Weekly Report – Week 2

Summary for the Week:

- Statistics Basic Definition
    - Event
    - Probability
    - Joint Probability
    - Conditional Probability
    - Bayes Rule
- Random Variables
    - Discrete random variables
    - Continuous random variable
- Distribution of random variables
    - Bernoulli distribution
    - Uniform distribution
    - Normal distribution
    - Central limit theorem
- Data Wrangling
- Feature extraction
- Text data representation
- Data VS Signal
    - What is data?
    - What is Signal?
    - Difference between data and signal?
- Encoding and Distribution
- Scaling and Normalization
- Statistics in python
- Data loading and saving in python
- Data exploration in python
- Scaling, encoding and distribution in python

Reference for the report:

- Statistics for Data Science | Beginner's Guide to Statistics for Data Science (analyticsvidhya.com)
- Chapter 14 Random variables | Introduction to Data Science (harvard.edu)
- What Is Data Wrangling? A Complete Introductory Guide (careerfoundry.com)
- Feature Extraction Definition | DeepAI
- Text Representation for Data Science and Text Mining | by Ivo Bernardo | Towards Data Science
- Data vs Signal | Difference between Data and Signal (rfwireless-world.com)
- An Introduction to Data Encoding and Decoding in Data Science (sitepoint.com)
- A Gentle Introduction to Statistical Data Distributions - MachineLearningMastery.com
- Scaling and Normalization | Kaggle

Mathematics and ML:

Math is the foundation of machine learning. Every decision, iteration, and outcome are based on math. We must comprehend the many mathematical procedures and concepts that form the basis of your outcomes if you want to get exceptional achievements.

Some key areas:

- Vectors
- Matrices
- Probability
- Python programming

2.1 Statistics: Basic Definitions

- **Event:**

  In probability, an event is defined as a collection of results from a random experiment.

For a coin toss experiment, sample space $\Omega = \{'head', 'tail'\}$. And event A could be either $\{'head'\}$ or $\{'tail'\}$.

For a dice roll experiment, sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ and event $A = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$.

- **Probability:**

  Probability is a measure of the possibility that an event will occur and is defined for an event. A number between 0 and 1 is used to quantify it.

  The probability of an event *A* occurring is denoted as *P(A)*. The probability of an event *A* not occurring is denoted as *P(A) = 1 – P(A)*.

- **Joint Probability:**

  It is possible to define probability collectively for several events.

  Consider the experiment where we toss two coins. In this case the probability of seeing "*heads for coin toss 1*" and "*heads for coin toss 2*" is an example for two events. If two events, *A* and *B* are independent then the joint probability is *P (A and B) = P(A) P(B)*.

$$P(headsFirstToss\ and\ headsSecondToss) = P(headsFirstToss) \times P(headsSecondToss) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

- **Conditional Probability:**

  The likelihood of an event *A* given the occurrence of another event *B* is known as conditional probability.

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

  is not zero.

- **Bayes Rule:**

  The bayes rule describes the likelihood of an occurrence *A* depending on the likelihood of an associated event *B*.

  Example: if cancer is related to age, using Bayes' rule information about the person's age can be used to more accurately assess the probability that the person has cancer.

Bayes' rule is mathematically stated as:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \text{ in which } P(B) \neq 0.$$

2.2 Random Variable:

A random variable is a variable whose potential values are generated by an unpredictable event.

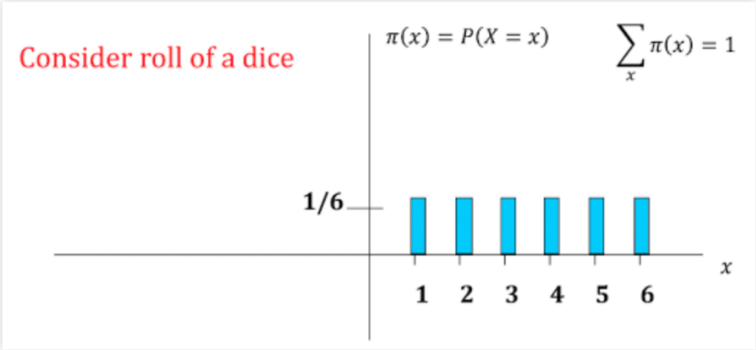
A random variable, then, is a function that can be used to assign probabilities to important events in a random experiment.

There are two categories of random variables:

- **Discreate Random Variable:**

  Discreate random variables have a finite number of possible values. Example, the number of dice sides, the number of emails received in an hour, etc.

