Assignment #2 [HW] - Horizons 25 (Data Science)

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 a numerical measure of how likely an event is to occur, ranging from 0 to 1.
   - 0: The event is impossible.
   - 0.5: The event is equally likely to occur or not.
   - 1: The event is certain.
     (Source: Slide 20)
2. What is the probability of rolling a '3' on a standard six-sided die?
   - Favorable outcome: Rolling a 3 → 1 outcome
   - Total possible outcomes: 6 (numbers 1–6)
   - Probability = 1/6≈0.167

     ≈0.167 or 16.7%
     (Source: Slide 21)
3. List the three main measures of central tendency discussed.
   - Mean
   - Median
   - Mode
     (Source: Slide 25)
4. What is the primary purpose of descriptive statistics?
   Descriptive statistics summarize and clarify the main features of a dataset, making complex data understandable and actionable. This step transforms raw numbers into insights, guiding decision-making in ML and AI projects
   (Source: Slide 25)
5. 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 highest and lowest values in a dataset.Range gives a quick sense of data spread, highlighting variability, which is crucial before deeper analysis

   Range = 100 − 60 = 40
   (Source: Slide 30)

6. What is the key difference between “Variance” and “Standard Deviation” in terms of their units and interpretability?
    - Variance uses squared units and measures the average squared deviation from the mean.
    - Standard Deviation is the square root of variance and retains the original units, making it more interpretable.
      (Source: Slide 31)

Medium (30%)

7. Why is understanding probability crucial in Machine Learning? Provide an example.

Probability is foundational in machine learning because it allows us to model uncertainty and make informed decisions based on data. Real-world data is often noisy, incomplete, or ambiguous, and probability helps quantify how confident we are in our predictions.

For example, in a spam email classifier, the model might output:

- “There is an 80% probability this email is spam.”

This doesn’t mean the email *is* spam, but that based on the model’s training, there’s a high likelihood. This probabilistic output allows for:

- Threshold tuning (e.g., only mark emails as spam if probability > 90%)
- Risk assessment (e.g., in medical diagnosis, a 95% probability of disease may warrant further testing)

Probability also underpins many ML algorithms:

- Naive Bayes uses Bayes’ Theorem.
- Logistic Regression outputs probabilities for classification.
- Bayesian Networks model dependencies between variables.

Key takeaway: Probability helps ML systems handle uncertainty and supports robust, data-driven decision-making.*(Example: Slide 23)*

---

8. When would you prefer to use the Median over the Mean? Provide an example.

The mean is sensitive to extreme values (outliers), while the median is more robust and represents the middle of the dataset.

You should prefer the median when:

- The data is skewed (not symmetrically distributed).

- There are outliers that could distort the average.

Example Scenario:
Imagine the incomes of 5 people in a neighborhood:

- ₹30,000, ₹35,000, ₹40,000, ₹45,000, ₹5,00,000
- Mean = ₹1,50,000 (heavily influenced by the one high income)
- Median = ₹40,000 (more representative of the typical resident)

In this case, the median gives a better sense of what most people earn.

Key takeaway: Choosing the right measure of central tendency ensures your summary statistics genuinely represent the data's reality

(Source: Slide 29)

---

9. What does "Data Exploration" mean in the context of statistics in Data Science & ML?

Data Exploration, also known as Exploratory Data Analysis (EDA), is the process of examining datasets to:

- Understand their structure
- Identify patterns, trends, and relationships
- Detect anomalies or outliers
- Assess data quality (missing values, inconsistencies)

This step is crucial before modeling because it helps you:

- Choose the right features
- Select appropriate models
- Avoid misleading results

Example Activities in EDA:

- Plotting histograms to see distributions
- Creating scatter plots to examine relationships
- Calculating summary statistics (mean, median, SD)

Key takeaway: Data exploration is the bridge between raw data and meaningful insights—it ensures that your modeling is grounded in a solid understanding of the data.

(Source: Slide 32)

---

10. How does the Friedreich's Ataxia (FRDA) case study highlight the importance of both data and methods?

The FRDA case study illustrates that data alone is not enough—you need the right statistical and machine learning methods to extract value from it.

In the case of FRDA, a rare genetic disease:

- Researchers collected biological data (e.g., gene expression).
- But to identify biomarkers (genes linked to the disease), they needed to apply:
    - Statistical tests to find significant differences
    - ML models to detect patterns and make predictions

This highlights a key principle in data science:

"Data is the raw material; methods are the tools that turn it into knowledge."

Without proper analysis, even rich datasets can remain inert and uninformative.

