Welcome to Online Engineering at George Washington University

Class will begin shortly

Audio: To eliminate background noise, please be sure your audio is muted. To speak, please click the hand icon at the bottom of your screen (**Raise Hand**). When instructor calls on you, click microphone icon to unmute. When you've finished speaking, ***be sure to mute yourself again***.

Chat: Please type your questions in Chat.

Recordings: Please note the recording of this class meeting will be available to download later today. The class recordings are to be used exclusively by registered students in this particular class. Releasing these recordings is strictly prohibited.

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

SEAS 8510
Analytical Methods for Machine Learning

Lecture 10

Dr. Zachary Dennis

Adapted from Rodriguez, Etemadi, Fall 2023, George Washington University

Agenda

9:00 – 9:15		Discussion Group
9:15 – 10:30		2nd Half Course Review
10:30 – 10:40		*BREAK (10 min)*
10:40 – 12:00		Applications of Machine Learning Math Foundations

Assignments

Last week: Discussion Summary due on 6/1 at 9 AM

This week: Final Exam opens 8 PM eastern on Saturday, 6/1 and must be started no later than 5 PM eastern on Monday, 6/3

Probability

How would you define probability?

Probability

- Concerns the study of uncertainty
- Fraction of times an event occurs
- Degree of belief about an event

Probability arises in two contexts:

1. In actual repeated experiments
    a) Ex: You record the color of 1,000 cars driving by and 57 of them are green. You estimate the probability of a car being green as 57/1,000 = 0.057
2. In idealized conceptions of a repeated process

    Ex. You consider flipping a coin. The expected probability of a head is 1/2 = 0.5.

    Ex. You need a model for how people's heights are distributed. You choose a normal distribution to represent the expected relative probabilities.

Why Probability?

Solving machine learning problems requires to deal with uncertain quantities, as well as with stochastic (non-deterministic) quantities

- Probability theory provides a mathematical framework for representing and quantifying uncertain quantities

There are different sources of uncertainty:

- Inherent stochasticity in the system being modeled
- For example, most interpretations of quantum mechanics describe the dynamics of subatomic particles as being probabilistic
- Incomplete observability
- Even deterministic systems can appear stochastic when we cannot observe all of the variables that drive the behavior of the system
- Incomplete modeling
- When we use a model that must discard some of the information we have observed, the discarded information results in uncertainty in the model's predictions
- E.g., discretization of real-numbered values, dimensionality reduction, etc.

Why Probability?

In supervised learning, want to predict something unknown (target) given something known (features). Depending on our objective, you might:

- Predict most likely value
- Predict value with smallest expected distance
- Quantify our uncertainty

In unsupervised learning, you often care about uncertainty

- Learn what's “normal” in order to detect anomalies

Random Variables

- Quantifying uncertainty requires the idea of a random variable
- A **random variable** is a function that maps outcomes of random experiments to a set of properties we are interested in
- A **probability distribution** is a description of how likely a random variable is to take on each of its possibles states
- Random Variables can be **discrete** or **continuous**

Sample Spaces and Events

- An **experiment** is any activity or process whose outcome is subject to uncertainty
- The **sample space** is the set of all possible outcomes of the experiment, denoted by
- The **event space** is the space of potential results of the experiment. Any collection (subset) of outcomes contained in the sample space ***S***
    - *An event is simple if it consists of exactly one outcome, it is compound if it consists of more than one outcome*

Example

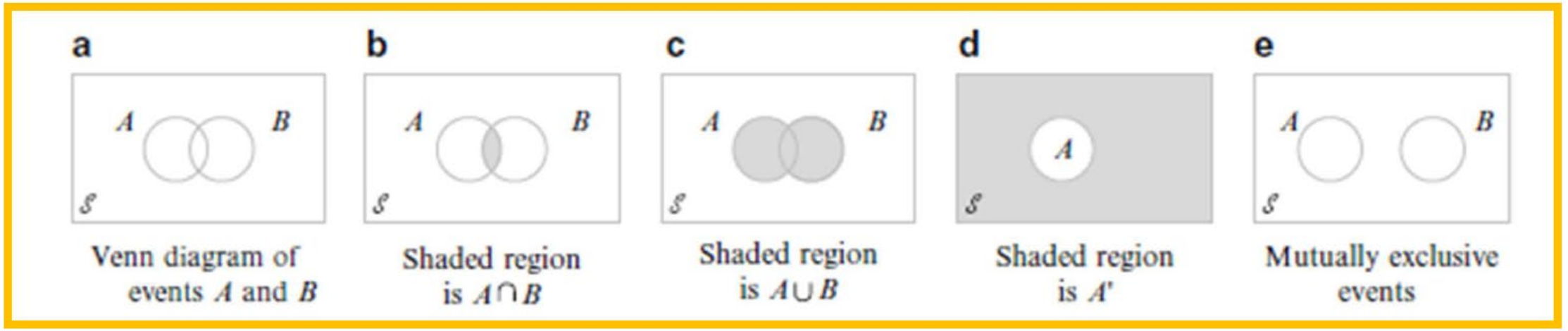
- **Experiment:**
  - Rolling an unbiased 6-sided die
- **Random Variable:**
  - # on the die facing up
- **Sample Space**
  - {1 , 2, 3, 4, 5, 6}
- **Probability of rolling a 6:**
  - P(6) = 1 / 6

Set Theory Related to Events

Definitions

- **Union** of two events A and B – denoted by A U B
    - A “or” B
- **Intersection** of two events A and B – denoted by A∩B
    - A “and” B
- **Complement** of an event A – denoted by A’ is the set of all outcomes in the sample space that are not contained in A
- When A and B have no outcomes in common, they are said to be disjoint or **mutually exclusive** events

Set Theory



Axioms of Probability

Given an experiment with sample space $\mathcal{S}$, the objective of probability is to assign to each event A a number P(A), called the probability of event A, which will give a precise measure of the chance that A will occur.

To ensure that the probability assignments will be consistent with our intuitive notions of probability, all assignments should satisfy the following axioms (basic properties) of probability

Axiom 1 For any event A, P(A) >= 0

Axiom 2 P($\mathcal{S}$) = 1

Axiom 3 If A1, A2, A3,…. Is an infinite collection of disjoint events, then

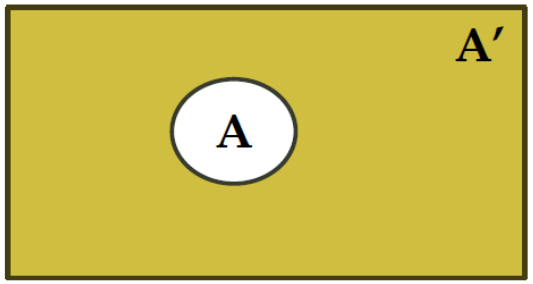
$$P(A_1 \cup A_2 \cup A_3 \ldots) = \sum_{i=1}^{\infty} P(A_i)$$

Probability Properties

Proposition

For any event A, $P(A) + P(A') = 1$, from which

$P(A) = 1 - P(A')$



- Useful because there are many situations where P(A') is more easily obtained than P(A)

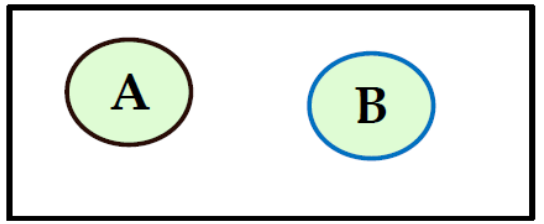
Probability Properties

Proposition

For any event A, $P(A) \leq 1$

When two events A and B are mutually exclusive,

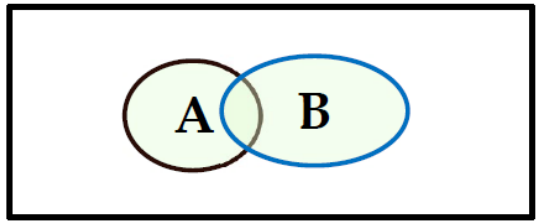
$$P(A \cup B) = \mathrm{P(A)} + \mathrm{P(B)}$$



Proposition

For any two events A and B

$$P(A \cup B) = \mathrm{P(A)} + \mathrm{P(B)} - \mathrm{P}((A \cap B)$$



Conditional Probability

The probabilities assigned to various events depend on what is known about the experimental situation when the assignment is made.

Subsequent to the initial assignment, partial information relevant to the outcome of the experiment may become available. Such information may cause us to revise some of our probability assignments.

For a particular event A we have used P(A) to represent the probability, assigned to A ; we now think of P(A) as the original, or unconditional probability, of the event A

Conditional Probability

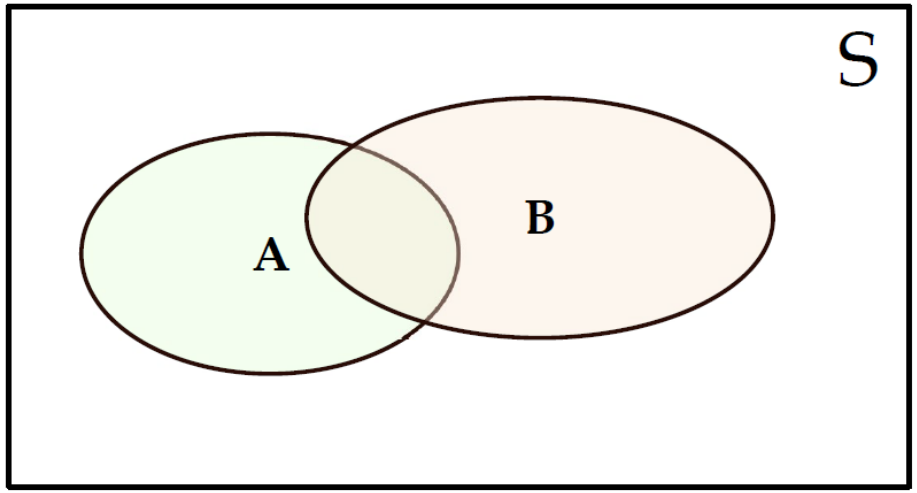
Now we'll examine how the information "event B has occurred" affects the probability assigned to A.

We will use the notation **P(A|B)** to represent the **conditional probability of A given that the event B has occurred**. B is the "conditioning event"

What is the complement to P(A|B)?