  It is defined using a *Probability Mass Function (PMF),* denoted as π(x).



But what if someone inquiries about the likelihood of rolling a die and receiving a number less than 5?

In such instances, we must use the Cumulative Distribution Function (CDF). The cumulative distribution function calculates the cumulative probability of a function. It can be defined as:

$$F(X) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i).$$

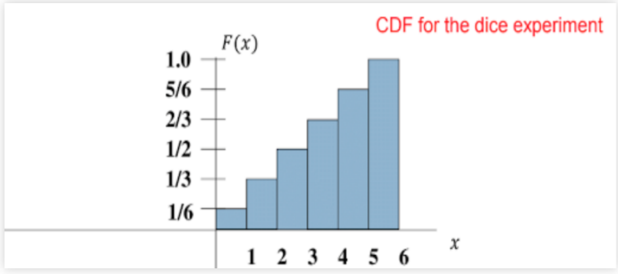


Figure. Cumulative distribution function in problem of rolling a dice.

The probability of seeing a number equal or less than five is $P(X \leq 5) = \frac{5}{6}$.

Based on the main property of probability, we can say the probability of seeing a number greater than five is $P(X > 5) = 1 - \frac{5}{6} = \frac{1}{6}$.

- **Continuous Random Variable:**

  Continuous random variables are defined using Probability Density Functions (PDF), denoted as *f(x)*. A probability density function is a statistical statement that specifies a probability distribution for a continuous random variable. PDF assigns a probability to a range of values of the random variable as *f(x) d(x) = P(x ≤ X ≤ x + dx)* integrating to 1.

  So we can say: $\int_{-\infty}^{+\infty} f(x)\,dx = 1$.

2.3 Distribution of Random Variables:

A probability distribution is a function that connects each outcome of a statistical experiment with its likelihood of occurrence.

Some of the important distributions:

- **Bernoulli Distribution:**

  The Bernoulli distribution is defined for a binary random variable with values X = 0 and X = 1.

  So $\pi(0) = P(X = 0) = p$ and $\pi(1) = P(X = 1) = 1 - p$.

  Sometimes, we use notations: $\pi(x) = B(x \parallel p)$ or $x \sim B(x \parallel p)$ where B means Bernoulli.

  For example, we can say a distribution over the outcome of an exam is Bernoulli. We may pass (x = 1) or fail (x = 0).

- **Uniform Distribution:**

  Uniform distribution can be defined for both discreate and continuous random values.

  - For discreate random variable:

$$\pi(x_i) = P(X = x_i) = \frac{1}{N}, \quad i = 1..N$$

- For continuous random variable:

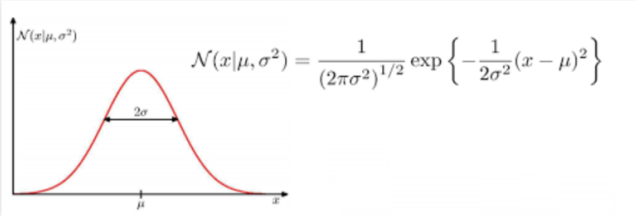
$$f(x) = \frac{1}{b - a}, a \leq x \leq b$$

- **Normal Distribution:**

The normal distribution is used to describe continuous random variables. It is by far the most popular distribution. A continuous random variable's normal distribution is defined as:

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$$

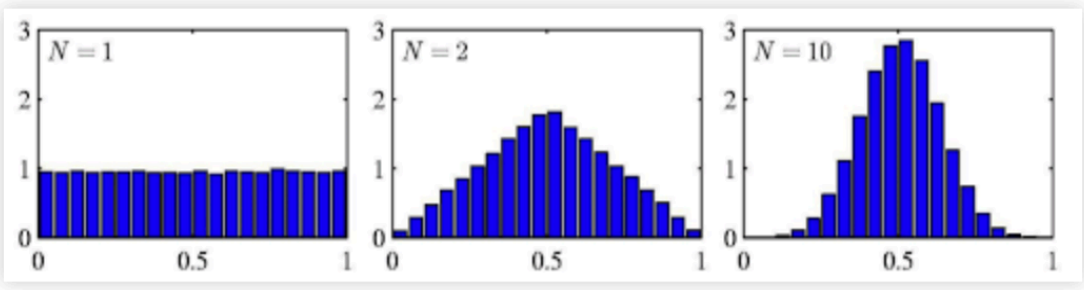
We denote $f(x) = N(x \mid \mu, \sigma^2)$ or $x \sim N(x \mid \mu, \sigma^2)$ where N means normal.



