COL866: Foundations of Data Science

Ragesh Jaiswal, IITD

Administrative Information

Administrative Information

- Course Instructor:
  - Ragesh Jaiswal
  - *Email*: `rjaiswal@cse.iitd.ac.in`
  - Office: SIT 403
- Course Time/Place:
  - Lectures: TBD
- Teaching Assistants: TBD

Administrative Information

- Grading Scheme
  1. *Homework* + *Quiz*: 20%
  2. *Minor*: 40% (two minors 20% each)
  3. *Major*: 40%
- Homework and Quizzes:
  - Gradescope: A paperless grading system. Use the course code **948VG9** to register. Please use your formal email address from IIT Delhi.
- Policy on cheating: Students using unfair means will be severely penalised.

Administrative Information

- Textbooks: We will follow this book available online.
  1. Foundations of Data Science by *Avrim Blum, John Hoproft, and Ravindran Kannan*.
- Course webpage: `http://www.cse.iitd.ac.in/~rjaiswal/2017/COL866/`.
  - The site will contain course information, references, homework, course slides etc. Please check this page regularly.

Introduction

- Why is a new foundational course in Computer Science required?
- Why doesn't foundations in Discrete Mathematics, Data Structures, and Algorithms suffice for modern information processing?

Introduction

- Why is a new foundational course in Computer Science required?
- Why doesn't foundations in Discrete Mathematics, Data Structures, and Algorithms suffice for modern information processing?
- Modern context:
    - Beyond worst case
    - Big data

Introduction

- Why is a new foundational course in Computer Science required?
- Why doesn't foundations in Discrete Mathematics, Data Structures, and Algorithms suffice for modern information processing?
- Modern context:
  - Beyond worst case
  - Big data
    - High dimensional data

High Dimension Space

- Our intuition about two or three dimension space does not usually carry over to larger dimensions.
- For example:
  - The volume of a unit ball goes to zero as dimension goes to infinity.
  - The volume of a unit ball is concentrated near its *surface* and is also concentrated at its *equator*.
  - If one generates a random point in d-dimensional space using a Gaussian to generate coordinates independently, the distance between all pair of points will *mostly* be the same when d is large.
    - Gaussian distribution has probability density function:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

High Dimension Space

- Our intuition about two or three dimension space does not usually carry over to larger dimensions.
- For example:
  - The volume of a unit ball goes to zero as dimension goes to infinity.
  - The volume of a unit ball is concentrated near its *surface* and is also concentrated at its *equator*.
  - If one generates a random point in *d*-dimensional space using a Gaussian to generate coordinates independently, the distance between all pair of points will *mostly* be the same when *d* is large.
    - Gaussian distribution has probability density function:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

    - This follows from the law of large numbers.

Theorem (Law of large numbers)

Let $x_1, x_2, ..., x_n$ be n independent samples of a random variable x. Then

$$\mathbf{Pr}\left[\left|\frac{x_1 + x_2 + ... + x_n}{n} - \mathbf{E}(x)\right| \geq \varepsilon\right] \leq \frac{\mathbf{Var}(x)}{n\varepsilon^2}.$$

- We will require the following two simple inequalities from probability theory.

Theorem (Markov's inequality)

Let x be a non-negative random variable. Then for $a > 0$,
$\mathbf{Pr}[x \geq a] \leq \frac{\mathbf{E}(x)}{a}$.

Theorem (Chebychev's inequality)

Let x be a random variable. Then for $c > 0$,
$\mathbf{Pr}[|x - \mathbf{E}(x)| \geq c] \leq \frac{\mathbf{Var}(x)}{c^2}$.

Theorem (Law of large numbers)

Let $x_1, x_2, ..., x_n$ be n independent samples of a random variable x. Then

$$\mathbf{Pr}\left[\left|\frac{x_1 + x_2 + ... + x_n}{n} - \mathbf{E}(x)\right| \geq \varepsilon\right] \leq \frac{\mathbf{Var}(x)}{n\varepsilon^2}.$$

- We will require the following two simple inequalities from probability theory.

Theorem (Markov's inequality)

Let x be a non-negative random variable. Then for $a > 0$, $\mathbf{Pr}[x \geq a] \leq \frac{\mathbf{E}(x)}{a}$.

Theorem (Chebychev's inequality)

Let x be a random variable. Then for $c > 0$, $\mathbf{Pr}[|x - \mathbf{E}(x)| \geq c] \leq \frac{\mathbf{Var}(x)}{c^2}$.

- A few more equalities:
  1. For any r.v. x, y, $\mathbf{E}(x + y) = ?$.
  2. For any r.v. x and any constant c, $\mathbf{Var}(x - c) = ?$.
  3. For any r.v. x and any constant c, $\mathbf{Var}(cx) = ?$.
  4. For any independent r.v. x, y, $\mathbf{Var}(x + y) = ?$.