Conditional Probability

Think of conditioning on B as redefining the sample space from S to B



Given that B has occurred, the relevant sample space is not longer S, but consists of outcomes in B. Event A has occurred if and only if one of the outcomes in the intersection occurred

Definition of Conditional Probability

Definition

For any two events A and B with P(B) > 0, the conditional probability of A given the B has occurred is defined by

$$P(A|B) = \frac{P\ (A \cap B)}{P(B)}$$

$$P(B|A) = \frac{P\ (A \cap B)}{P(A)}$$

Note that in the first, P(B)>=0

In the second, P(A)>=0

Multiplication Rule

Since,

$$P(A|B) = \frac{P\ (A \cap B)}{P(B)}$$

$$P\ (A \cap B) = \mathrm{P(A|B)} * \mathrm{P(B)}$$

This rule is important because it is often the case that $P(A \cap B)$ is desired, whereas both $P(B)$ and P(A|B) can be specified from the problem description.

Expanding the Multiplication Rule

The multiplication rule can be expanded to more than 2 events.

For example,

$$P\left(A_1 \cap A_2 \cap A_3\right) = \mathrm{P}(A_3|A_2 \cap A_1) * \mathrm{P}(A_1 \cap A_2)$$
$$= \mathrm{P}(A_3|A_2 \cap A_1) * \mathrm{P}(A_2|A_1) * \mathrm{P}(A_1)$$

Where A1 occurs first, followed by A2, and finally A3

Bayes' Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

The terms are referred to as:

- P(A), the prior probability, initial degree of belief for A
- P(A|B), the posterior probability, the degree of belief after incorporating the knowledge of B
- P(B|A), the likelihood of B given A
- P(B), the evidence

$$Posterior\ Probability = \frac{likelihood * prior\ probability}{evidence}$$

Bayes' Theorem

Let A_1, ….., A_k be a collection of mutually exclusive and exhaustive events with $P(A_i)>0$ for i=1…., k. Then for any other event B, for which P(B) > 0

$$P(A_j \mid B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B \mid A_j)P(A_j)}{\sum_{i=1}^{k} P(B \mid A_i)P(A_i)} \quad j = 1, \ldots, k$$

Independence

In our examples, it was frequently the case that P(A|B) differed from the unconditional probability P(A), indicating that the information “B has occurred” resulted in a change in the chance of A occurring. Often the chance that A will occur or has occurred is not affected by knowledge that B has occurred

Definition

Two events A and B are independent if P(A|B) = P(A) and are dependent otherwise.

If A and B are independent, then so are the following pairs of events:

- A’ and B
- A and B’
- A’ and B’

Naïve Bayes

Family of classification algorithms

- Set of features (X1, X2, X3, …. Xn)
- Predicting Class Y

Naïve Bayes “naively” assumes all features are independent.

Algorithm Steps:

- Create frequency tables
- Create likelihood tables
- Calculate posterior probability using Bayes’ theorem
- Predicts higher probability

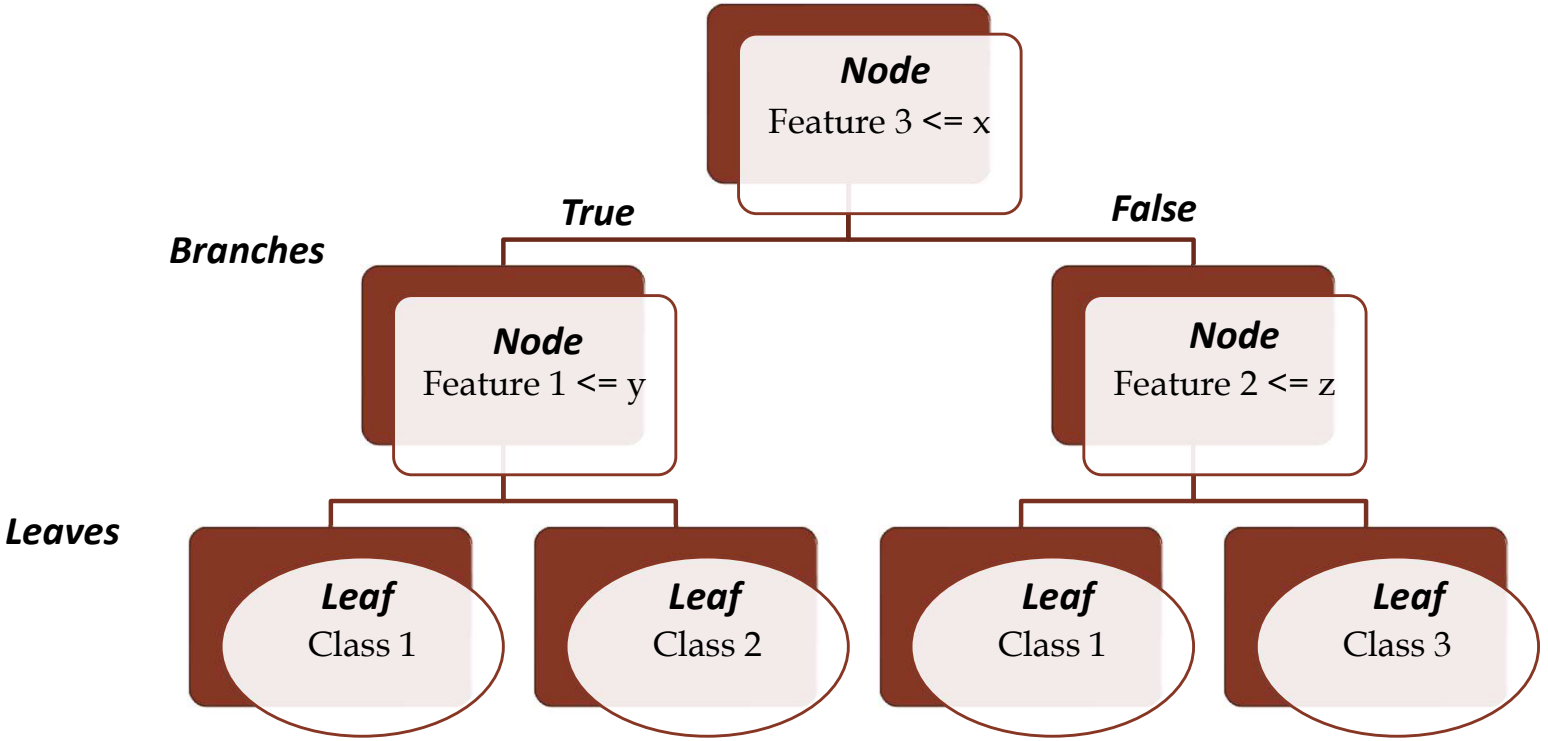
Decision Trees

- Decision Trees are versatile Machine Learning algorithms that can perform both classification and regression tests, and even multioutput tasks
- Powerful algorithms capable of fitting complex data sets
- Decision Trees are also the fundamental components of Random Forests, which are among the most powerful Machine Learning algorithms available today.

(Geron, 2019, p. 175)

Decision Tree Terminology

- Tree-shaped diagram
- Each branch represents a possible decision, occurrence or reaction



Construction of Decision Trees

- **Top Down Approach:**
  - Evaluate each attribute for usefulness classifying
  - Select best attribute as root of the tree
  - Split based on attribute to produce a subset of data
  - Repeat on subsets to find next nodes considering only attributes that have not been selected already

- **How do you decide which attribute is most useful?**

Attribute Selection Measures

- **How do you decide which attribute is most useful?**
  - Entropy
  - Information Gain
  - Gini Index

- **Common Algorithms and their Methods:**
  - ID3 (Iterative Dichotomizer 3) – Entropy, Information Gain
  - C4.5 (Successor of ID3) – Entropy, Information Gain
  - CART (Classification and Regression) – Gini Index

Entropy

- Entropy is a measure of impurity or randomness
- Node is pure when value 0 or 1

$$E(S) = \sum_{i=1}^{n} -p_i log_2 p_i$$

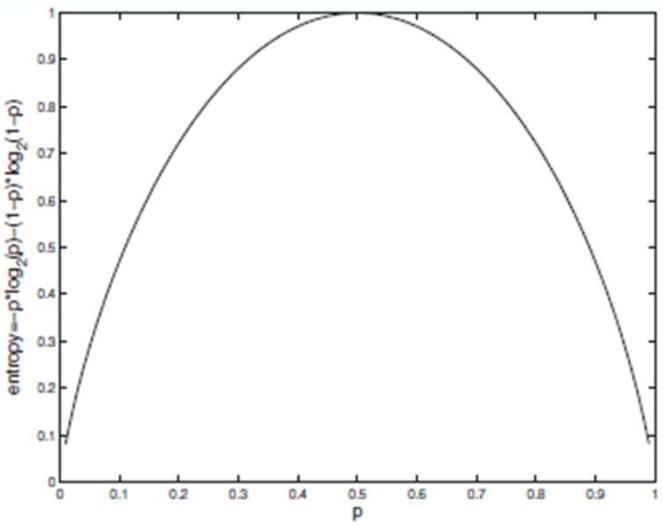


Figure 9.2 Entropy function for a two-class problem.

- S is the current state
- pi is the probability of an event i of state S or percentage of class i in a node of state S

(Alpaydin, 2014, p. 217)

Information Gain

- How much information an attribute provides for the target variable

$$Information\ Gain = Entropy_{parent} - Entropy_{children}$$

$$Info\ Gain\ (A; B) = E(A) - \sum_{b} P(B = b) * E(A|B = b)$$

Using Decision Trees to Make Predictions

Start at the root node (depth 0, at the top)

Is the petal length of the flower smaller than 2.45 cm?