One of the reasons for the normal distribution's popularity is that many natural events mimic a normal distribution.

- **Central Limit Theorem:**

According to the Central limit theorem, if you have a population with a mean and standard deviation and take sufficiently large random samples from it, the distribution of the sample means will be nearly normally distributed.



2.4 Data Wrangling:

Data wrangling, also known as data munging, is the act of preparing a dataset for analysis by cleaning, manipulating, and organizing it. This is a critical phase in the data science pipeline that frequently requires a combination of manual and automated methods.

Some of the common tasks involved in data wrangling are:

- Identifying and correcting errors and inconsistency in data
- Handling missing or incomplete values
- Combing multiple datasets
- Converting the data into a format that is suitable for analysis
- Identifying and removing outliers
- Aggregating the data into useful summary statistics

Overall, data wrangling is an important phase in the data science process because it helps us to transform raw, unstructured data into a format suited for analysis and insights.

2.5 Feature Extraction:

One of the most important processes in machine learning is feature extraction. As we have already established, a complete representation of data is considered information. Similarly, for ML modeling, we must create a set of features from the raw data, each of which provides information about the target variable.

- Computers understands numbers better:

  The first step would be finding the features that can be represented as numbers. To use an image as input data for the algorithm, it needs to be represented in numerical vector of features.
- Creating a Model:

  We begin by dividing the image into smaller blocks. Each block, we can compute features of our choices:
  - Color averaged across the block
  - Shapes within the block
  - Textures in the block
  - Radiance or brightness

After converting the image to numbers, the resulting feature matrix can be input into a suitable computer algorithm for categorization as indoor or outdoor.

2.6 Text Data Representation:

Computers can only comprehend numbers. Before you can feed words, images, and ideas into a computer for processing, they must first be converted into numbers.

Data representation is a critical step in constructing models from massive amounts of data. Before using Machine Learning, data must be represented by 'features' known as attributes or parameters. It is critical to select the appropriate features while developing a model.

2.7 Data VS Signal:

- **What is Data?**

  Data is a collection of facts, such as numbers, words, measurement, observations or just description of things. It is the raw material that is used to create information. Data can be collected from a variety of sources, such as surveys, experiments, observations, and measurements.

  Data is used to create information. Information is data that has been processed and organized in a way that makes it meaningful. Information can be used to make decisions, solve problems, and understand the world.

- **What is Signal?**

  A signal is a piece of data that contains useful information. The signal is typically what we want to understand, while the noise is everything else that gets in the way.

- **Difference between data and signal?**

  Data and Signal are two terms that are often used interchangeably, but they have different meanings.

  - Data is a collection of facts, such as numbers, words, measurement, observations or just description of things. It is the raw material that is used to create information.
  - Signal is a specific type of data that contain useful information. The signal is typically what we want to understand, while the noise is everything else that gets in the way.

2.8 Encoding and Decoding

- **Encoding:**

  Unlike features with quantitative values, some features have categorical values that the machine cannot interpret. Encoding techniques are employed to convert to integer numbers to solve this problem.

  Some of the several well – known techniques of encoding such as OrdinalEncoding, One-Hot Encodings and LabelEncoding.

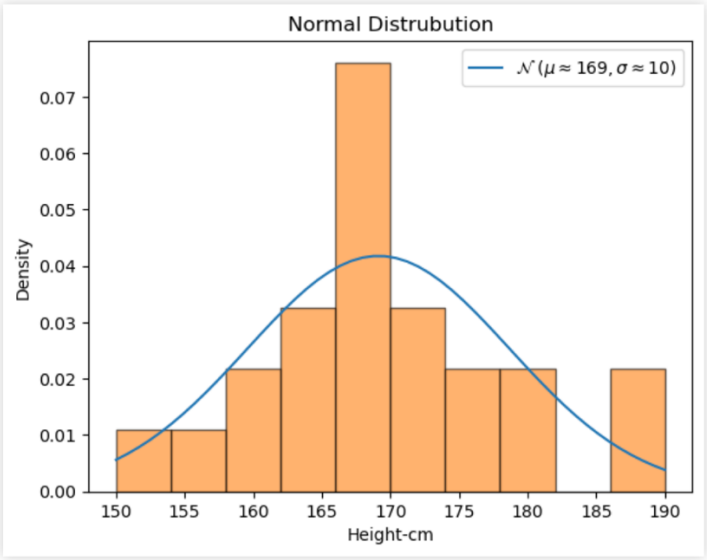
- **Distribution:**

  The distribution of values in a dataset is referred to as its structure. This is critical because the distribution of values has a substantial impact on the performance of a machine learning system.