Key takeaway: The power of data is unlocked through the application of rigorous methods—statistics and ML are essential to derive actionable insights.
(Source: Slide 11)

Hard (20%)

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

In a city, house prices can vary dramatically due to factors like location, size, amenities, and neighborhood prestige. For example:

- A small apartment in a suburban area might cost ₹50 lakhs.
- A luxury penthouse in a prime location might cost ₹10 crores.

This wide range of values leads to a high standard deviation, which means the data points (house prices) are spread out far from the mean.

If you only consider the mean (average) price, it might be skewed by a few extremely high-priced properties. For instance:

- Suppose most homes cost between ₹50–80 lakhs, but a few luxury homes cost ₹5–10 crores.
- The mean might come out to ₹1.5 crores, which doesn't reflect what most people actually pay.

This is why relying solely on the mean can be misleading in such cases. A better approach is to also look at:

- Median: The middle value, which is more robust to outliers.

- Standard Deviation: To understand how much variation exists in the data.
  (Context: Slide 29 & 31)

12. What does a “Volcano Plot” show in the context of biomarker discovery? What do “up-regulated” and “down-regulated” mean?

A Volcano Plot is a type of scatter plot used in genomics and bioinformatics to identify genes that show significant changes in expression between two conditions (e.g., healthy vs. diseased).

- X-axis: $\log_2$(fold change)
  This shows how much a gene’s expression level has changed.
  - Positive values → Up-regulated (gene is more active in the condition being studied)
  - Negative values → Down-regulated (gene is less active)
- Y-axis: $-\log_{10}$(adjusted p-value)
  This represents the statistical significance of the change.
  - Higher values mean the result is more statistically significant (i.e., less likely due to chance).
- Colored dots:
  These typically highlight genes that are both significantly different in expression and statistically reliable.
  - For example, red dots might represent genes that are highly up-regulated and statistically significant.

This is crucial in identifying biomarkers—genes that could indicate the presence or progression of a disease.

Part 2: Machine Learning Fundamentals

Easy (50%)

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

Arthur Samuel, a pioneer in artificial intelligence, defined Machine Learning as:

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

This definition emphasizes the shift from traditional rule-based programming to systems that adapt and improve from experience. Samuel’s work on a self-learning

checkers program was one of the earliest examples of this concept in action.
(Source: Slide 35)

---

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

The three primary types of Machine Learning are:

1. Supervised Learning – Learning from labeled data (e.g., predicting house prices).
2. Unsupervised Learning – Discovering patterns in unlabeled data (e.g., customer segmentation).
3. Reinforcement Learning – Learning by interacting with an environment and receiving feedback (e.g., training a robot or playing a game).
   (Source: Slide 39)

These categories cover most real-world ML applications and help determine the right approach based on the problem and data available.

---

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

In Supervised Learning, the model learns from labeled data. The two main types of tasks are:

- Classification: Predicts discrete categories or classes.
  - *Example*: Classifying emails as Spam or Not Spam.
    (Source: Slide 41)
- Regression: Predicts continuous numerical values.
  - *Example*: Predicting the price of a house based on features like size and location.
    (Source: Slide 43)

The key difference lies in the type of output: categories vs. continuous values.

---

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

The main goal of Unsupervised Learning is to uncover hidden patterns, structures, or relationships in data without using labeled outputs.

This is useful when we don't know what to look for in the data, such as:

- Grouping similar customers (clustering)
- Reducing complexity in high-dimensional data (dimensionality reduction)

It enables exploratory analysis and pattern discovery in large, unstructured datasets.
(Source: Slide 44)

---

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

PCA stands for Principal Component Analysis.

Its primary purpose is dimensionality reduction—it simplifies complex datasets by transforming them into fewer dimensions (called principal components) while preserving as much important variance (information) as possible.

This helps:

- Visualize high-dimensional data
- Speed up learning algorithms
- Reduce noise and redundancy

Example: In a dataset with 100 features, PCA might reduce it to 2 or 3 key components that still capture the essence of the data.
(Source: Slide 46)

Medium (30%)

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

Traditional Programming follows a rule-based approach:

- Input = Data + Explicit Rules (written by a programmer)
- Output = Result (computed by applying rules to data)

Machine Learning, on the other hand, flips this paradigm:

- Input = Data + Results (examples or labels)
- Output = Model or Rules (learned automatically)

Insight:
In traditional programming, humans define the logic. In ML, the system learns the logic from examples. This makes ML especially powerful for tasks where writing rules is hard (e.g., recognizing faces, detecting fraud).
(Source: Slide 36)

---

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

Expanded Explanation:
The core idea of ML is that a model can learn patterns from labeled examples and generalize to new, unseen data.

Analogy:
Just like a child learns to recognize a cat by seeing many pictures labeled “cat,” a machine learning model is trained on many labeled images. Over time, it learns to identify features (like ears, whiskers, tail) that are common in cats.

