bar{P} \underset{\frac{1}{n}\sum}{=} \{p_i\} \underset{\text{realize}}{\leftarrow} \{X_i\} \underset{\text{iid}}{\sim} f$$

Step 1: Select n iid variables $\{X_i\}$ corresponding to set of n independent observation $\{p_i\}$

Step 2: Take their average to estimate the mean of f.

Central Limit Theorem

- Goal: Estimate how well the sample mean approximates the true mean
- The sample mean is dependent to the data we select ($\{X_i\}$ is randomly selected). Therefore it's not precisely equal to the mean of f.
- Central Limit Theorem

Central Limit Theorem: Consider n iid random variables $X_1, X_2, \ldots, X_n$, where each $X_i \sim f$ for any fixed distribution f with mean μ and bounded variance σ^2. Then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges to the normal distribution with mean $\mu = \mathbf{E}[X_i]$ and variance σ^2/n.

Which means: The mean of random sample mean will converge to the normal distribution with mean $\mathrm{E}[X]$ as the observation increases.

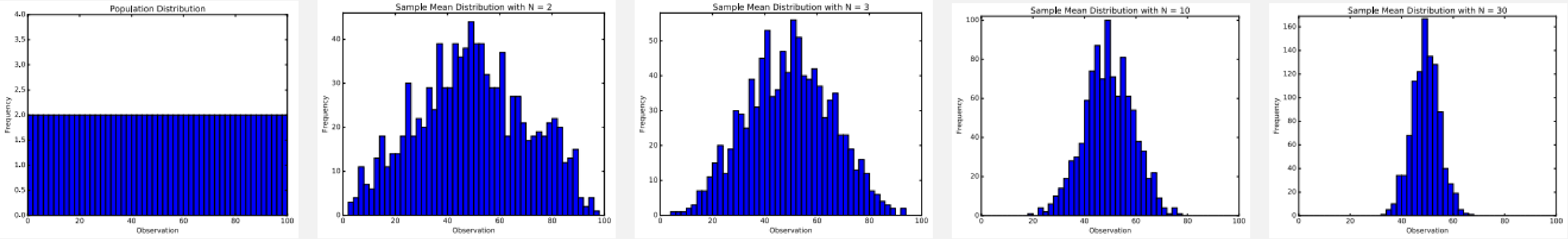
$\{p_i\}$

Central Limit Theorem

- Example

Example: Central Limit Theorem

Consider f as a uniform distribution over $[0, 100]$. If we create n samples $\{p_1, \ldots, p_n\}$ and their mean $\bar{P}$, then repeat this 1000 times, we can plot the output in histograms:



We see that starting at $n = 2$, the distributions look vaguely looks normal (in the technical sense of a normal distribution), and that their standard deviations narrow as n increases.

Central Limit Theorem Example

Rolling n dice, each governed by the same probability distribution. Each face of a dice has probability 1/6.

mean

$$\mu = E(x_i) = \frac{1+2+3+4+5+6}{6} = 3.5$$

variance

$$\frac{\sigma^2}{n} = \frac{(2.5^2 + 1.5^2 + 0.5^2) * 2}{6} \approx 2.92$$

standard deviation

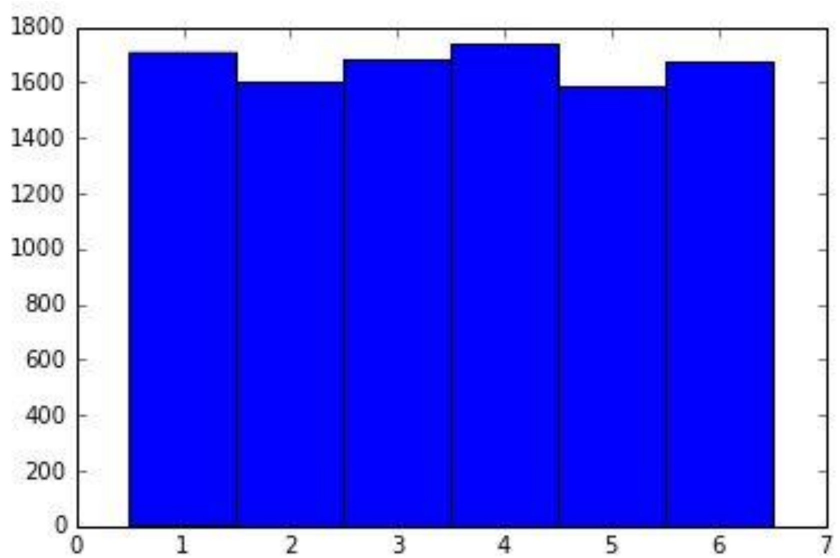
$$\sqrt{\frac{\sigma^2}{n}} \approx 1.71$$