- If it is, then you move down to the left (depth 1, left). This is a leaf node and the decision tree predicts that your flower is an *Iris Setosa*

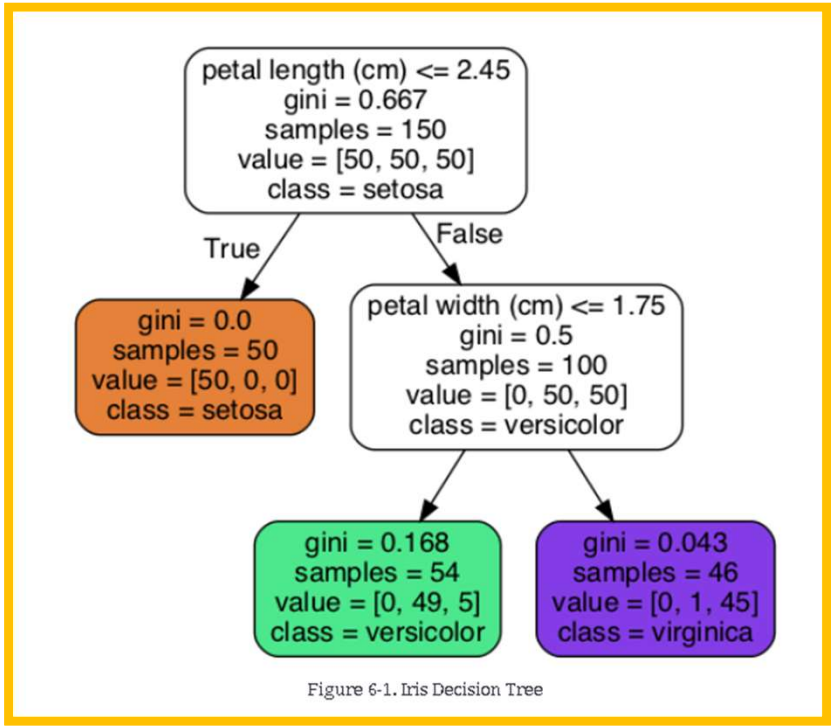


Figure 6-1. Iris Decision Tree

If it is greater than 2.45 cm…

Move down to the right to child node (depth 1, right)

Is the petal width smaller than 1.75 cm?

- Yes > then your flower is most likely an *Iris Versicolor* (depth 2, left)
- No > then your flower is likely *Iris Virginica* (depth 2, right)

(Geron, 2019, p. 176)

Using Decision Trees to Make Predictions

- A node's **samples** attribute is how many training instances it applies to
- How many training instances have petal length of more than 2.45 cm?

  100
- A node's **value** attribute tells you how many training instances of each class this node applies to
  - This node (petal length greater than 2.45 cm and width less than 1.75 cm) applies to: 0 setosa, 49 versicolor, and 5 virginica
- A node's **gini** attribute measures its impurity
  - A node is "pure" and has gini = 0 when all training instances it applies to belong to the same class

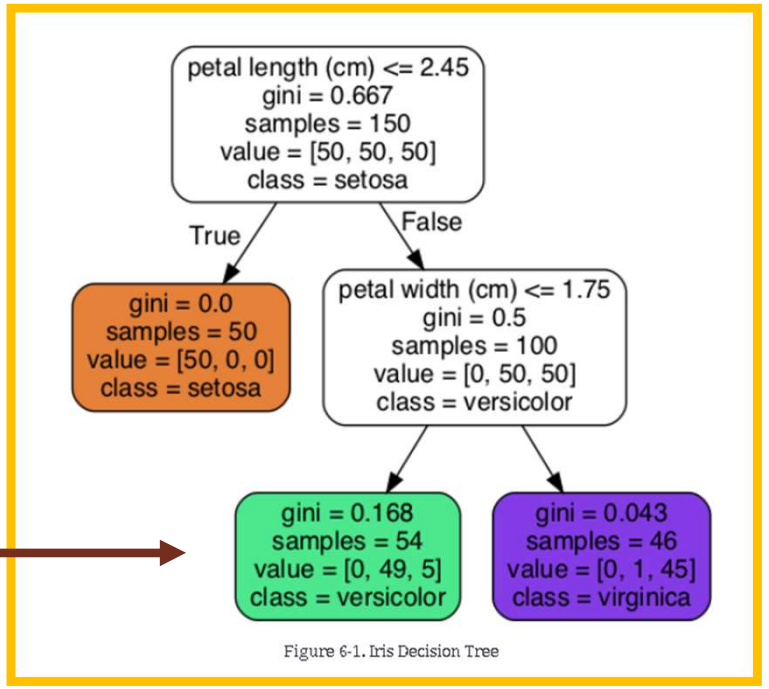


Figure 6-1. Iris Decision Tree

(Geron, 2019, p. 177)

Gini Impurity

$$G_i = 1 - \sum_{k=1}^{n} {p_{i,k}}^2$$

- $p_{i,k}$ is the ratio of class *k* instances among the training instances in the *ith* node
- Gini = 0 when the node is “pure” and all training instances it applies to belong to the same class

(Geron, 2019, p. 177)

Estimating Class Probabilities

A Decision Tree can also estimate the probability that an instance belongs to a particular class k

First it traverses the tree to find the leaf node for this instance, and then it returns the ratio of training instances of class k in this node.

If you ask it to predict the class, it outputs the class with the highest probability

Example instance: Petal length = 5 cm, Petal width = 1.5 cm

- What is probability the flower is Iris Setosa?

= 0 / 54 = 0%

- Iris Versicolor?

= 49 / 54 = 90.7%

- Iris Virginica?

= 5 / 54 = 9.3%

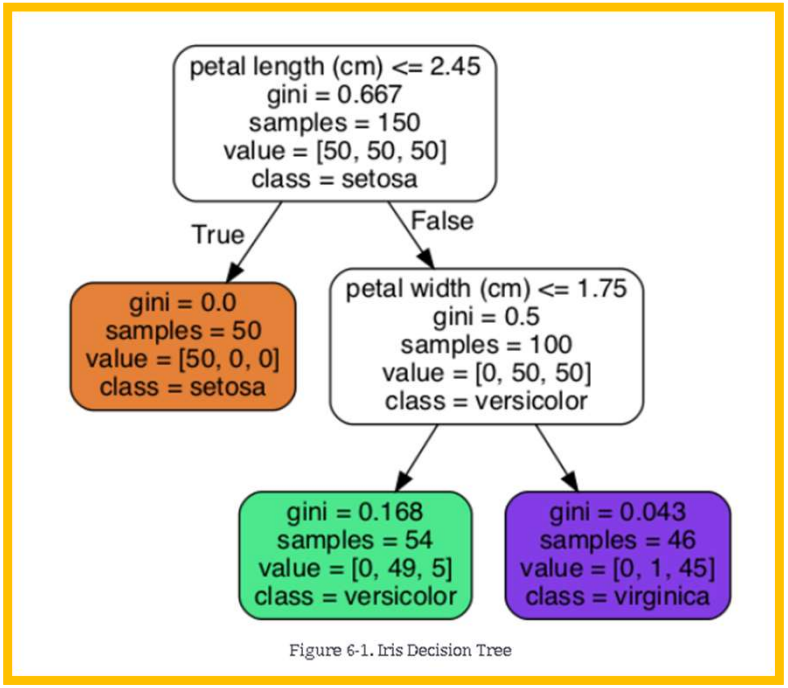
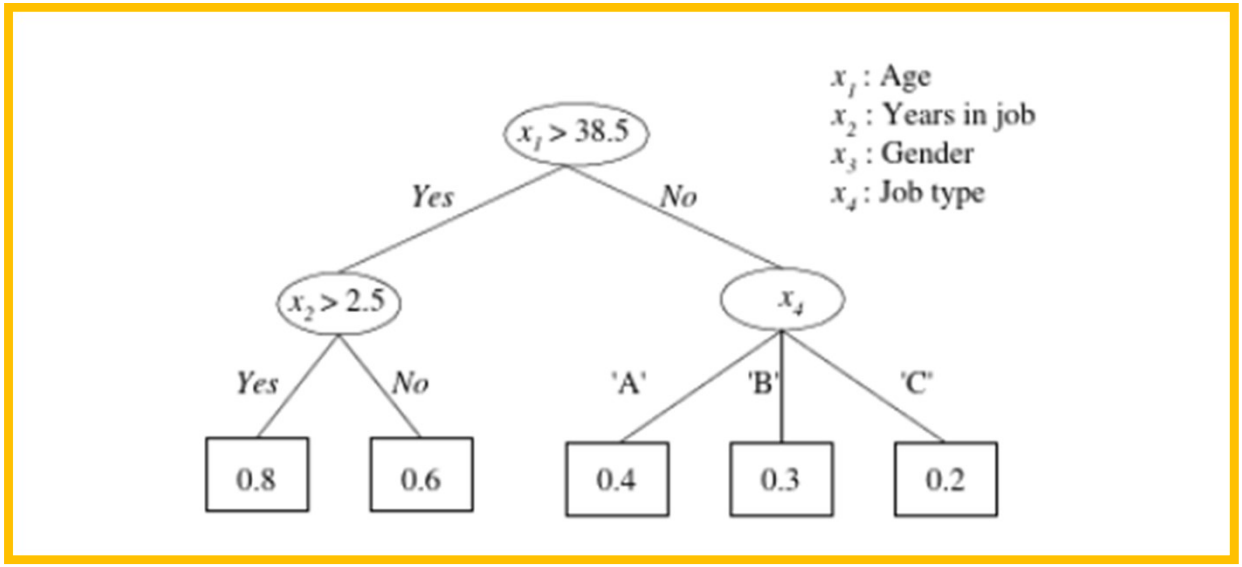


Figure 6-1. Iris Decision Tree

(Geron, 2019, p. 178)

Rule Extraction from Trees

- The decision tree can be converted to IF-THEN rules by tracing the path from the root node to each leaf node in the tree



R1: IF (age $>$ 38.5) AND (years-in-job $>$ 2.5) THEN $y = 0.8$
R2: IF (age $>$ 38.5) AND (years-in-job $\leq$ 2.5) THEN $y = 0.6$
R3: IF (age $\leq$ 38.5) AND (job-type = 'A') THEN $y = 0.4$
R4: IF (age $\leq$ 38.5) AND (job-type = 'B') THEN $y = 0.3$
R5: IF (age $\leq$ 38.5) AND (job-type = 'C') THEN $y = 0.2$

(Alpaydin, 2014, p. 225)

Regularization Hyperparameters

Decision Trees:

- Make very few assumptions about the training data
- Tend to overfit if they are not constrained
  - The tree structure will adapt itself to the training data, fitting it very closely

- To avoid overfitting to the training data, you need to restrict the decision tree's freedom during training (i.e. regularization)

- The regularization hyperparameters depend on the algorithm used, but generally you can at least restrict the maximum depth of the decision tree
  - Default max_depth = None allows for unlimited depth. Reducing max_depth will regularize the model and reduce the risk of overfitting

(Geron, 2019, p. 181)

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Naïve Bayes

Naïve Bayes

Family of classification algorithms

- Set of features (X1, X2, X3, …. Xn)
- Predicting Class Y

Naïve Bayes “naively” assumes all features are independent.