Once trained, the model can:

- Recognize new cat images it hasn’t seen before
- Distinguish cats from other animals

Insight:
This process of learning from examples is what enables ML systems to adapt and improve with more data. It’s the foundation of supervised learning.
(Source: Slide 38)

---

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

Expanded Explanation:
In Reinforcement Learning (RL), an agent is an autonomous entity that:

- Interacts with an environment
- Takes actions
- Receives feedback in the form of rewards or penalties

The agent’s goal is to maximize cumulative reward over time by learning which actions lead to better outcomes.

Example:
In a video game, the agent might try different moves. If a move leads to winning points, it gets a reward. Over time, it learns a strategy (policy) that helps it win more often.

Insight:
Unlike supervised learning, RL doesn’t rely on labeled data. Instead, it learns through trial and error, making it ideal for dynamic environments like robotics, game playing, or

autonomous driving.
(Source: Slide 47)

---

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

Supervised Learning Algorithms:

1. Linear Regression – Used for predicting continuous values (e.g., predicting house prices).
2. Decision Trees / Random Forests – Used for both classification and regression tasks. They split data based on feature values to make predictions.

Unsupervised Learning Algorithm:

- K-Means Clustering – Groups similar data points into clusters based on feature similarity (e.g., customer segmentation).

Insight:
These algorithms are foundational tools in ML. Supervised algorithms learn from labeled data, while unsupervised ones uncover hidden patterns without labels. Choosing the right algorithm depends on the problem type and data characteristics.
(Source: Slides 49–50)

Hard (20%)

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?

Expanded Explanation:

Both Data Preprocessing and Feature Engineering are foundational to building effective machine learning models. They are marked as “IMPORTANT!” because they directly influence the quality of the input, which in turn affects the accuracy, robustness, and generalizability of the model.

🔹 Data Preprocessing involves:

- Handling missing values
- Removing or correcting outliers
- Normalizing or scaling features

- Encoding categorical variables

If skipped or done poorly:

- The model may learn from noise or irrelevant patterns
- It may fail to converge or produce biased predictions
- Performance metrics like accuracy or precision may be misleading

🔹 Feature Engineering involves:

- Creating new features from raw data (e.g., extracting “day of week” from a timestamp)
- Selecting the most informative features
- Reducing dimensionality to avoid overfitting

If done incorrectly:

- The model may overfit (memorize noise instead of learning patterns)
- It may underfit (miss important trends)
- Important relationships in the data may be overlooked

Insight:
These steps are not just technical chores—they are strategic decisions that shape how the model “sees” the data. In many real-world projects, better preprocessing and feature engineering outperform complex models.

(Source: Slide 51)

---

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?

Expanded Explanation:

This is a False Positive error:

- The model predicted “Spam”, but the actual label was “Not Spam”.

Why is this problematic?

- Critical information is lost: You might miss a deadline, exam update, or assignment.
- Trust in the system decreases: Users may stop relying on the spam filter if it misclassifies important emails.

- Recovery is difficult: Important emails may be deleted or buried in the spam folder, especially if users don't check it regularly.

Broader Implications:

- In domains like healthcare, a false positive could mean unnecessary anxiety or treatment.
- In finance, it could block legitimate transactions.

Insight:
In classification tasks, the cost of different errors is not always equal. False positives in spam detection can be more damaging than false negatives, depending on the context. That's why many ML systems are tuned not just for accuracy, but for precision, recall, and cost-sensitive performance.

(Source: Slide 52)

Part 3: Artificial Intelligence Concepts

Easy (50%)

24. What is the broad definition of Artificial Intelligence (AI) provided in the slides?
Artificial Intelligence (AI) is broadly defined as a branch of computer science focused on building systems that can perform tasks typically requiring human intelligence—such as reasoning, learning, perception, decision-making, and language understanding.

John McCarthy, who coined the term in 1956, described AI as:

"The science and engineering of making intelligent machines."

Insight:
This definition highlights AI's dual nature: it's both a scientific pursuit (understanding intelligence) and an engineering challenge (building intelligent systems). AI aims to replicate or simulate human cognitive functions in machines.

(Source: Slide 54)

25. According to the concentric circles diagram, what is the relationship between AI, Machine Learning (ML), and Deep Learning (DL)?
The relationship is hierarchical and nested, often visualized as concentric circles:

- AI is the broadest field, encompassing all efforts to create intelligent behavior in machines.

- Machine Learning (ML) is a subset of AI that focuses on enabling systems to learn from data rather than being explicitly programmed.
- Deep Learning (DL) is a subset of ML that uses multi-layered neural networks to model complex patterns in large datasets.