Reference: https://zhuanlan.zhihu.com/p/25241653 and https://www.albany.edu/~jr853689/CentralLimitTheoremForDice.htm
You can do Dice Experiment on http://www.randomservices.org/random/apps/DiceExperiment.html

Central Limit Theorem Example

1. Rolling a dice for 10000 times.
   Draw the histogram of random data.

```
>>> import numpy as np
>>> random_data = np.random.randint(1, 7, 10000)
>>> random_data.mean()
'3.5199'
>>> random_data.std()
'1.7116085971973851'
```

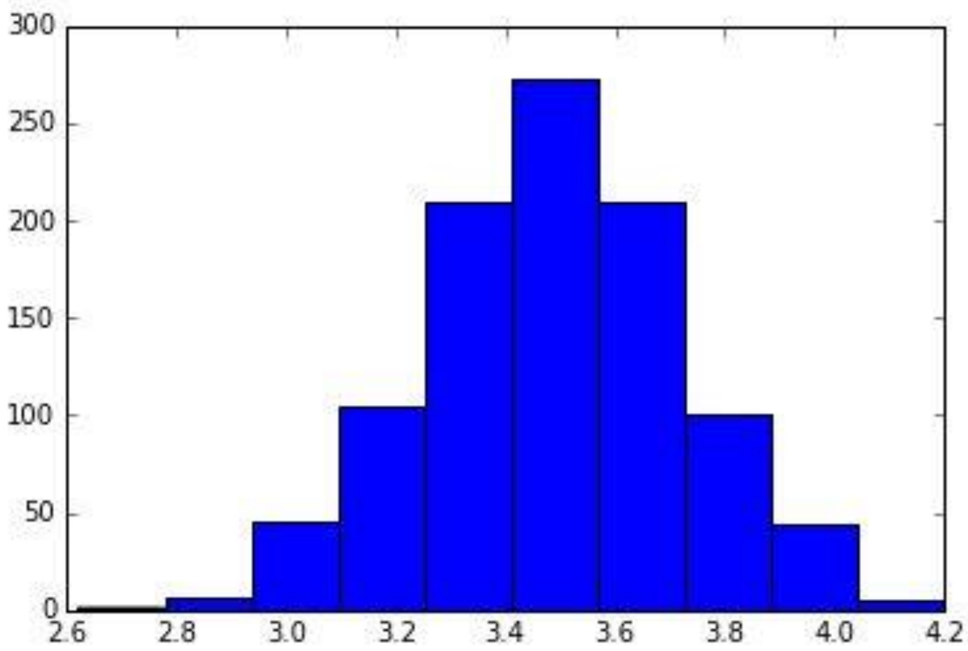


2. Choose ten scores from the generated data randomly and calculate their mean and variance.

```
[>>> sample1 = []
[>>> for i in range(0, 10):
[...     sample1.append(random_data[int(np.random.random() * len(random_data))])
[...
[>>> np.mean(sample1)
 '2.5'
[>>> np.std(sample1)
 '1.8027756377319946'
```

3. Now choose 1000 groups, each group has 50 samples. Draw the histogram of the 1000 mean scores.

```
>>> samples_mean = []
>>> samples_std = []
>>>
>>> for i in range(0, 1000):
...     sample = []
...     for j in range(0, 50):
...         sample.append(random_data[int(np.random.random() * len(random_data))])
...     samples_mean.append(np.mean(sample))
...     samples_std.append(np.std(sample))
...
>>> np.mean(samples_mean)
'3.51656'
>>> np.mean(samples_std)
'1.692681775503283'
```



- population average: 3.5 is the "average of all possible rolls of a fair die."
- The output and data distribution illustrate the Central Limit Theorem

Remaining Mysteries

- What does convergence mean?

Convergence refers to what happens as some parameter increases. (eg. n goes to infinity then the distribution will be more precise.)

- How we formalize the error when estimating?

The distance between $\overline{P}$ and μ is more than ε, with probability at most δ .

We call that "probably approximately correct" (PAC).

- How we describe the distribution of f more detailedly?

We discuss some very common cencentration of measure tools - To provide the upper bounds to state the PAC bounds.