  A dataset can contain a variety of distributions, including normal, uniform, and skewed distributions. The most frequent is the normal distribution, which is defined by a bell-shaped curve symmetrical around the data's mean. A uniform distribution has values that are uniformly dispersed across the data range, whereas a skewed distribution has values that are not evenly distributed and are instead focused on one side of the range. In general, the distribution of values in a dataset can have a wide range of effects on the effectiveness of a machine learning system. For example, if the values are not evenly distributed, the algorithm may be biased toward certain values, resulting in poor performance. On the other hand, if the data are normally distributed, the algorithm may find it easier to learn and anticipate.

$$\mu = \frac{\sum h_i}{N}$$

$$\sigma = \sqrt{\frac{(h_i - \mu)^2 + \ldots + \left((h_N - \mu)^2\right)}{N}}$$



2.9 Scaling and Normalization

- **Scaling**

  Scaling is the process of turning a collection of values to a new range of values in machine learning. We have various features in a dataset. The raw or unscaled features can have varying ranges, which can cause issues when training a model.

Student	Height	Weight
1	6.149	181.915
2	4.337	193.951
3	5.687	158.260
4	4.660	182.327
5	5.573	182.569
6	6.337	199.538
7	5.882	169.008
8	5.732	185.017
9	6.715	175.552
10	4.931	167.091

- **Normalization**

  Normalization is the process of transforming the values of a dataset to a common scale. This is often done to improve the performance of machine learning algorithms, which can be sensitive to the scale of the input feature.

  The common method of normalization is min-max normalization that scales the values of the dataset to a range of 0 to 1. This is done by subtracting the minimum value from each value in the dataset and then dividing by the difference between minimum and maximum values.

$$v' = \frac{v - min(v)}{max(v) - min(v)}$$

Student	Height	Weight
1	0.573	0.573
2	0.000	0.865
3	0.567	0.000
4	0.136	0.583
5	0.520	0.589
6	0.841	1.000
7	0.650	0.260
8	0.587	0.648
9	1.000	0.419
10	0.250	0.214

2.10 Statistics in Python

- **Random Variable**

```
import numpy as np
B = np.random.randn(4,3)
print('An example of a random matrix is:')
print(B)
```

The output of the above cell when executed will be:

```
An example of a random matrix is:
[[-0.39584372 -0.11740709  0.47348398]
 [ 0.10150287 -1.48922449  1.1987325 ]
 [-0.48927572 -0.04080551  1.19190827]
 [ 0.77228966  0.74481305 -0.0233208 ]]
```

2.11 Data Loading and Saving

- **Loading random data into pandas DataFrame**

```
import pandas as pd
import numpy as np
df = pd.DataFrame(np.random.randint(0,2000,size=(10, 6)),
columns=list('ABCDEF'))
df.head()
```

The output of the above cell when executed will be:

	A	B	C	D	E	F
0	290	958	211	1777	1880	321
1	1804	1191	47	687	1415	272
2	1053	516	717	230	1452	909
3	1771	709	500	654	478	1955
4	674	438	1286	879	1799	1371

- **Saving DataFrame**

```
df.to_csv("df.csv",index=None)
```

2.12 Data Exploration

```
import pandas as pd
df=pd.read_csv("data/Advertising.csv")
df.info()
```

The output of the code will be:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   TV         198 non-null    float64
 1   Radio      198 non-null    float64
 2   Newspaper  199 non-null    float64
 3   Sales      197 non-null    float64
 4   Types      200 non-null    object
dtypes: float64(4), object(1)
memory usage: 7.9+ KB
```

- **Data Description**

```
df.describe()
```

The output of the above code will be:

	TV	Radio	Newspaper	Sales
count	198.000000	198.000000	199.000000	197.000000
mean	148.223232	23.361111	30.508543	14.055330
std	85.463201	14.889023	21.824034	5.240709
min	0.700000	0.000000	0.300000	1.600000
25%	75.150000	9.925000	12.700000	10.400000
50%	150.650000	23.450000	25.600000	12.900000
75%	219.475000	36.575000	45.100000	17.400000
max	296.400000	49.600000	114.000000	27.000000

- **Finding the missing values and replacing them**