Insight:
This structure shows how Deep Learning is a powerful technique within ML, which itself is a key pathway to achieving broader AI goals. Most modern AI breakthroughs (e.g., image recognition, language models) are driven by deep learning.

(Source: Slide 55)

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

the three types of AI based on capability are:

1. Artificial Narrow Intelligence (ANI) – Performs specific tasks very well (e.g., Siri, ChatGPT, AlphaGo).
   This is the type of AI we currently have.
2. Artificial General Intelligence (AGI) – Human-level intelligence across a wide range of tasks. Still theoretical.
3. Artificial Superintelligence (ASI) – Surpasses human intelligence in all domains. Highly speculative.

Insight:
Understanding these types helps us track AI's evolution. While ANI is already transforming industries, AGI and ASI raise profound questions about ethics, control, and the future of human-machine collaboration.

(Source: Slide 57)

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

The two foundational areas of AI are:

1. Natural Language Processing (NLP) – Enables machines to understand and generate human language.
   Applications: Chatbots, translation, sentiment analysis.
2. Knowledge Representation & Reasoning – Focuses on how machines store and use information to make logical decisions.
   Applications: Expert systems, semantic search.

(Source: Slide 58)

Medium (30%)

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

Source: Slide 59

Natural Language Processing (NLP) is a foundational area of AI that focuses on enabling computers to understand, interpret, and generate human language in a meaningful way.

It combines linguistics, computer science, and machine learning to bridge the gap between human communication and machine understanding.

Example Application (Slide 59):

- Sentiment Analysis – Determining whether a product review is positive, negative, or neutral. This is widely used in customer feedback analysis, social media monitoring, and brand reputation management.

Insight:
NLP powers many everyday technologies—like chatbots, voice assistants, and translation tools—and is central to how AI interacts with humans in natural, intuitive ways.

---

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

Source: Slide 62

Generative AI refers to a class of AI models that can create new content—such as text, images, music, or code—rather than just analyzing or classifying existing data.

- Traditional AI: Focuses on tasks like classification, detection, or prediction based on existing data.
    - *Example*: Identifying whether an image contains a cat.
- Generative AI: Learns the underlying patterns of data and uses them to generate entirely new outputs.
    - *Example*: Creating a realistic image of a cat that doesn't exist in the real world.

Example from Slide 62:

- DALL·E – A generative model that creates images from text prompts (e.g., “an astronaut riding a horse in a futuristic city”).

Insight:
Generative AI is revolutionizing creativity and content production. It’s not just about automation—it’s about augmentation, enabling humans to co-create with machines in fields like design, storytelling, and scientific discovery.

Hard (20%)

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.

Source: Slide 63

Improved Answer:

AI models learn by identifying patterns in historical data. However, if that data reflects existing societal biases, the model can amplify and perpetuate those biases—often in subtle and systemic ways.

How Bias Enters:

- Historical bias: If past decisions were biased (e.g., hiring mostly men), the model learns to replicate that pattern.
- Sampling bias: If the training data underrepresents certain groups, the model may perform poorly on them.
- Label bias: If human-labeled data reflects subjective judgments, those biases are encoded into the model.

Hypothetical Example:

A company trains a hiring algorithm on resumes of past successful employees. If the historical data favored male candidates due to unconscious bias, the model may learn to prioritize male-associated features (e.g., certain names, schools, or job titles), leading to fewer women being shortlisted, even when equally qualified.

Insight:
Bias in AI isn’t just a technical flaw—it’s a social and ethical issue. It can reinforce discrimination at scale, making it harder to detect and correct. That’s why fairness audits, diverse datasets, and inclusive design are critical in AI development.

---

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?

Source: Slide 63

In high-stakes domains like healthcare, finance, and criminal justice, AI decisions can have life-altering consequences. Explainability—also known as model interpretability—refers to our ability to understand why a model made a particular decision.

Why It Matters:

- Trust: Clinicians and patients are more likely to trust AI if they understand its reasoning.
- Accountability: If an AI system makes a harmful or incorrect decision, we need to trace its logic to assign responsibility.
- Bias Detection: Transparent models help identify if decisions are being influenced by irrelevant or discriminatory factors.
- Regulatory Compliance: In many sectors, laws require that automated decisions be explainable (e.g., GDPR’s “right to explanation”).

Example:

If an AI system diagnoses a patient with cancer, a doctor needs to know which features (e.g., tumor size, genetic markers, imaging patterns) influenced that decision. This allows the doctor to validate or challenge the AI’s output, ensuring patient safety.

Insight:
Explainability is not just a technical feature—it’s a moral and legal necessity. As AI systems become more complex (e.g., deep neural networks), developing tools for interpretability (like SHAP, LIME, or attention maps) becomes essential for responsible AI deployment.