Content

Part A: Sampling and Estimation

Part B: Probably Approximately Correct (PAC)

Part C: Concentration of Measure : Markov, Chebyshev Inequality and Chernoff-Hoeffding Inequality

Probably Approximately Correct (PAC)

- Concentration of measure bounds

Introduce three most common concentration of measure bounds, provide increasingly strong bounds, but requires increasing information about underlying f

a. Markov Inequality

b. Chebyshev Inequality

c. Chernoff-Hoeffding Inequality

- Basic form of PAC bound

$$\mathbf{Pr}[|\bar{X} - \mathbf{E}[\bar{X}]| \geq \varepsilon] \leq \delta.$$

Content

Part A: Sampling and Estimation

Part B: Probably Approximately Correct (PAC)

Part C: Concentration of Measure : Markov, Chebyshev Inequality and Chernoff-Hoeffding Inequality

Markov Inequality

➢ Theorem

Let X be a non-zero, random variable, and a>0 then

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

Or equivalently $\Pr\big[|X| \geq aE[|X|]\big] \leq \frac{1}{a}$.

PAC bound with $\in = a - E[|X|]$ and $\delta = E[|X|]$/a. Then the bound can be rephrased as:

$$\Pr[|X - E[|X|]| \geq \in] \leq \frac{E[|X|]}{\in + E[|X|]}$$

Reference: https://www.cs.purdue.edu/homes/egrigore/Fall12/lect2.pdf

Markov Inequality

- Conclusion: It provide weak bounds but it only requires only expected value $E[|X|]$ of f
- Example

For the toss of n fair coins let X_i denote the event that the i^{th} coin lands heads. Then $E[\sum X_i] = \frac{n}{2}$ so the probability that more than 2/3's of the coins come up heads is

$$\Pr\left[\frac{2n}{3}\ come\ up\ heads\right] <= \frac{n/2}{2n/3} = \frac{3}{4}$$

Reference: https://www.cs.purdue.edu/homes/egrigore/Fall12/lect2.pdf

Chebyshev's Inequality

➢ Theorem

$$\Pr[|X - E[X]| \geq \epsilon] \leq \frac{Var[|X|]}{\epsilon^2}$$

where $Var[X] = E[(X - E[X])^2]$. The bound is $\delta = \frac{Var[|X|]}{\epsilon^2}$.

➢ This bound is typically stronger than the Markov because δ decreases quadratically in ϵ instead of linearly.

➢ Conclusion: Compared with Markov inequality, it has two property:

1. Have stronger PAC bounds, because δ decreases quadratically in ϵ .
2. Support negative value of X by square function.

Chebyshev's Inequality

- **Proof** Let random variable $Y = |X - E[X]|$. Using Markov's inequality we have that

$$\Pr[Y \geq a\ ] = Pr[Y^2 \geq a^2\ ] \leq \frac{E[Y^2]}{a^2} = \frac{Var[X]}{a^2}$$

- Example

Example: Chebyshev for IID Samples

Recall that for an average of random variables $\bar{X} = (X_1 + X_2 + \ldots + X_n)/n$, where the X_is are iid, and have variance σ^2, then $\mathbf{Var}[\bar{X}] = \sigma^2/n$. Hence

$$\mathbf{Pr}[|\bar{X} - \mathbf{E}[X_i]| \geq \varepsilon] \leq \frac{\sigma^2}{n\varepsilon^2}.$$

Consider now that we have input parameters ε and δ, our desired error tolerance and probability of failure. If can draw $X_i \sim f$ (iid) for an unknown f (with known expected value and variance σ), then we can solve for how large n needs to be: $n = \sigma^2/(\varepsilon^2\delta)$.

Chernoff/Hoeffding Inequality

- Theorem

Let $X_1, \dots, X_n$ be independent random variables in the interval [0, 1] and let $X = \sum X_i$. Then

$$\Pr[|X - E[X]| \geq \mathrm{t}\,] \leq 2\exp(-2t^2/n)$$

Equivalently

$$\Pr[|X - E[X]| \geq \epsilon \mathrm{E[X]}\,] \leq 2\exp(-2\epsilon^2 E[X]^2/n)$$

and

$$\Pr[|X - E[X]| \geq \epsilon \mathrm{n}\,] \leq 2\exp(-2\epsilon^2 n)$$

- The bound is

$$\delta = 2\exp(-2\epsilon^2 n)$$

For desired error tolerance ϵ and failure probability δ, we can set

$$n = \left(\frac{1}{2\epsilon^2}\right)\ln\left(\frac{2}{\delta}\right)$$

Although this has similar relationship with ϵ and δ, but the dependence of n on δ is exponentially less for this bound.