Algorithm Steps:

1. Create frequency tables
2. Create likelihood tables
3. Calculate posterior probability using Bayes’ theorem
4. Predicts higher probability

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Random Forests

Ensemble Method

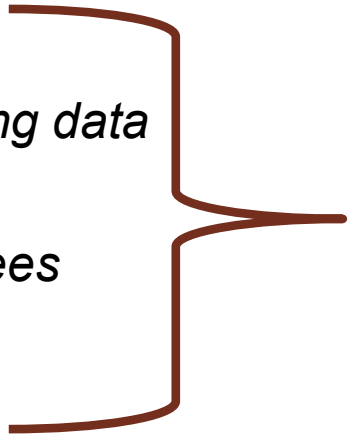
Example of an ensemble method:

Train a group of Decision Tree Classifiers

Each on a random subset of the training data

Make predictions using each of the individual trees

Then predict the class the gets the most votes

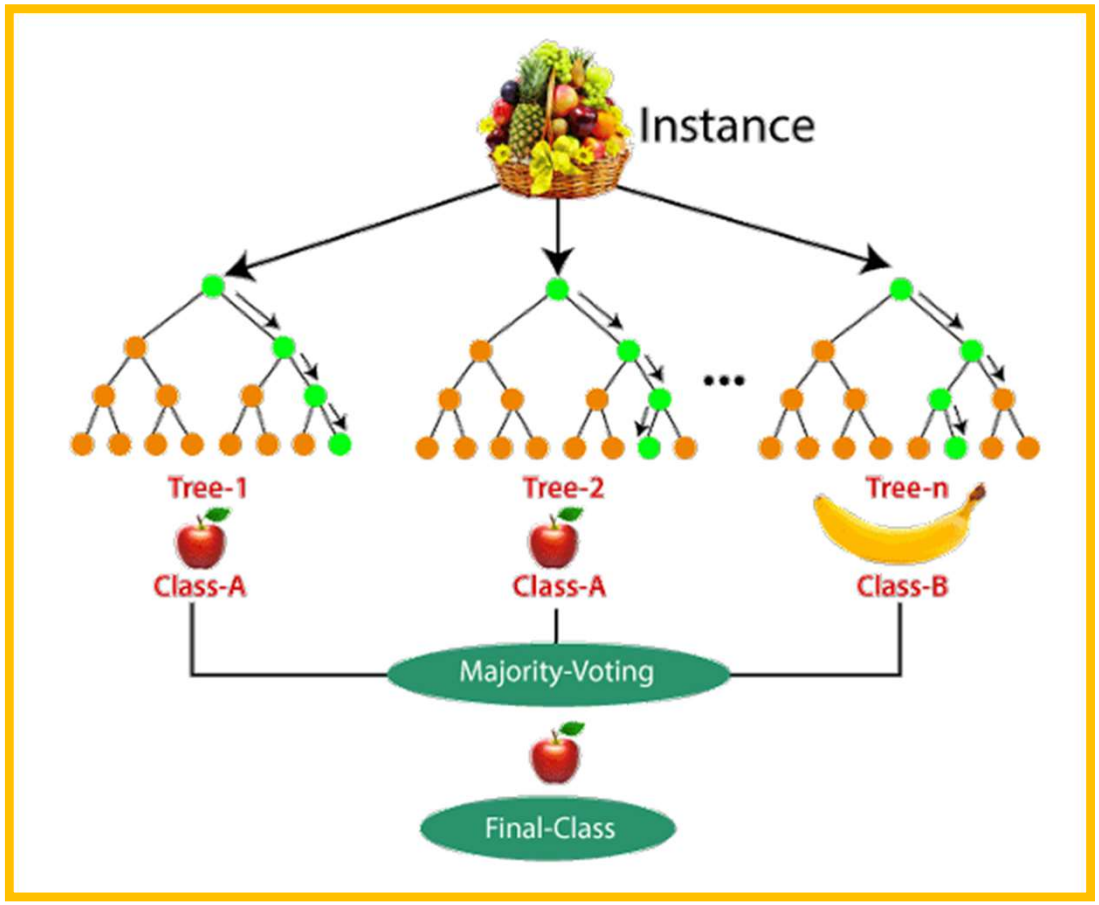


Random Forest

Often use ensemble method near end of a project, after already building a few good predictors. Combine into even better predictor.

(Geron, 2019, p. 189)

Random Forests



- Random Forest is an ***ensemble*** of Decision Trees, generally trained via the bagging method (or sometimes pasting)
- Random Forest introduces extra randomness when growing trees; instead of searching for very best feature when splitting a node, it searches for the best feature among a random subset of features
- Results in greater tree diversity, which trades a higher bias for a lower variance, generally yielding an overall better model

(Geron, 2019, p. 197-8)
Image: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Discrete Distributions

Random Variables

Definition

For a given sample space S of some experiment, a **random variable** (rv) is any rule that associates a number with each outcome in S . In mathematical language, a random variable is a function whose domain is the sample space and whose range is the set of real numbers.

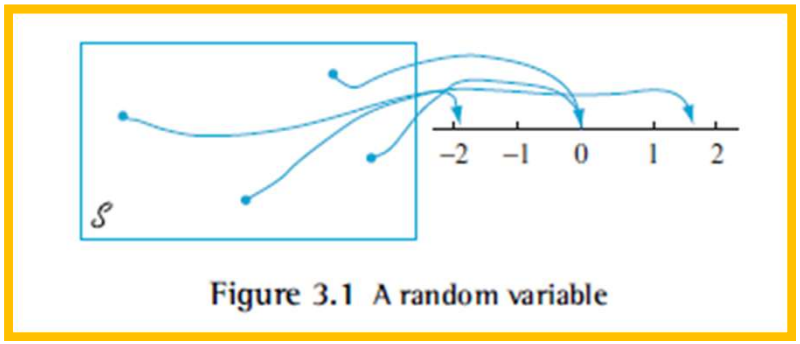


Figure 3.1 A random variable

(Devore & Berk, 2018, p.98)

Two Types of Random Variables

Definition

A **discrete** random variable is a rv whose possible values either constitute a finite set or else can be listed in an infinite sequence in which there is a first element, a second element, and so on ("countably infinite").

A random variable is **continuous** if *both* of the following apply:

1. Its set of possible values consists either of all numbers in a single interval on the number line (possibly infinite in extent, e.g., from $-\infty$ to ∞) or all numbers in a disjoint union of such intervals (e.g., $[0, 10] \cup [20, 30]$)
2. No possible value of the variable has positive probability, that is, $P(X=c)=0$ for any possible value c

(Devore & Berk, 2018, p.98)

Two Types of Random Variables

Definition

A **discrete** random variable is a rv whose possible values either constitute a finite set or else can be listed in an infinite sequence in which there is a first element, a second element, and so on ("countably infinite").

A random variable is **continuous** if *both* of the following apply:

1. Its set of possible values consists either of all numbers in a single interval on the number line (possibly infinite in extent, e.g., from $-\infty$ to ∞) or all numbers in a disjoint union of such intervals (e.g., $[0, 10] \cup [20, 30]$)
2. No possible value of the variable has positive probability, that is, $P(X=c)=0$ for any possible value c

(Devore & Berk, 2018, p.98)

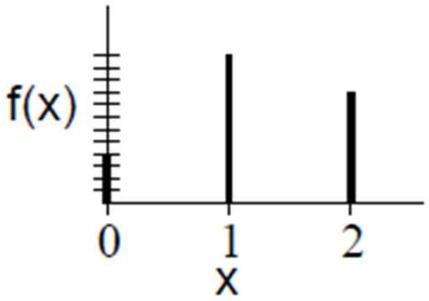
Representing Random Variables

Probability Mass Function (pmf)

When variables take on values that are countable or listable from smallest to largest they are called Discrete Random Variables

A pmf, f(x), of a random variable X is a function representing the values of a random variable with their associated probabilities – can be specified in tabular, graphic or equation form

X	f(x)
0	0.16
1	0.48
2	0.36



$$f(x) = \binom{2}{x}(0.6)^x(0.4)^{2-x}$$

$$x = 0,1,2$$

(Devore & Berk, 2018, p.101)

Representing Random Variables

Cumulative Distribution Function (cdf)

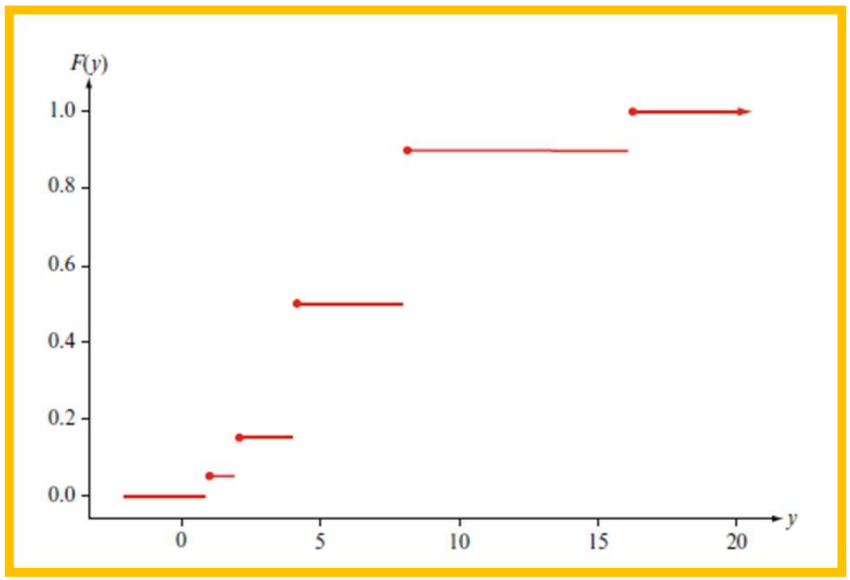
A CDF, F(x), of a random variable X is a function representing P(X<=x)