```
print("Check any values are null--
>",df.isnull().values.any())
print("How many values are null in individual columns or
attribute\n",df.isnull().sum())
df2=df.copy().bfill()
print("Check any values are null--
>",df2.isnull().values.any())
print("How many values are null in individual columns or
attribute\n",df2.isnull().sum())
```

2.13 Scaling, Encoding and Distribution

- **Categorical value encoding**

```
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
integer_encoded = label_encoder.fit_transform(df2["Types"])
df2["interger label"]=integer_encoded
```

- **Data Distribution**

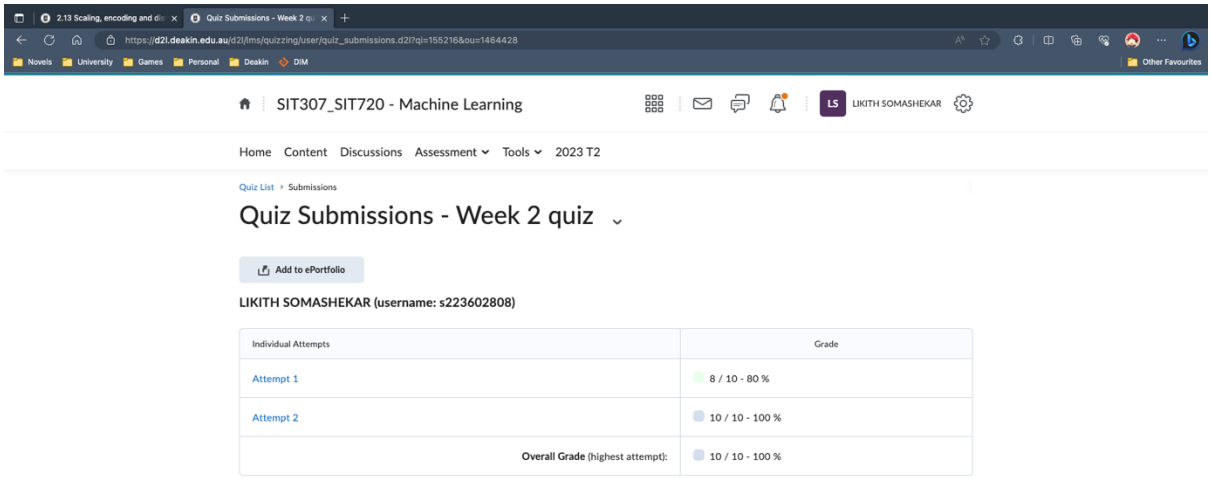
```
import matplotlib.pyplot as plt
df2[['TV', 'Radio', 'Newspaper', 'Sales']].hist()
plt.tight_layout()
plt.show()
```

- **Scaling**

```
df3=df2.copy()
df3[['TV', 'Radio', 'Newspaper', 'Sales']]=(df3[['TV', 'Radio',
'Newspaper', 'Sales']]-df3[['TV', 'Radio', 'Newspaper',
'Sales']].mean())
/df3[['TV', 'Radio', 'Newspaper', 'Sales']].std(ddof=0)
print("--------------------------")
print(df3[['TV', 'Radio', 'Newspaper', 'Sales']].mean())
print("STD:")
print("--------------------------")
print(df3[['TV', 'Radio', 'Newspaper', 'Sales']].std(ddof=0))

df3[['TV', 'Radio', 'Newspaper', 'Sales']].hist()
plt.tight_layout()
plt.show()
```

Week 2 Quiz


2.13 Scaling, encoding and di
Quiz Submissions - Week 2 qu
https://d2l.deakin.edu.au/d2l/lms/quizzing/user/quiz_submissions.d2l?qi=155216&ou=1464428
Novels
University
Games
Personal
Deakin
DIM
Other Favourites
SIT307_SIT720 - Machine Learning
LS
LIKITH SOMASHEKAR
Home
Content
Discussions
Assessment
Tools
2023 T2
Quiz List
Submissions
Quiz Submissions - Week 2 quiz
Add to ePortfolio
LIKITH SOMASHEKAR (username: s223602808)
Individual Attempts
Grade
Attempt 1
8 / 10 - 80 %
Attempt 2
10 / 10 - 100 %
Overall Grade (highest attempt):
10 / 10 - 100 %