Chernoff-Hoeffding Inequality

- Consider set of iid random variables $X_1, X_2, \ldots, X_n$ where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Now we assume each X_i lies in a bounded domain $[b,t]$.

$$\mathbf{Pr}[|\bar{X} - \mathbf{E}[\bar{X}]| > \varepsilon] \leq 2\exp\left(\frac{-2\varepsilon^2 n}{\Delta^2}\right)$$

The bound is

$$\delta = 2\exp\left(\frac{-2\varepsilon^2 n}{\Delta^2}\right)$$

For desired error tolerance ε and failure probability δ , we can set

$$n = (\Delta^2/(2\varepsilon^2))\ln(2/\delta)$$

Although this has similar relationship with ε and δ , but the dependence of n on δ is exponentially less for this bound.

The Derivation of Three Basic Inequality

Suppose that $\mathbf{Z}$ has a finite mean and that $\mathbb{P}(\mathbf{Z} \geq \mathbf{0}) = \mathbf{1}$. Then, for any $\epsilon > \mathbf{0}$,

$$\mathbb{E}(\mathrm{Z}) = \int_0^\infty z dP(\mathrm{z}) \geq \int_\epsilon^\infty z dP(\mathrm{z}) \geq \epsilon \int_\epsilon^\infty dP(\mathrm{z}) = \epsilon \mathbb{P}(\mathrm{Z} > \epsilon)$$

which yields *Markov's inequality*:

$$\mathbb{P}(\mathrm{Z} > \epsilon) \leq \frac{\mathbb{E}(\mathrm{Z})}{\epsilon}$$

An immediate consequence of Markov's inequality is *Chebyshev's inequality*

$$\mathbb{P}(|\mathrm{Z} - \mu| > \epsilon) = \mathbb{P}(|\mathrm{Z} - \mu|^2 > \epsilon^2) \leq \frac{\mathbb{E}(\mathrm{Z} - \mu)^2}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}$$

Where $\mu = \mathbb{E}(\mathrm{Z})$ and $\sigma^2 = Var(Z)$.

Reference: http://www.stat.cmu.edu/~larry/=sml/Concentration.pdf

The Derivation of Three Basic Inequality

If $\mathbf{Z_1, \ldots, Z_n}$ are iid with mean μ and variance $\boldsymbol{\sigma^2}$ then, since $\mathbf{Var}(\overline{\mathrm{Z_n}}) = \frac{\sigma^2}{n}$, Chebyshev's inequality yields

$$\mathbb{P}(|\overline{\mathrm{Z_n}} - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

While this inequality is useful, it does not decay exponentially fast as $\mathbf{n}$ increases.

To improve the inequality, we use *Chernoff 's method*: for any t > 0,

$$\mathbb{P}(\mathrm{Z} > \epsilon) = \mathbb{P}(\mathrm{e^Z} > \mathrm{e}^\epsilon) = \mathbb{P}(\mathrm{e^{tZ}} > \mathrm{e}^{t\epsilon}) \leq \mathrm{e}^{-\mathrm{t}\epsilon}\mathbb{E}(\mathrm{e^{tZ}})$$

Chernoff-Hoeffding Inequality If $\mathbf{Z_1, \ldots, Z_n}$ are independent with $\mathbb{P}(\mathbf{a} \leq \mathrm{Z_i} \leq$

Conclusion

	Markov	**ChebyShev**	**Chernoff-Hoeffding**
Strength of PAC bound	Weak (Linealy)	Meduim (Quadratically)	Strong (Exponentially)
Information requires	E[x]	E[x] and Var[x]	Bound of X
Value of X	Larger than zero	Arbitrary	Arbitrary

Application 1: Approximating the fraction of 1's in a binary string

Suppose we want to estimate the fraction of 1's in a given string $\mathbf{S} \subset \{0, 1\}^{\mathbf{n}}$.

That is, we wish to find a fast randomized algorithm that, given ϵ and string $\mathbf{S}$ outputs a value $\mathbf{V}$ such that $|\mathbf{V} - \textbf{fraction of 1's}| < \epsilon$ with probability $\mathbf{2/3}$.

Algorithm. Pick $\mathbf{k} = 1/\epsilon^1$ uniformly random indices in the string $\mathbf{S}$ and output the fraction of 1's in the sample.