For discrete random variables this is a step function

pdf

y	1	2	4	8	16
$p(y)$	.05	.10	.35	.40	.10

cdf

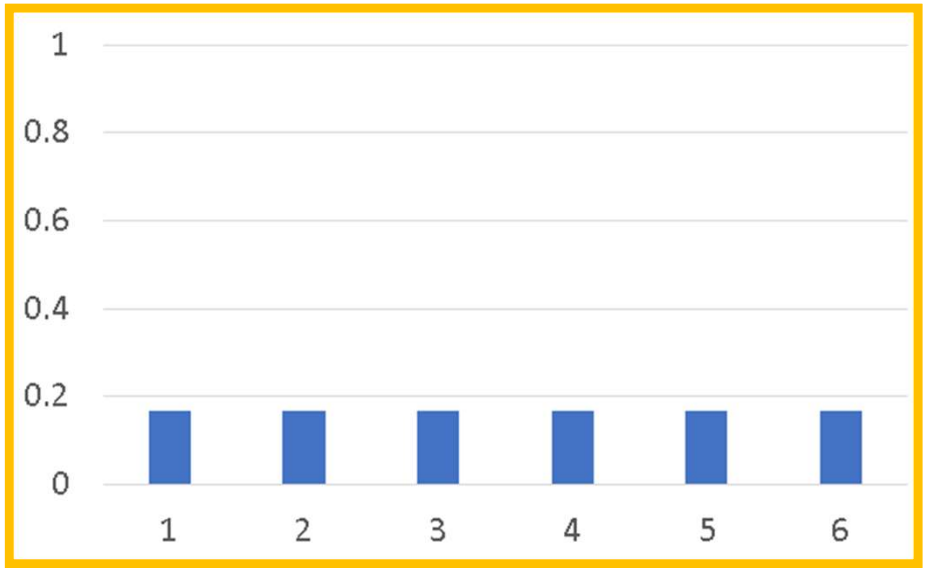


(Devore & Berk, 2018, p.105-106)

Uniform Distribution

The probabilities of each outcome are **evenly distributed** across the sample space

Ex: Rolling a fair die has 6 discrete, equally probable outcomes



Notation: $X \sim U(n)$

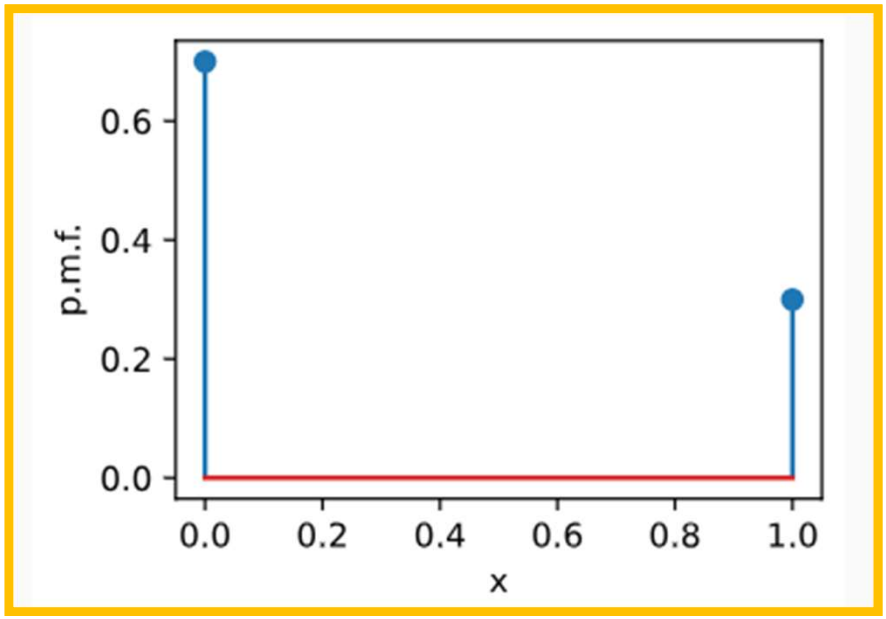
The probability of each value i $\in \{1, 2, \ldots . n\}$ is

$$p_i = \frac{1}{n}$$

Bernoulli Distribution

Any random variable whose only possible value are 0 and 1 is called a **Bernoulli random variable**

Notation: $X \sim Bernoulli(p)$



p = 0.3
1 – p = 0.7

https://d2l.ai/chapter_appendix-mathematics-for-deep-learning/distributions.html

Binomial Distribution

Binomial means there are two discrete, mutually exclusive outcomes

- Heads **or** tails
- On **or** off
- Defective **or** non defective
- Success **or** failure

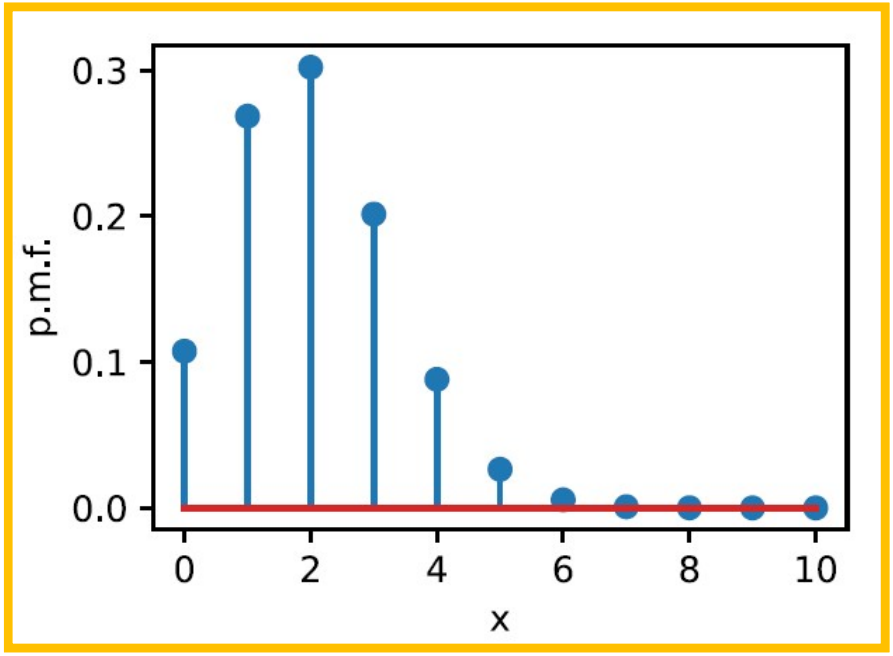
4 Criteria for an experiment to follow the Binomial Distribution:

- The experiment consist of a sequence of n smaller experiments called trials, where n is fixed in advance.
- Each trial can result in one of the same two possible outcomes, we generically denote by success (S) or failure (F).
- Trials are independent, so that the outcome of any particular trial does not influence the outcome on any other trial,
- The probability of success P(S) is constant from trial to trial; we denote this as probability p.

(Devore & Berk, 2018, p.128)

Binomial Distribution

- Performing a sequence of *n* independent experiments, each of which has probability *p* of succeeding, where $p \in \{0, 1\}$
- The probability of getting *k* successes in *n* trials is $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- Notation: $X \sim Binomial(n, p)$



https://d2l.ai/chapter_appendix-mathematics-for-deep-learning/distributions.html

Binomial Distribution

Example:

If you roll a die 16 times,

- What is the probability that a five comes up three times?

$p(x = 3)$

```
from scipy.stats import binom
binom.pmf(3,16,1/6)
```

```
0.24231376033713137
```

- What is the probability that a five comes up at least three times?

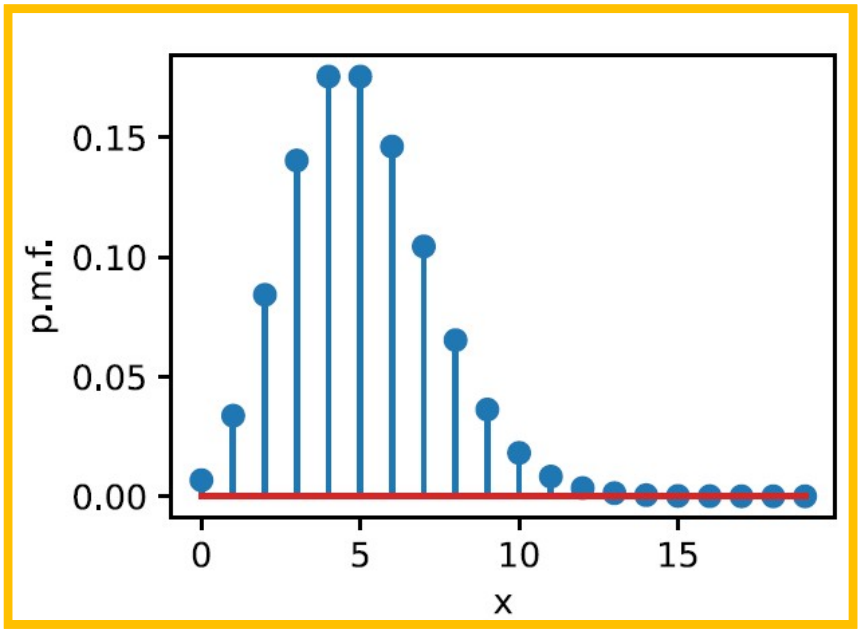
$p(x \geq 3) = 1 - p(x \leq 2)$

```
1 - binom.cdf(2, 16, 1/16)
```

```
0.07420726082533868
```

Poisson Distribution

- A number of events occurring independently in a fixed interval of time with a known rate λ
- A discrete random variable X with states $\{0, 1, 2, \dots\}$ has probability $P(X = k) = \frac{\lambda^X e^{-\lambda}}{X!}$
- The rate λ is the average number of occurrences of the event
- Notation: $X \sim Poisson(\lambda)$



https://d2l.ai/chapter_appendix-mathematics-for-deep-learning/distributions.html

Poisson Distribution

Example:

The number of arrivals at a store can be modeled as Poisson process with an average arrival rate of 10 customers per hours. What is the probability in the next 30 minutes that:

a) 3 customers arrive

```
from scipy.stats import poisson

poisson.pmf(3, 5)
```

```
0.1403738958142805
```

b) More than 4 customers arrive

```
1- poisson.cdf(4, 5)
```

```
0.5595067149347874
```

c) Less than 6 customers arrive

```
poisson.cdf(5, 5)
```

```
0.615960654833063
```