Theorem (Law of large numbers)

Let $x_1, x_2, ..., x_n$ be n independent samples of a random variable x. Then

$$\mathbf{Pr}\left[\left|\frac{x_1 + x_2 + ... + x_n}{n} - \mathbf{E}(x)\right| \geq \varepsilon\right] \leq \frac{\mathbf{Var}(x)}{n\varepsilon^2}.$$

- We will require the following two simple inequalities from probability theory.

Theorem (Markov's inequality)

Let x be a non-negative random variable. Then for $a > 0$, $\mathbf{Pr}[x \geq a] \leq \frac{\mathbf{E}(x)}{a}$.

Theorem (Chebychev's inequality)

Let x be a random variable. Then for $c > 0$, $\mathbf{Pr}[|x - \mathbf{E}(x)| \geq c] \leq \frac{\mathbf{Var}(x)}{c^2}$.

- A few more equalities:
  1. For any r.v. x, y, $\mathbf{E}(x + y) = \mathbf{E}(x) + \mathbf{E}(y)$.
  2. For any r.v. x and any constant c, $\mathbf{Var}(x - c) = \mathbf{Var}(x)$.
  3. For any r.v. x and any constant c, $\mathbf{Var}(cx) = c^2\mathbf{Var}(x)$.
  4. For any independent r.v. x, y, $\mathbf{Var}(x + y) = \mathbf{Var}(x) + \mathbf{Var}(y)$.

Theorem (Law of large numbers)

Let $x_1, x_2, ..., x_n$ be n independent samples of a random variable x. Then

$$\mathbf{Pr}\left[\left|\frac{x_1 + x_2 + ... + x_n}{n} - \mathbf{E}(x)\right| \geq \varepsilon\right] \leq \frac{\mathbf{Var}(x)}{n\varepsilon^2}.$$

Proof

- We have:

$$\begin{aligned}
\mathbf{Pr}\left[\left|\frac{x_1 + x_2 + ... + x_n}{n} - \mathbf{E}(x)\right| \geq \varepsilon\right] &\leq \frac{\mathbf{Var}\left(\frac{x_1 + x_2 + ... + x_n}{n}\right)}{\varepsilon^2} \\
&= \frac{1}{n^2\varepsilon^2} \cdot \mathbf{Var}(x_1 + ... + x_n) \\
&= \frac{1}{n^2\varepsilon^2} \cdot \mathbf{Var}(x_1) + ... + \mathbf{Var}(x_n) \\
&= \frac{\mathbf{Var}(x)}{n\varepsilon^2}.
\end{aligned}$$

Theorem (Law of large numbers)

Let $x_1, x_2, ..., x_n$ be n independent samples of a random variable x. Then

$$\mathbf{Pr}\left[\left|\frac{x_1 + x_2 + ... + x_n}{n} - \mathbf{E}(x)\right| \geq \varepsilon\right] \leq \frac{\mathbf{Var}(x)}{n\varepsilon^2}.$$

Proof

- We have:

$$\begin{aligned}
\mathbf{Pr}\left[\left|\frac{x_1 + x_2 + ... + x_n}{n} - \mathbf{E}(x)\right| \geq \varepsilon\right] &\leq \frac{\mathbf{Var}\left(\frac{x_1+x_2+...+x_n}{n}\right)}{\varepsilon^2} \\
&\quad \text{(using Chebychev's inequality)} \\
&= \frac{1}{n^2\varepsilon^2} \cdot \mathbf{Var}(x_1 + ... + x_n) \\
&= \frac{1}{n^2\varepsilon^2} \cdot \mathbf{Var}(x_1) + ... + \mathbf{Var}(x_n) \\
&\quad \text{(using independence)} \\
&= \frac{\mathbf{Var}(x)}{n\varepsilon^2}.
\end{aligned}$$

Theorem (Law of large numbers)

Let $x_1, x_2, ..., x_n$ be n independent samples of a random variable x. Then

$$\mathbf{Pr}\left[\left|\frac{x_1 + x_2 + ... + x_n}{n} - \mathbf{E}(x)\right| \geq \varepsilon\right] \leq \frac{\mathbf{Var}(x)}{n\varepsilon^2}.$$

- The above theorem gives a sense of how concentrated the sum of independent random variables is around the mean value.
- Such tail bounds are extremely useful in randomised analysis.
- Here is a general theorem for sum of independent random variables.

Theorem (Master tail bounds theorem)

Let $x = x_1 + ... + x_n$, where $x_1, ..., x_n$ are mutually *independent random variables with zero mean and variance at most σ^2. Let $0 \leq a \leq \sqrt{2}n\sigma^2$. Assume that $|\mathbf{E}(x_i^s)| \leq \sigma^2(s!)$ for $s = 3, 4, ..., \lfloor \frac{a^2}{4n\sigma^2} \rfloor$. Then*

$$\mathbf{Pr}(|x| \geq a) \leq 3e^{-\frac{a^2}{12n\sigma^2}}.$$

End