HORIZONS-25-A2-HW

ASHOKA HORIZONS : APPLIED DATA SCIENCE WITH ML AND AI

Name - Tushar Joshi / Week#2 / Assignment-2

Part - 1 : Probability and Statistics

- Easy

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

- Probability is a mathematical tool that calculates the likelihood of an event to occur, in other words probability is the measure of how likely an event is to occur. If the probability is 0 then it signifies that the event is impossible to happen. If probability is 0.5, then it means the chances of the event to occur or to not occur is equal. And probability = 1, defines that the event is sure to happen.

2. What is the probability of rolling a '3' on a standard six-sided die? Show the favorable outcomes and total possible outcomes.

- ⅙ or 0.166……7.
- Favorable Outcomes = {6}
- Total Possible Outcomes = {1,2,3,4,5,6}

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

- Mean
- Median
- Mode

4. What is the primary purpose of descriptive statistics?

- The primary purpose of descriptive statistics is to summarize, and describe the main features of a dataset in a more clear, organized and

presentable way. What it helps in, is reducing large amounts of data to key value and visuals so to make better sense from it.

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

- Range shows - how spread out the values in a dataset are. In simple words it is the difference between the highest and the lowest values.
- In the given test score data, Range = Highest Value - Lowest Value i.e. 100-60 = 40.

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

- Variance is the average of the squared deviation from the mean, while Standard deviation is the square root of variance. The unit of standard deviation is the same as the unit of the original data while the unit of variance is square of the unit of original data. Standard deviation can be more easier to understand and compare directly while variance is less intuitive (harder to relate to original data).Variance gives you the spread in squared units (more mathematical), while Standard deviation gives you the spread in original units (more practical).

- Medium

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

- Understanding probability is crucial when working with machine learning models because real world data is often noisy and uncertain. Also machine learning models often give output in probabilities. Probability also helps in quantifying confidence in our findings and predictions. And it also works as a foundation in many statistical tests and machine learning algorithms.
- Example - In a machine learning model that detects spam emails, the model can give an output like 80% probability that this email is spam.

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

- Use the median when your dataset contains outliers or is skewed, as the mean can be distorted by extreme values.
- Example scenario: In reporting house prices in a city where most homes cost ₹50 lakh but a few cost ₹15 crore, the mean will be much higher than what most people actually pay. The median gives a better picture of a "typical" home price.

9. 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 refers to the process of analyzing and summarizing the main characteristics of a dataset—such as patterns, distributions, and relationships between variables. It helps data scientists gain insights, detect anomalies, and make informed decisions about how to preprocess or model the data effectively.

10. 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).

- A case study like the one on Friedreich's Ataxia (FRDA) shows how crucial both data and methods are in solving real-world problems. The quality and accuracy of data help researchers understand the disease, while methods like statistics and machine learning allow them to identify patterns, predict outcomes, and support diagnosis or treatment. It highlights that meaningful insights come not just from collecting data, but from analyzing it with the right techniques.

- Hard

11. 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?

- The standard deviation might be very large in a house price dataset if the prices vary widely — for example, if some houses are small apartments while others are luxury villas. This wide spread means that the “average” (mean) house price could be misleading, as it might not represent what most people actually pay. A few extremely expensive houses can raise the mean, making homes seem more expensive than they are for the typical buyer.

12. 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?

- A volcano plot visually highlights which biomarkers (like genes) show both a strong change and high statistical significance between two conditions. The x-axis shows how much a biomarker’s level changes (up or down), while the y-axis shows how confident we are in that change. “Up-regulated” means more active, and “down-regulated” means less active in one condition compared to the other.

Part - 2 : Machine Learning Fundamentals

- Easy

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

- “A field of study that gives computers the ability to learn without being explicitly programmed.”

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

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

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

- Classification predicts categories (e.g., spam or not spam), while regression predicts continuous values (e.g., house prices).

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

- To discover hidden structures, patterns, or relationships in data without using labeled outcomes.

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

- PCA (Principal Component Analysis) is a dimensionality reduction technique used to simplify data while preserving key information.

- Medium

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

- In traditional programming, we provide rules and data to get answers. In ML, we provide data and answers (examples) and let the system learn the rules(models).

19. Briefly describe the core idea of "Learning from Examples" in Machine Learning, using the cat recognition analogy.

- Just like a child learns what a cat looks like by seeing many examples, ML learns patterns from labeled data to recognize unseen instances of cats.

20. What is an "agent" in the context of Reinforcement Learning, and how does it learn?

- An agent interacts with an environment and learns by receiving rewards or penalties based on its actions – like a dog learning with treats.

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

- Supervised: Logistic Regression, Random Forest
- Unsupervised: K-Means Clustering

- Hard

22. 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?

- If data isn't cleaned or relevant features aren't created, the model may learn from noise, perform poorly, or make incorrect predictions. These steps ensure data quality and model accuracy.

23. 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?

- This is a False Positive – predicting "spam" when it's actually "not spam." It's problematic because important communication could be missed.

Part - 3 : Artificial Intelligence

- Easy

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

- AI is the field of computer science that creates systems capable of performing tasks that normally require human intelligence

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

- AI is the broadest field. ML is a subset of AI focused on learning from data. DL (Deep Learning) is a subset of ML that uses neural networks

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

- ANI (Narrow AI), AGI (General AI), ASI (Superintelligence). We currently have ANI (e.g., Siri, AlphaGo).

27. Name two key areas that are considered "Foundations of AI."

- Natural Language Processing (NLP)
- Computer Vision.

- Medium

28. Briefly explain the difference between AI "Thinking Humanly" and "Acting Rationally" as goals of AI, according to Russell & Norvig's categories.

- Thinking Humanly means mimicking how humans think (like reasoning or memory). Acting Rationally focuses on choosing the best action to achieve goals using logic, even if it doesn't think like a human

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

- NLP enables computers to understand and generate human language. Example: Sentiment analysis or chatbots like ChatGPT.

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

- Generative AI creates new content (e.g., images, text, code). In contrast, traditional AI just analyzes or classifies existing data. Example: DALL·E generating images from text.

- Hard

31. 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.

- AI can learn human biases present in training data. For instance, a hiring algorithm trained on biased resumes may unfairly prefer male candidates over equally qualified females.

32. 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?

- In critical fields like healthcare, we need to know *why* an AI makes a .diagnosis or decision. Without transparency, it’s hard to trust or correct the model when lives are at stake.