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Continuous Distributions

Continuous Distributions

Continuous random variables take on values in a continuum

Examples:

- Car Speed [0, 150]
- Gas Price [$3, $4]

Probability Density Function

- For a continuous random variable P(X=x*) = 0 for any specific value x*
- The pdf, f(x), is used to express probability of intervals

$$P(a \leq X \leq b) = \int_a^b f(x\_dx$$

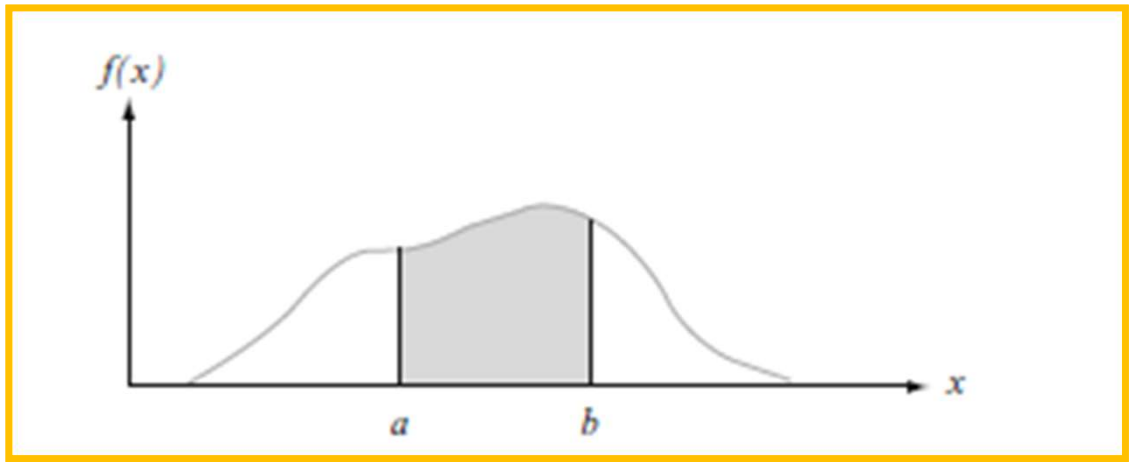
- Note that for a continuous random variable

  P(a < X < b) = P(a <= X < b) = P(a < X <= b) = P(a <= X <= b)

(Devore & Berk, 2018, p.160)

Continuous Distributions

That is, the probability that X takes on a value in the interval [a, b] is the area above this interval and under the graph of the density function (illustrated below).



The graph of f(x) is often referred to as the **density curve**.

(Devore & Berk, 2018, p.160)

Probability Distributions for Continuous Variables

For f(x) to be a legitimate pdf, it must satisfy the following two conditions:

1. $f(x) \geq 0$ for all x

2. $\int_{-\infty}^{\infty} f(x)dx = [\text{area under the entire graph of f(x)}] = 1$

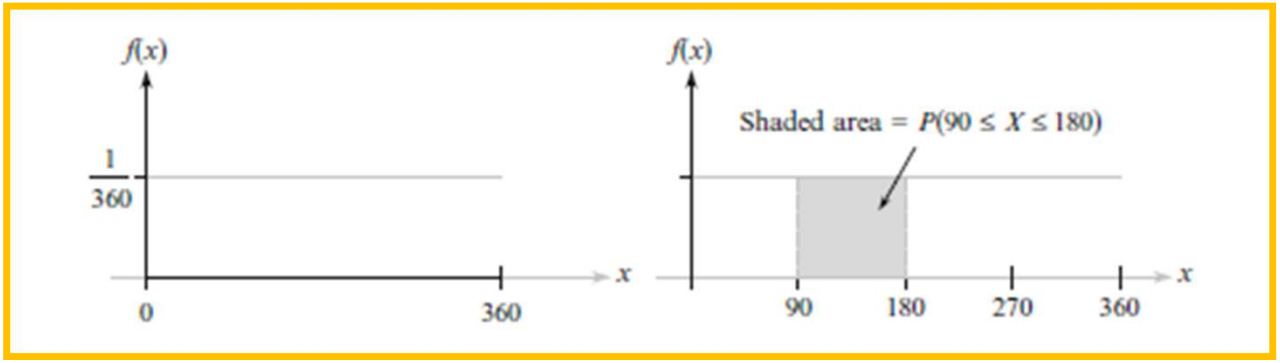
(Devore & Berk, 2018, p.161)

Continuous Uniform Distribution

Definition

A continuous rv X is said to have a uniform distribution on the interval [A, B] if the pdf of X is:

$$f(x; A, B) = \begin{cases} \dfrac{1}{B - A} & A \le X \le B \\ 0 & Otherwise \end{cases}$$



(Devore & Berk, 2018, p.161)

Normal Distribution

The **normal distribution** is the most important one in all of probability and statistics. Many numerical populations have distributions that can be fit very closely by an appropriate normal curve.

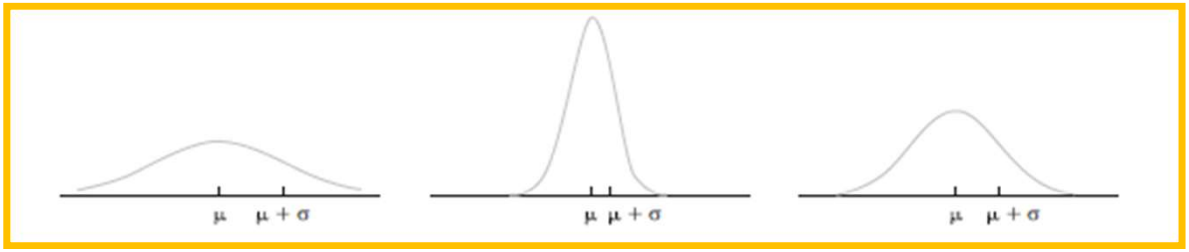
Definition

A continuous rv X is said to have a normal distribution with parameters μ and σ (or μ and σ^2), where $-\infty < \mu < \infty$ and $0 < \sigma$, if the pdf of X is

$$f(x;\ \mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/(2\sigma^2)} \qquad -\infty < x < \infty$$

Mean and Variance:

$$E(X) = \mu$$
$$V(X) = \sigma^2$$



(Devore & Berk, 2018, p.179)

Normal Distribution Applications

- Often an assumption in ML algorithms
  - Gaussian Naïve Bayes – assumes likelihood of features is normal
  - Linear Regression – assumes residuals normal
- Often an assumption in Statistical Tests
  - ANOVA (Analysis of Variance) – assumes residuals normal
  - T-tests – assumes populations samples from normal
- Featuring scaling - standardization

Normal Distribution Example

Example:

The time that is takes a drive to react to brake lights on a decelerating vehicle is critical in helping to avoid rear- end collisions

The article “Fast-Rise Brake Lamp as a Collision-Prevention Device” (*Ergonomics*, 1993: 391-395) suggests that reaction time for an in-traffic response to a brake signal from standard break lights can be modeled with a normal distribution having a mean value of 1.25 sec and a standard deviation of .46

(Devore & Berk, 2018, p.186)

Exponential Distribution

Definition

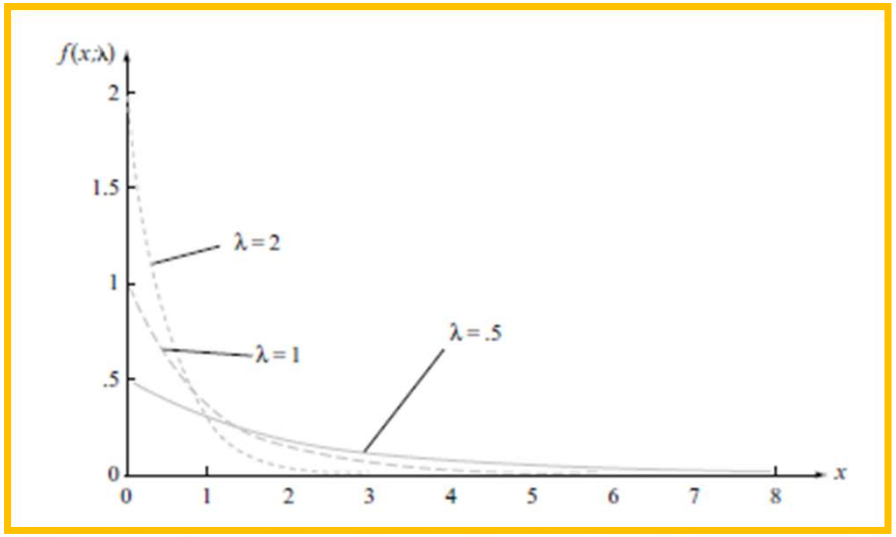
X is said to have an **exponential distribution** with parameter λ $(\lambda > 0)$ if the pdf of X is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & otherwise \end{cases}$$

$$F(x;\lambda) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

$$\mu = \frac{1}{\lambda}$$

$$\sigma^2 = \frac{1}{\lambda^2}$$



(Devore & Berk, 2018, p.198)

Exponential Distribution

The exponential distribution is frequently used as a model for the distribution of **times between the occurrence of successive events**, such as customers arriving at a service facility or calls coming into a switchboard.