Analysis. Let $X_1, \dots, X_k$ be random variables indicating if a 1 was found in the string position for the $\mathbf{i}^{\mathbf{th}}$ index selected $(1 \le \mathbf{i} \le \mathbf{k})$. Then by Chernoff's bound

$$\Pr\left[\left|\frac{X}{k} - \frac{E[X]}{k}\right| > \epsilon\right] \le 2\mathrm{e}^{-2\epsilon^2 k} = 2e^{-2} < \frac{1}{3}$$

as $\epsilon^2 k = 1$.

meaning that we output a good estimate (i.e. within ϵ from the true fraction of 1's in the string) with probability > 2/3.

Application 2: Improving a random algorithm's correctness

Suppose we are given a randomized algorithm A which on each input x from some domain D outputs a 0 or 1 answer and it is correct with probability p = 2/3. Let algorithm B run A for t times and output the majority answer.

Show that algorithm B is correct (on each input) with probability greater than $1 - 2^{-ct}$ for some constant c (that is, $\forall x \in D, \Pr[B(x) = f(x)] \geq 1 - 2^{-ct}$.)

Analysis. Let X1, . . ., Xt be indicator variables such that Xi = 1 if A outputs the correct answer in the i^{th} step. Therefore, $E[X_i] = p = 2/3$. Set $X = \sum X_i$, that is X is the random variable counting the number of correct answers, and notice that $E[X] = 2t/3$.

$$\begin{aligned}
&\Pr[\text{B outputs incorrect answer}]\\
&= \Pr\left[\text{A outputs incorrect answer more than } \frac{t}{2}\ times\right] = \Pr\left[X < \frac{t}{2}\right]\\
&\leq \Pr\left[X - \frac{2t}{3} < \frac{t}{2} - \frac{2t}{3}\right]\\
&= \Pr\left[X - \frac{2t}{3} < -\frac{t}{6}\right] \leq \Pr\left[\left|X - \frac{2t}{3}\right| > \frac{t}{6}\right]\\
&\leq 2e^{-\frac{2t^2\left(\frac{1}{6}\right)^2}{t}} = 2^{-ct}
\end{aligned}$$

Example

Example: Uniform Distribution

Consider a random variable $X \sim f$ where $f(x) = \{\frac{1}{2}$ if $x \in [0, 2]$ and 0 otherwise.$\}$, i.e, the Uniform distribution on $[0, 2]$. We know $\mathbf{E}[X] = 1$ and $\mathbf{Var}[X] = \frac{1}{3}$.

- Using the Markov Inequality, we can say $\mathbf{Pr}[X > 1.5] \leq 1/(1.5) \approx 0.6666$ and $\mathbf{Pr}[X > 3] \leq 1/3 \approx 0.33333$.
  or $\mathbf{Pr}[X - \mu > 0.5] \leq \frac{2}{3}$ and $\mathbf{Pr}[X - \mu > 2] \leq \frac{1}{3}$.
- Using the Chebyshev Inequality, we can say that $\mathbf{Pr}[|X - \mu| > 0.5] \leq (1/3)/0.5^2 = \frac{4}{3}$ (which is meaningless). But $\mathbf{Pr}[|X - \mu| > 2] \leq (1/3)/(2^2) = \frac{1}{12} \approx 0.08333$.

Now consider a set of $n = 100$ random variables $X_1, X_2, \ldots, X_n$ all drawn iid from the same pdf f as above. Now we can examine the random variable $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. We know that $\mu_n = \mathbf{E}[\bar{X}] = \mu$ and that $\sigma_n^2 = \mathbf{Var}[\bar{X}] = \sigma^2/n = 1/(3n) = 1/300$.

- Using the Chebyshev Inequality, we can say that $\mathbf{Pr}[|\bar{X} - \mu| > 0.5] \leq \sigma_n^2/(0.5)^2 = \frac{1}{75} \approx 0.01333$, and $\mathbf{Pr}[|\bar{X} - \mu| > 2] \leq \sigma_n^2/2^2 = \frac{1}{1200} \approx 0.0008333$.
- Using the Chernoff-Hoeffding bound, we can say that $\mathbf{Pr}[|\bar{X} - \mu| > 0.5] \leq 2\exp(-2(0.5)^2 n/\Delta^2) = 2\exp(-100/8) \approx 0.0000074533$, and $\mathbf{Pr}[|\bar{X} - \mu| > 2] \leq 2\exp(-2(2)^2 n/\Delta^2) = 2\exp(-200) \approx 2.76 \cdot 10^{-87}$.

Thanks!

Wanjun Zhong, Yupan Huang