Data Science Bootcamp, 9th January 2017

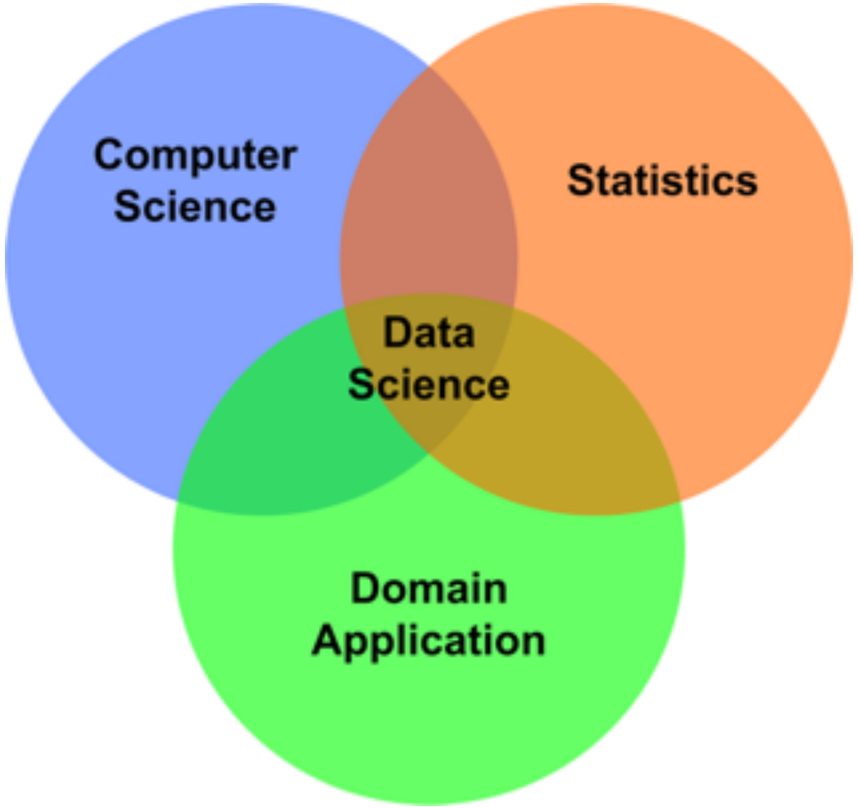
Data Summarization & Visualization
Basic Concepts in Probability

Kriste Krstovski

COLUMBIA UNIVERSITY
Data Science Institute

What is Data Science?

- Interdisciplinary field that lies between computer science statistics and domain application



What is Data Science?

- Computer Science:
  - Algorithms, machine learning, complexity theory, representation, visualization, natural language processing, image processing, information retrieval, optimization, working with big data, etc.
- Statistics:
  - Probability, inference, bayesian statistics, probabilistic modeling, statistical thinking, etc.
- Domain Application:
  - Problems that require domain expertise
  - Involves domain specific data

What is Data Science?

- Apply methods from computer science and statistics in a particular domain
- Build tools that help explore and interpret data, uncover patterns, etc.
- Develop solutions for domain specific problems
- It involves interaction and collaboration between experts in different fields
- The interaction may ultimately lead to creating new approaches across various disciplines

In This Section

- **Data summarization and visualization**
- Basic concepts in probability

Data Summarization

- Summarization depends on the nature of the data
- Numerical
    - Discrete or continuous
    - Mean, median, variance, quantiles, etc.
- Categorical/Ordinal
    - Frequency of occurrence, percentage, etc.

Data Visualization

- Essentially first step in understanding your data
- Powerful exploration tool
- Helps develop intuitions about solving a problem
- Interpretation of the results
- Ways to communicate the results to the general public

In This Section

- Data summarization and visualization
- **Basic concepts in probability**

Basic Concepts in Probability

- Probability function:

  $$p(X)$$

  $$\forall X : 0 \leq p(X) \leq 1 \quad \sum p(X) = 1$$

- Joint Probability:

  $$p(X,Y) = p(Y,X)$$

- Conditional Probability:

  $$p(X|Y)$$

Basic Concepts in Probability

- Independence:

$$p(X,Y) = p(X)p(Y)$$

$$p(X|Y) = p(X)$$

$$p(Y|X) = p(Y)$$

Basic Concepts in Probability

- Sum Rule:

$$p(X) = \sum_{Y} p(X, Y)$$

- Product Rule:

$$p(X, Y) = p(Y|X)p(X)$$

$$p(X, Y) = p(X|Y)p(Y)$$

Basic Concepts in Probability

- Bayes' Rule:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

- Using the sum rule:

$$p(X) = \sum_Y p(X|Y)p(Y)$$

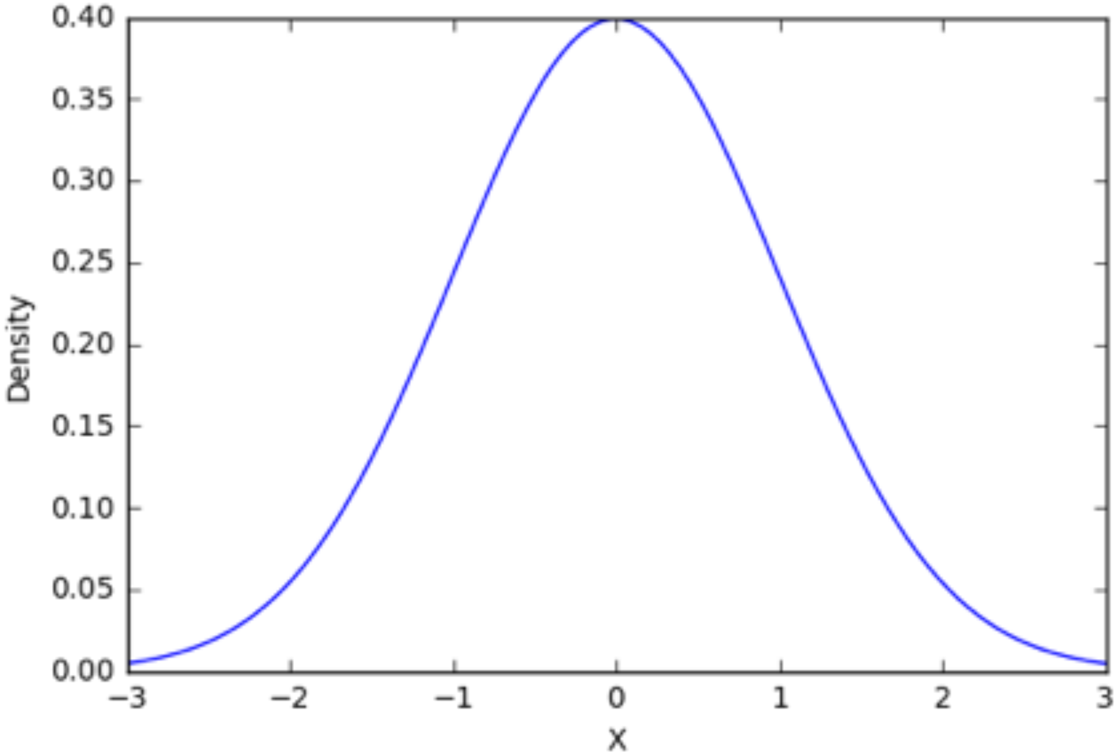
$$p(Y|X) = \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)}$$

Common Probability Distributions

- Gaussian Distribution:

$$X \sim \mathcal{N}(\mu,\, \sigma^2)$$

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



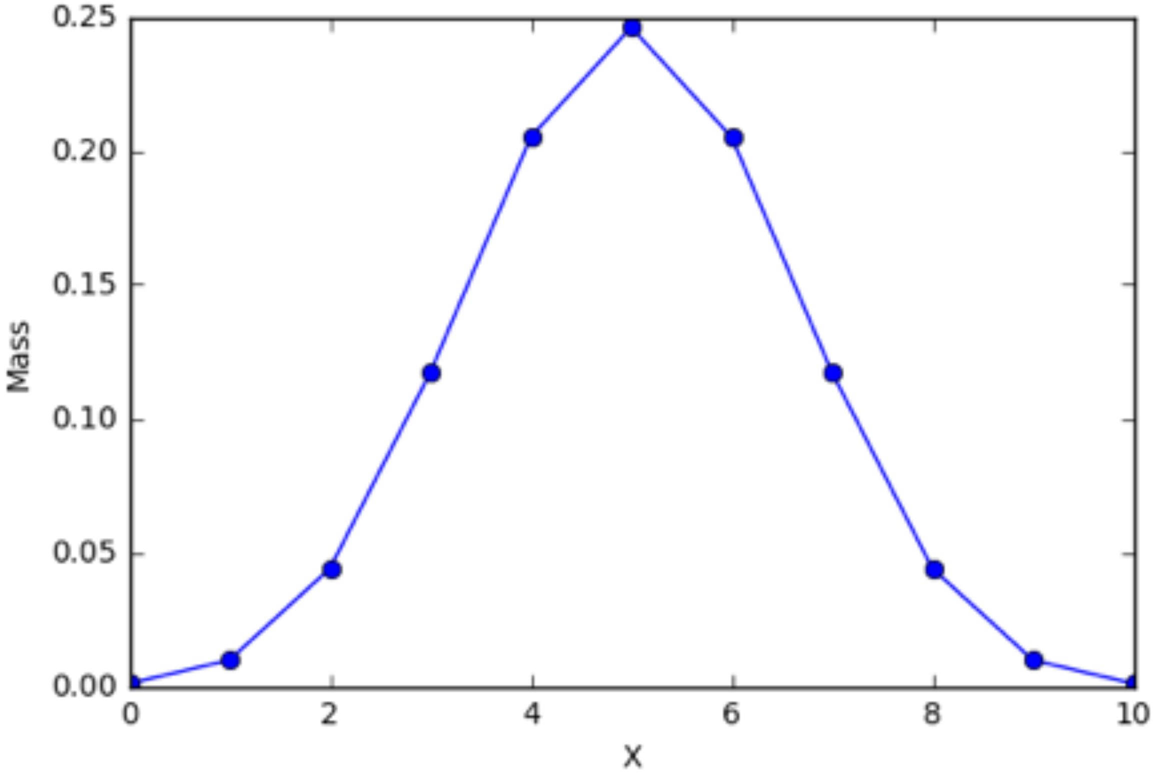
Common Probability Distributions

- Binomial Distribution:

$$X \sim B(n, p)$$

$$p(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

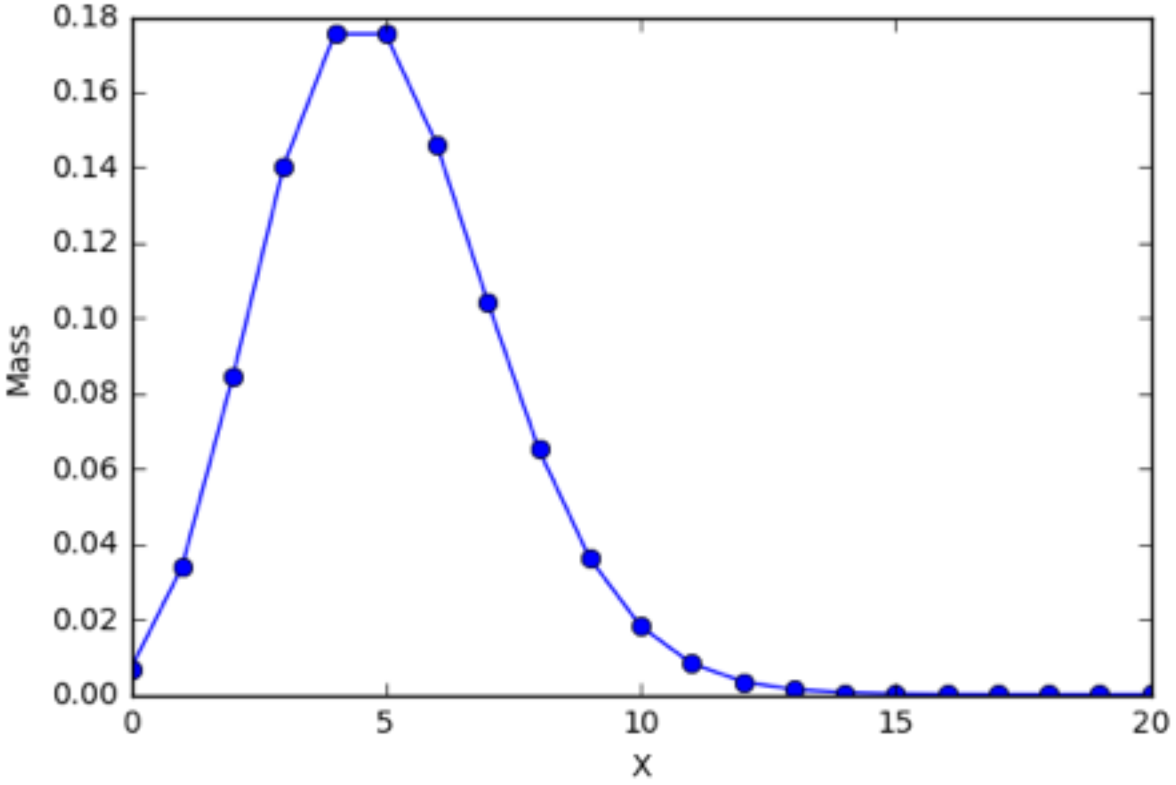


Common Probability Distributions

- Poisson Distribution:

$$X \sim P(\lambda)$$

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$



Two Lab Session

- Cover simple summarization and visualization of different data sets
- Implementation and computation of basic concepts in probability
- Python: Pandas package