Assignment #2 – Homework 2

**Course:** Ashoka Horizons: Applied Data Science with ML and AI  
**Instructors:** Rintu Kutum, Gautam Ahuja  
**Student:** Advay  
**Date:** 30th May 2025

Part 1: Probability and Statistics Easy (50%)

1. Define probability in your own words. What do probabilities of 0, 0.5, and 1 signify for an event?

Probability is the estimation of the likeliness an event is to occur, often represented on a scale of 0-1 or in percentages. Probabilities of 0, 0.5, and 1, each signify the likeliness’ of impossible, even chance, and certain respectively.

1. What is the probability of rolling a ’3’ on a standard six-sided dice? Show the favorable outcomes and total possible outcomes.

The probability of rolling a ‘3’ on a standard six-sized dice is 1/6 or 0.1666666667 ≈ 0.17.

1. List the three main measures of central tendency discussed.

Range, Standard Deviation, and Variance.

1. What is the primary purpose of descriptive statistics?

The primary purpose of descriptive statistics is to summarize and organize data in a way that makes it easier to understand. Instead of looking at all individual data points, descriptive statistics give us a quick overview through measures like mean, median, mode, range, and standard deviation. It helps to describe the main features of a dataset without making any predictions or assumptions beyond the data itself.

1. Define “Range” as a measure of dispersion. How is it calculated using the example test scores: 60, 70, 80, 90, 100?

Range is the difference between the maximum value and the minimum value in a particular dataset. It aims to measure the vastness of a dataset, in terms of its spread of values. Example using the data given above: Range = 100-60 = 40.

1. What is the key difference between “Variance” and “Standard Deviation” in terms of their units and interpretability?

The key difference between variance and standard deviation is mainly in their units and how easy they are to interpret. Variance is the average of the squared differences from the mean, so its units are the square of whatever units the data is in. For example, if the data is in meters, the variance will be in square meters.

On the other hand, standard deviation is the square root of the variance, which means its units are the same as the original data (like meters). Because of this, standard deviation is usually more interpretable and easier to relate back to the data, while variance is more useful in calculations and theoretical contexts.

Medium (30%)

1. Explain why understanding probability is crucial when working with Machine Learning models. Give one example from the slides.

Probability is crucial when working with Machine Learning Models as they are models which are predicting possible outcomes from the data fed to them. Probability is central when making decisions using any ML model (Hypothesis Testing, pg. 32 of Faculty Lecture W2), as it is the key concept which is driving the data analysis and extrapolating the data for us to show the likeliest favourable outcome based on the patterns and trends it analyses.

1. When would you prefer to use the Median over the Mean to describe the central tendency of a dataset? Provide an example scenario.

Median is best to be used in data which has outliers or is skewed, as mean can be influenced by the outliers while calculating as it takes all values. Whereas, median excludes the outliers due it picking the middle most value in sorted ascending or descending order of the data set. For example, if there is a data set and it is skewed, the mean would consider the extreme values of the outlier and misrepresent the actual central tendency of the data.

1. The slides mention “Data Exploration” as a reason why statistics is important in Data Science & ML. Explain what this means in a sentence or two.

Data Exploration is the discovery of patterns and trends, which reveal either causal or correlated relationships between variables being examined. These patterns and trends which are being evaluated is important to unearth newer connections in our world, the main crux of data science to generalise the data for humans to evaluate their surroundings better through data. Whilst in ML, data exploration is one of the main means for models to learn, especially in unsupervised models, which learn through data exploration, to make decisions and predict outcomes. Statistics is the main tool of data exploration, and rationalise data.

1. Briefly describe how a Case Study, like the one presented on Friedreich’s Ataxia (FRDA), highlights the importance of both data and methods (like statistics/ML).

The case study on Friedreich’s Ataxia (FRDA) really showed how important both data and the right methods are in making real progress. It wasn’t just about having a lot of data, it was about collecting the right kind of data, like genetic information, clinical symptoms, and progression over time. But that data alone wouldn’t have meant much without the use of statistical tools and machine learning techniques to find patterns, identify biomarkers, and make predictions. It showed me that without careful methods, even good data can be useless, and without good data, even the best methods won’t help. It is that both the model and the data matter in creating quality predictions and insights for a particular field. The two have to work together to generate meaningful insights, especially in something as complex as rare diseases.

Hard (20%)

1. Imagine a dataset of house prices in a city. Why might the standard deviation be very large? How could this affect your interpretation of the “average” house price if you only looked at the mean?