The reason for this is that the exponential distribution is closely related to the Poisson process

(Devore & Berk, 2018, p.198)

Gamma Distribution

Definition

A continuous random variable X is said to have a **gamma distribution** if the pdf of X is

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Where the parameters $\alpha \; and \; \beta \; \alpha > 0, \; \beta > 0.$

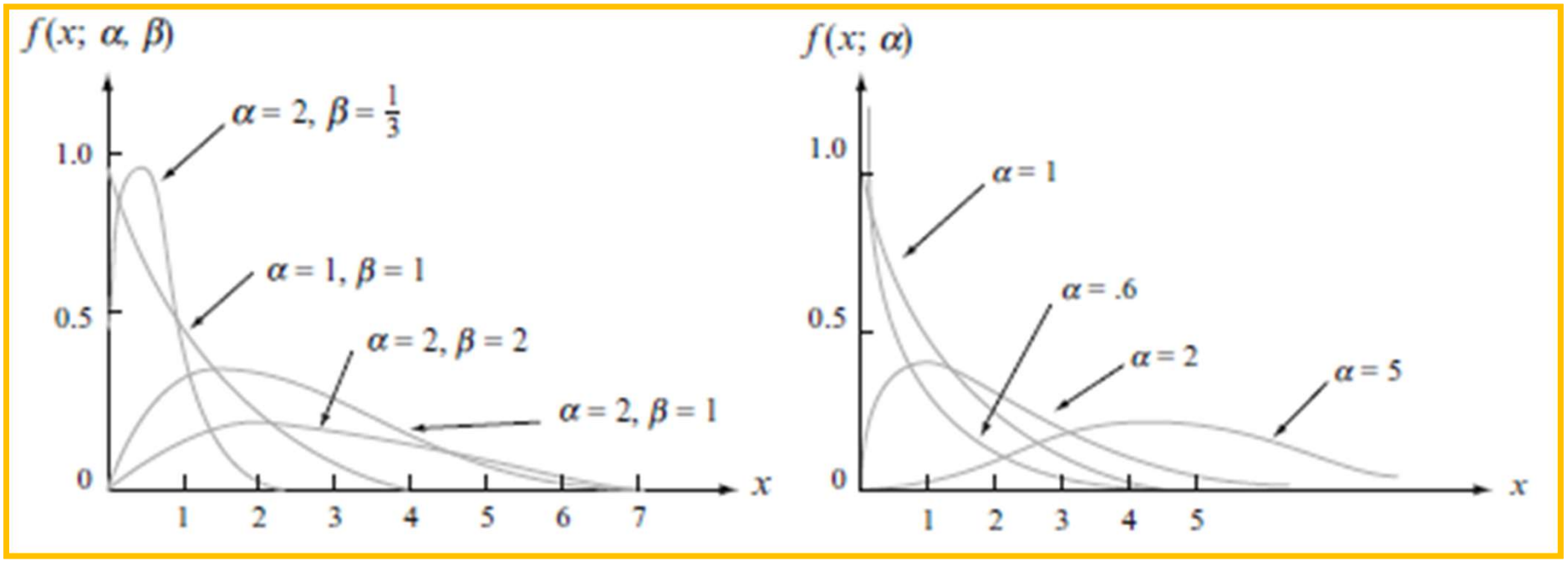
Mean and Variance:

$$E(X) = \mu = \alpha\beta$$
$$V(X) = \sigma^2 = \alpha\beta^2$$

The **standard gamma distribution** has β=1

(Devore & Berk, 2018, p.195)

Gamma Distribution



Gamma Density Curves

Standard Gamma Density Curves

The **exponential distribution** results from taking $\alpha = 1 \; and \; \beta = 1/\lambda$

(Devore & Berk, 2018, p.196)

Probability Distribution Percentile

- Let X be some continuous r.v., and p be a probability of interest.
- Sometimes we are interested in finding q_p such that

$$F_X(q_p) = P(X \leq q_p) = p$$

  where the smallest value of q_p for which this is true is the p-th quantile (or 100p-th percentile) of the distribution for X. The median of a distribution is its 50th percentile

- **Example**: If exam scores are distributed normally with mean and std. dev. of 80 and 5, what is the 90th percentile score?

```
1 norm.ppf(0.9, loc=80, scale=5)
```

```
86.407757827723
```

Understanding Data

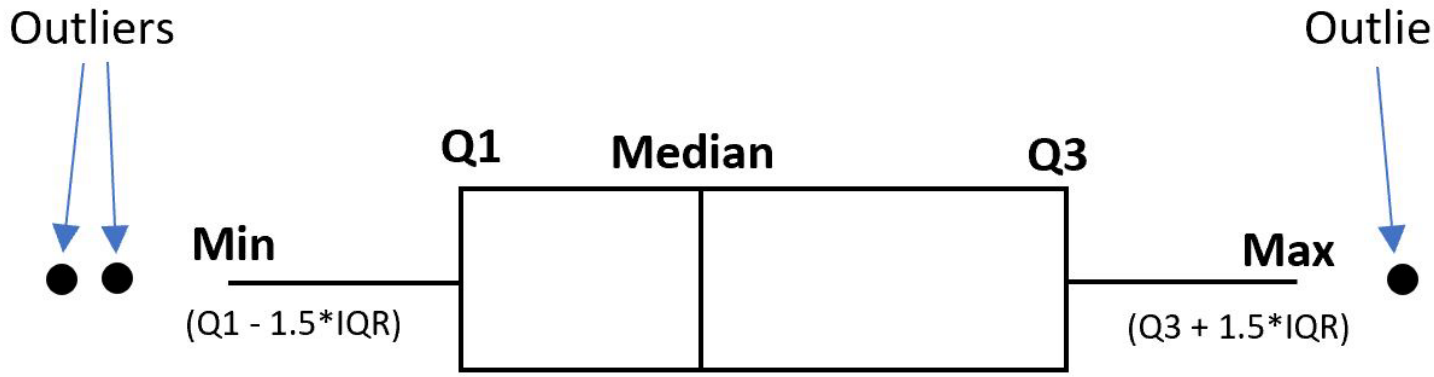
- In machine learning, understanding your data is the first step toward building effective models. Let's explore key data statistics and metrics.
- **Mean**: The mean is a measure of central tendency. It represents the average value of a dataset.
- **Median:** Another measure of central tendency. It represents the middle value when data is sorted (or the average of two middle values). Less affected by outliers compared to the mean
- **Variance**: Variance measures the spread or dispersion of data points. A higher variance indicates greater data variability.

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- **Standard Deviation**: Standard deviation is the square root of the variance, s_x. It provides a standardized measure of data dispersion.

Understanding Data

- **Quantiles:** Divide data into equal-sized subsets. Common quantiles include quartiles (25th, 50th, 75th percentiles). Useful for understanding data distribution and identifying outliers.
- **Interquartile Range (IQR):** IQR measures the spread of data around the median. It is the difference between the 75th percentile (Q3) and the 25th percentile (Q1). Useful for identifying outliers and assessing data variability.
- **Skewness:** Quantifies the asymmetry of the data distribution.
  - Positive skew: Data is skewed to the right (tail on the right).
  - Negative skew: Data is skewed to the left (tail on the left).



https://www.statology.org/how-to-read-box-plot-with-outliers/

Understanding Data

- **Covariance:** Covariance measures the degree to which two variables change together.
  - Positive covariance: Variables move in the same direction.
  - Negative covariance: Variables move in opposite directions.

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- **Correlation:** Correlation is a standardized measure of covariance. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation). Helps assess the linear relationship between two variables.

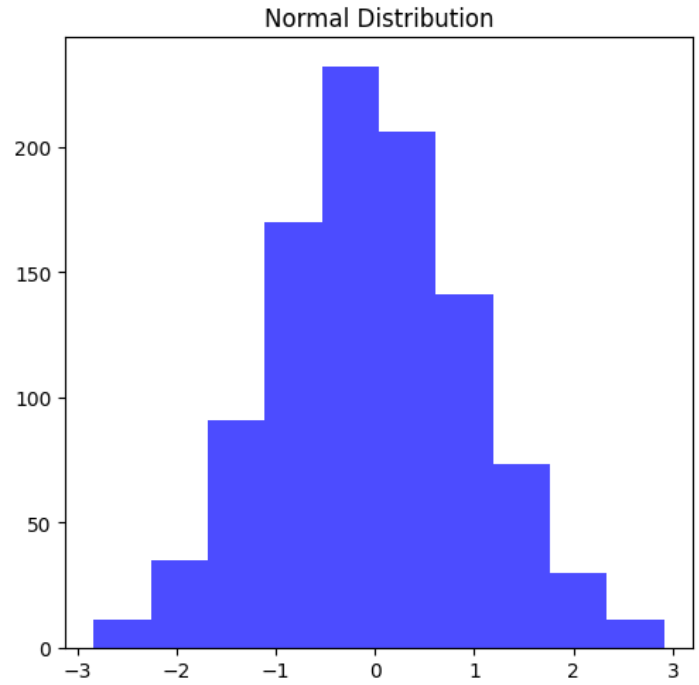
$$r = \frac{\mathrm{Cov}(X,Y)}{s_X s_Y}$$

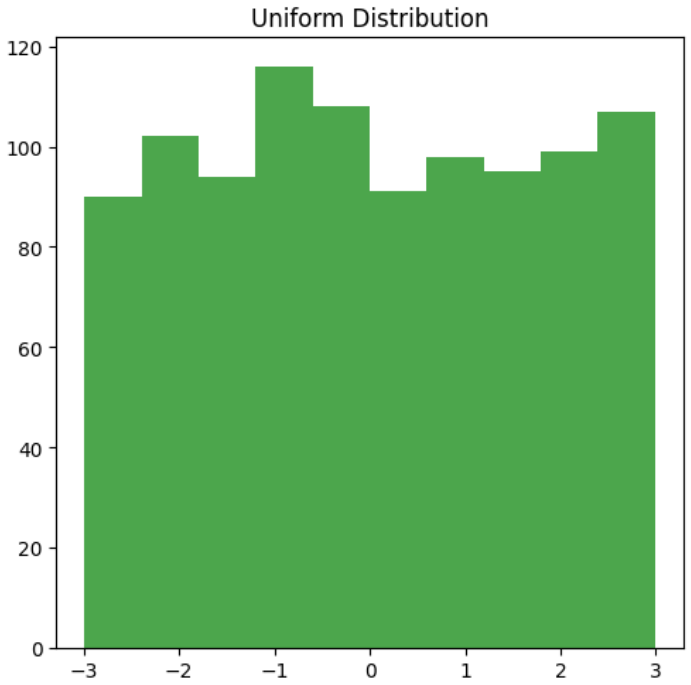
Understanding Data

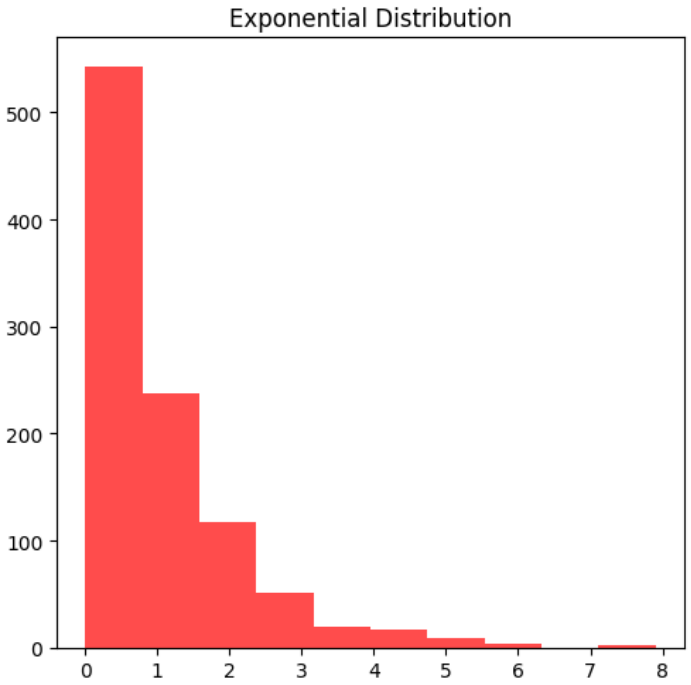
```
mean = np.mean(data)
variance = np.var(data, ddof = 1) #Denominator of N-1, sample var
std_dev = np.std(data)
median = np.median(data)
Q1 = np.percentile(data, 25)
Q3 = np.percentile(data, 75)
cov_matrix = np.cov(data_X, data_Y)
corr_matrix = np.corrcoef(data_X, data_Y)

from scipy.stats import iqr, skew
iqr(data)
skew(data)
```