A city is an agglomeration of many kinds of individuals earning different incomes. These translate into different kinds of houses and localities. Real estate prices of different neighbourhoods can influence the average house price. The spread of the prices would also be large due to the wide variety of houses which may be present in a city. Therefore, the standard deviation of house prices in a city might be very large if there’s a lot of variation in the types of houses, like small apartments and huge luxury homes being in the same dataset. This wide spread means the prices are all over the place. If you only looked at the mean (average) price, it could give a misleading impression because a few extremely expensive houses could pull the mean up, making it seem like the “average” house is more expensive than what most people are actually paying, in an average neighbourhood. The data would be filled of outliers like expensive houses and localities, or even slums on the other extreme end of cheap housing. That’s why it’s important to look at the standard deviation or even the median to understand the data better.

1. The slides show a “Volcano Plot” in the context of discovering biomarkers. Without needing to understand all the biology, what do you think the plot is trying to show based on its axes (“log2(fold change)” and “−log10(adjusted p-value)”) and the colored dots? What might “up-regulated” and “down-regulated” mean in simple terms?

Even without understanding the biology, the volcano plot seems to be showing how different certain genes or biomarkers are between two groups, like healthy and diseased patients. The x-axis, “log2(fold change),” shows how much something increases or decreases—so values far to the left or right mean bigger changes. The y-axis, “−log10(adjusted p-value),” tells us how statistically significant those changes are—the higher the point, the more confident we are that the change isn’t just random. Coloured dots probably represent the most important or significant biomarkers. “Up-regulated” would mean that a gene or protein is more active or expressed more in the disease group, and “down-regulated” means it’s less active compared to the normal group.

Part 2: Machine Learning Fundamentals Easy (50%)

1. What is Arthur Samuel’s 1959 definition of Machine Learning?

“Machine Learning: A field of study that gives computers the ability to learn without being explicitly programmed.” – Arthur Samuel 1959.

1. List the “Big Three” types of Machine Learning.

Supervised Learning, Unsupervised Learning, and Reinforcement Learning.

1. In supervised learning, what is the difference between “Classification” and “Regression” tasks? Give one example of each from the slides.

Classification tasks are predicting a category or class of a piece of stimuli through the data given. The output is mostly discrete. An example from the slides could be a simple ML model trying to predict if an email is spam or not.

While, Regression tasks are predicting a continuous value. The output is a number. An example from the slides for a regression task is predicting the price of a house or even the amount of sales a company may make.

1. What is the main goal of Unsupervised Learning, according to the slides?

The main goal of unsupervised learning is to find hidden patterns, structures, or relationships, and make discoveries from the data.

1. What does PCA stand for and what is its primary purpose in unsupervised learning?

PCA is Principal Component Analysis, is a dimensionality reduction method. Its primary purpose is to simplify data, and speed up learning with visualisation. It can help by reducing the number of features while preserving the core information required for the learning or model for the evaluation being processed.

Medium (30%)

1. Explain the difference between traditional programming and machine learning in terms of their inputs and outputs.

In traditional programming, we give the computer both the rules (the code) and the input, and it follows those rules to produce an output. But in machine learning, we give the computer the input and the output (examples), and it figures out the rules or patterns on its own. So ML is more about learning from data instead of being told exactly what to do.

1. Briefly describe the core idea of “Learning from Examples” in Machine Learning, using the cat recognition analogy.

The idea of "learning from examples" in machine learning means that the model sees many examples of a situation and learns patterns from them. For example, in cat recognition, we don’t give the computer rules like “a cat has pointy ears and whiskers.” Instead, we show it thousands of pictures of cats and not-cats, and it learns the common features of what makes a cat. Over time, it gets better at telling whether a new image is a cat based on those learned patterns.

1. What is an “agent” in the context of Reinforcement Learning, and how does it learn?

In reinforcement learning, an “agent” is the thing that makes decisions and takes actions in an environment to achieve a goal. It learns by trying different actions and getting feedback in the form of rewards or penalties. Over time, it figures out which actions lead to the best long-term rewards and improves its behaviour based on that experience.

1. List two common ML algorithms for Supervised Learning and one for Unsupervised Learning mentioned in the slides.

Supervised Learning – Linear Regression, Logistic Regression

Unsupervised Learning – K-Means

Hard (20%)

1. The “Machine Learning Workflow” includes “Data Preprocessing” and “Feature Engineering.” Why do you think these steps are marked as “IMPORTANT!” and what kind of problems might occur if they are not done properly?

Data Preprocessing and Feature Engineering are marked as important because they directly affect how well the machine learning model performs. If the data is messy, incomplete, or contains errors, the model might learn the wrong patterns or give poor results. Preprocessing makes sure the data is clean, consistent, and usable, while feature engineering helps highlight the most relevant parts of the data for the model to learn from. If these steps are skipped or done badly, the model could be confused, biased, or totally inaccurate, even if it’s a very advanced algorithm.

1. Consider the spam email detection example. If a spam filter incorrectly marks an important email from your school as spam, what type of error is this in the context of classification (e.g., False Positive, False Negative)? Why might this type of error be particularly problematic?

If the spam filter marks an important email from your school as spam, that’s a false positive, it predicted something was spam when it wasn’t. This type of error is especially problematic because it can cause you to miss critical information, like deadlines, assignments, or announcements. In general, false positives in spam detection can lead to a loss of trust in the system and frustration if important emails keep getting filtered out.

Part 3: Artificial Intelligence Concepts Easy (50%)

1. What is the broad definition of Artificial Intelligence (AI) provided in the slides?

AI is the a broad field of computer science focused on creating systems that can perform tasks that usually require human intelligence.

1. According to the concentric circles diagram, what is the relationship between AI, Machine Learning (ML), and Deep Learning (DL)?

Deep learning is essential for machine learning, as it includes taking is a depth of data which will help train ML models better. After which, if the DL was data-rich, the ML models would be trained to give more and more accurate results. These results would be helpful to AI models which are based on resolving issues, which use ML models and DL processes to create results or responses, based on the deep variety of data learnt constructing the results.

1. List the three types of AI based on capability discussed in the slides. Which type do we have today?

Types of AI based on capability: Artificial Narrow Intelligence, Artificial General Intelligence, Artificial Superintelligence.

In today’s world we have artificial narrow intelligence.

1. Name two key areas that are considered “Foundations of AI.”

Search & Problem Solving, along with Knowledge Representation & Reasoning are two key foundations of AI.

Medium (30%)

1. Briefly explain the difference between AI “Thinking Humanly” and “Acting Rationally” as goals of AI, according to Russell & Norvig’s categories.

Thinking Humanly is the ability to conduct activities such as decision-making, problem solving, and learning. It means that AI systems try to think the way humans do, mimicking human reasoning, emotions, or other processes. It focuses on understanding and replicating how the human brain works. On the other hand, Acting Rationally is about creating system that make the best possible decisions to achieve a goal, regardless of whether they think like a human. It’s more focused on logic and effectiveness than on copying human behaviour. So, the main difference is that one is about how the AI thinks like a human, the other is about what it does ie. acting in the most logical and goal-orientated way, unlike humans which can be influenced by emotions and beliefs.

1. What is Natural Language Processing (NLP)? Give one example application mentioned.

Natural Language Processing enables computers to understand and generate human language. It focuses on human interaction with computers. A prime example could be the Google Translate, which focuses on machine translation of human languages in itself.

1. What is Generative AI, and how does it differ from AI models that only analyze existing data? Give an example.

Generative AI is what the name suggests, it focuses on the ability to generate content rather than just analysing existing data. A prime example is Open AI by ChatGPT, where it not only processes data, but is able to create responses out of scratch, or even videos from AIs like Sora.

Hard (20%)

1. The slides discuss “Ethical Considerations in AI,” including “Bias.” Explain how an AI model might learn biases from data and give a hypothetical example of an unfair outcome that could result.

An AI model can learn biases if the data it's trained on reflects unfair patterns from the real world. For example, if a hiring algorithm is trained on past company hiring data, and that data shows a pattern of mostly hiring men for tech roles, the model might learn to prefer male candidates, even if it's not told to. A hypothetical unfair outcome would be the model automatically ranking equally qualified female applicants lower just because it “learned” that pattern from biased data. This would be not only unfair but also harmful in reinforcing existing discrimination.

1. The concept of “Explainability” or “Transparency” in AI is becoming increasingly important. Why do you think it’s important to understand \*how\* an AI model makes its decisions, especially in critical applications like healthcare?

Explainability is really important, especially in areas like healthcare, because people need to trust and understand how decisions are being made. If an AI model recommends a treatment or diagnoses a disease, doctors and patients have the right to know why that decision was made. Without transparency, it’s hard to tell if the model is using the right reasons or just picking up on random patterns. It also becomes difficult to fix mistakes or hold someone accountable. In critical situations where lives are involved, blind trust in a “black box” model is risky; understanding the reasoning behind decisions helps make AI safer and more reliable.