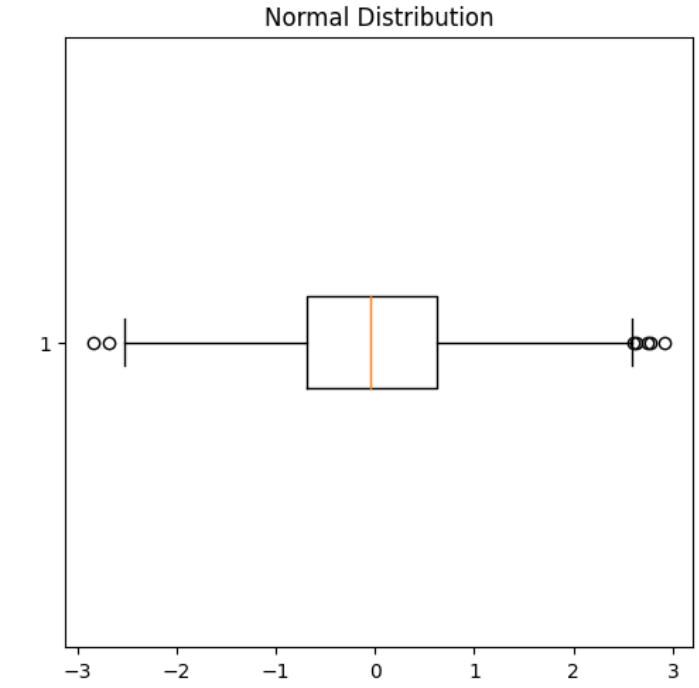
- `import matplotlib.pyplot as plt`
- `plt.hist(data)`
- `plt.boxplot(data)`

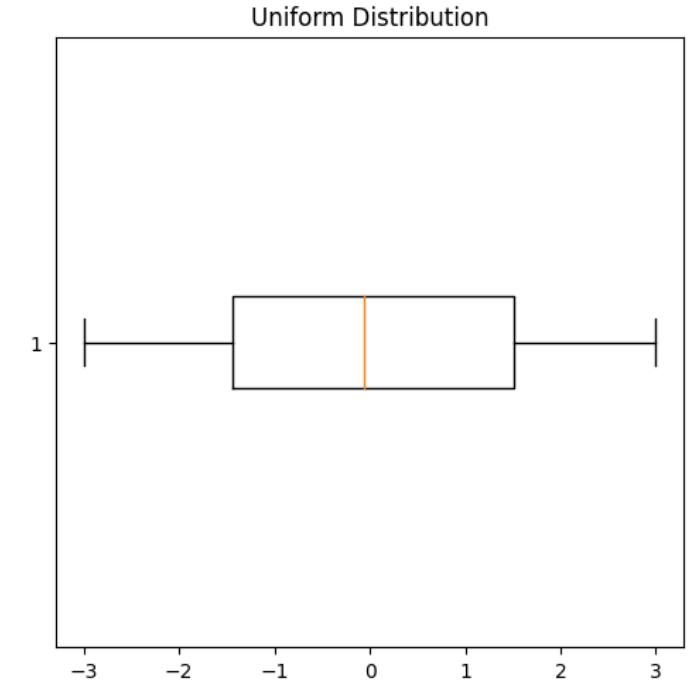
Understanding Data

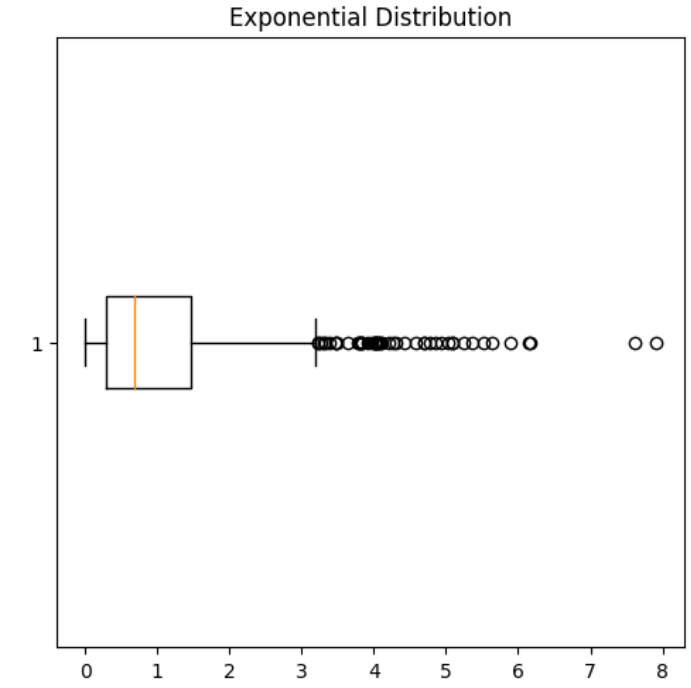












Hypothesis Testing

- A fundamental tool in data analysis. Used to make informed decisions based on data.
- To determine if there is enough evidence to reject a hypothesis about a population parameter.
- Involves formulating hypotheses, collecting data, and drawing conclusions
- Two types of hypotheses:
  - Null Hypothesis (H_0): Represents the status quo or no effect.
  - Alternative Hypothesis (H_a): Represents the proposed effect or change.
- Example:
  - H_0: There is no difference in test scores before and after a training program.
  - H_a: There is a significant improvement in test scores after the training program.

Hypothesis Testing Error

- Type 1 Error (False Positive): Occurs when a null hypothesis that is actually true is rejected.
  - It represents a situation where the test incorrectly detects an effect or difference that doesn't exist.
  - The probability of Type 1 error is denoted as "alpha" (α) and is typically set as the significance level (e.g., 0.05) in hypothesis testing.
  - Minimizing Type 1 error is important when the cost or consequences of making a false positive decision are high.
- Type 2 Error (False Negative): Occurs when a null hypothesis that is actually false is not rejected.
  - It represents a situation where the test fails to detect a real effect or difference that exists.
  - The probability of Type 2 error is denoted as "beta" (β).
  - Minimizing Type 2 error is crucial when failing to detect a true effect can have significant implications, such as in medical testing or quality control.
- Example: There is a glass of water on a table. Null: It is water. Alternative: It is H2SO4. What type of error is more dangerous?

Hypothesis Testing Steps

- Data Collection: Gather relevant data through experiments, surveys, or observations.
- Formulate Hypotheses: Define null and alternative hypotheses based on research questions.
- Select Significance Level (α): Determine the acceptable level of Type I error (false positive).
- Perform Statistical Test: Choose an appropriate statistical test (e.g., t-test, chi-square, ANOVA) based on data type and research question.
- Determine P-value: Calculate the p-value, which represents the probability of obtaining results as extreme as those observed, assuming the null hypothesis is true.
  - Smaller p-values indicate stronger evidence against the null hypothesis.
- Make a Decision: Compare the p-value to the chosen significance level (α).
  - If p-value $< \alpha$, reject the null hypothesis.
  - If p-value $\geq \alpha$, fail to reject the null hypothesis.

T-tests

- A T-test is a statistical test used to compare the means of two groups.
- It helps determine if there are significant differences between the groups.
- **Types of T-Tests**:
  - Independent samples t-test: Compares means between two different groups.
  - Paired sample t-test: Compares means from the same group at different times.
  - One-sample t-test: Tests the mean of a single group against a known mean.
- **Criteria for Use**:
  - Normally distributed data.
  - Scale (interval or ratio) data.
  - Random sampling from the population.
- **Common Applications**:
  - Comparing test scores of two different groups of students.
  - Assessing the effect of a treatment in a before-and-after study.
  - Testing hypotheses in experimental research.

ANalysis Of VAriance (ANOVA)

- ANOVA is a statistical method used to test the differences between two or more means.
- Commonly used when comparing three or more groups.
- Types of ANOVA include:
  - One-way ANOVA: Tests the effect of a single factor.
  - Two-way ANOVA: Tests the effect of two independent variables.
  - MANOVA (Multivariate ANOVA): Tests multiple dependent variables.
- Assumptions:
  - Normal distribution of the dependent variable.
  - Homogeneity of variances (equal variances across groups).
  - Independent observations
- **Post-hoc Analysis**: To determine which specific groups differ after an ANOVA indicates a significant difference exists.
  - Helps in interpreting the results of an ANOVA by providing detailed pairwise comparisons.
- Most Common: Tukey's Honestly Significant Difference (HSD) Test

Chi-Square Test of Independence

- The Chi-Square Test of Independence is a statistical method used to determine if there is a significant association between two categorical variables.
- Purpose:
  - To test the independence of two variables.
  - Commonly used in feature selection, hypothesis testing, and market research.
- Formulate Hypotheses:
  - Null Hypothesis (H0): Variables are independent.
  - Alternative Hypothesis (H1): Variables are not independent.
- If the p-value is less than α, reject H0.