7 Feb 2025

Date
Page

Pattern Case-Study — Hypothesis Testing

STATEMENT- A company wants to promote welness program for employees. They introduce a daily 10 mins meditation program and want to test if it significantly reduces stress levels after 4 weeks.

Step-1 Formulating Hypothesis Null and Alternative

NULL- There is no significant difference in stress due to meditation.

ALTERNATIVE - Meditation significantly reduces stress.

Step-2 Choosing the right test

Since we are comparing stress levels before and after meditation for the same employees, a paired t-test is appropriate.

Step-3 Setting the significant Level

Typically we choose $\alpha = 0.05$ (5% significance level).

step-4 Collecting Data

→ Measuring employee's stress level before the program (Pre-test).

→ Conduct daily 10 minutes meditation session for 4-weeks.

→ Measuring employee's stress levels after the program (Post-test).

→ Compute the difference in stress levels for each employee.

Step-5 Performing the Hypothesis test

→ Calculate the mean and standard deviation of the differences (Post-test − Pre-test).

→ Compute the t statistic using the formula,

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

where, $\bar{d}$ = mean difference
s_d = standard deviation of differences
n = sample size

→ Compare the t-statistic with the critical t-value form the t-distribution table at $n-1$ degrees of freedom

→ If p-value < 0.05, reject the null hypothesis.

Step-6 Conclusion

→ If we reject the null hypothesis, we conclude that meditation significantly reduces stress.

→ If we fail to reject the null, we conclude that there is no significant evidence to prove that meditation reduces stress.

Take-Home Test

A. Theory Based Question

1. A pattern is a recurring structure, trend, or regularity in data that can be recognised and analysed. Patterns help in identifying relationships, making predictions and understanding behaviours in various domains like machine learning, data science, image processing.

Example: Stock Market Trends — In financial data patterns like 'Heads and shoulders' or 'moving averages' help traders predict prices movements. For instance, if a stock price movement consistently rises after reaching a certain low points, this support level pattern can be used to anticipate future price trends.

2. Detection of Pattern Recognition & Difference from Anomaly Detection

Pattern recognition is the process of identifying regularities, structures or patterns in data using algorithms and statistical techniques. It involves classification input data into predefined categories based on similarity and learning models.

Difference From Anomaly Detection:

Pattern Recognition is a process of identifying regularities, structures or pattern in data usage algos and statistical techniques. It involves classifying input

data input predefined categories based on similarities and learnt model. While pattern recognition focuses on identifying and classification recurring patterns, anomaly detection aims to find deviations or outliers that do not conform to expected patterns.

	Feature	Pattern Recognition	Anomaly Detection
1.	Objective	Identifies and classifies patterns	Detects rare or unusual instances
2.	Data focus	Recognized common trends & structures	Finds deviations from normal behaviour
3.	Example	Handwritting recognition, speech recognition.	Fraud detection, network intrusion detection.

3. 3 Pattern Recognition Techniques,

1. Statistical Pattern Recognition,

→ Uses Probability and statistical models to classify data.

→ Example ÷ Naive Bayes classifier for spam detection

2. Structure Pattern Recognition,

→ Represents data as relationships (graphs, trees) to

recognize patterns.

→ Example : Parsing natural language sentences in NLP.

3. Neural-Network-Based Pattern

→ Uses deep learning models to learn complex patterns from large datasets.

→ Example : Convolutional Network Neural for Image recognition.

4. Supervised and Unsupervised Learning

Pattern recognition can be achieved through Supervised Learning and Unsupervised Learning depending on whether labelled data is available.

Feature	Supervised Learning	Unsupervised Learning
→ Defination	Learning from labelled data where the correct output is provided.	Learns patterns from unlabelled data without predefined outputs.
→ Objective	Classification & regression (predicting categories & values)	Clustering & pattern discovery (finding hidden structures).
→ Training Data.	Requires labelled data.	Required unlabelled data.

Example Algorithm	Decision Trees, SVM, Neural Networks.	K-Means, DBSCAN, Hierarchical clustering
Example Use Case.	Email spam detection (spam vs non spam)	Customer segmentation (grouping similar customers.)

Example in Real life :-

→ Supervised learning :- A CV Ranking system trained on resumes labeled as 'Highly suitable', 'Moderately suitable', or not 'suitable' based on past hiring decisions.

→ Unsupervised learning :- A job application clustering system that groups resumes based on similarities in skills, experiences and education without predefined labels.

5. Overfitting in Pattern Recognition

Overfitting occurs when a pattern recognition model learns not only the underlying patterns in the training data but also the noise and random fluctuations.

This makes the model perform exceptionally well on the training data but poorly on unseen (test) data, reducing its generalization ability.

Key characteristics of Overfitting :

1. High training accuracy, low test accuracy.
2. Model memorizes training data instead of learning general patterns.
3. Poor performance on new, unseen data.

Example in Pattern Recognition :

Suppose we are training a resume ranking model for your CV analyzer project. If the model is overfitted,

It will perfectly rank resumes correctly in the training sets.

It it fails to rank new resumes correctly because it memorized specific keywords from the training resumes rather than learning general hiring patterns.

Prevention of Overfitting :

1. **Regularization :** Techniques like L1 / L2 regularization prevents accessive complexity.
2. **Cross-Validation :** Splitting data into multiple subsets to ensure performance consistency.
3. **Pruning :** Removing less important branches on DT.

4. Dropout : Randomly deactivating neurons to prevent reliance on specific patterns.

5. Increase Training Data : More diverse data reduces the likelihood of memorization.

6. Anomaly Detection

1. Intrusion Detection in Neural Networks, Anomaly detection can identify unauthorised access or cyberattacks by monitoring network traffic. If a system usually receives 10 login attempts per hour but suddenly experience 1000 login attempts per hour but suddenly experiences 10000 login attempts, this unusual spike may indicate a brute-force attack or hacking attempt.

7. Anomaly Detection Calculation Over Time,

Anomaly in time-series data are identified by detecting significant deviations from normal trends. This can be done using,

→ Statistical methods (eg. moving average, Z-score, Standard deviation).

→ Machine learning models (e.g. LSTMs, Isolation forest, Autoencoders).

Time-Series example: Detecting Server Down Time,

Imagine a web server that normally handles 5,000 requests per hour,

Step-1: Establish Normal Behavior,

→ Compute a rolling average of request per hour.

→ Use a threshold (eg. mean ± 3 standard deviations) to define normal behaviour.

2. Step-2: Monitor in Real-Time,

→ If the server suddenly receives only 100 requests, the drop is far below the lower threshold, signaling a potential server failure or cyberattack.

Step-3: Alert & Response

→ The system raises an alert and triggers security measures to prevent downtime.

3. Pattern Recognition: Probability theory play a crucial role in pattern recognition by providing the framework to model uncertainity, make predictions, and classify data based on statistical principles, It helps in:

→ Quantifying Uncertainity: Pattern recognition often deals with incomplete, noisy, or ambiguous data.

Probability theory allows the model to make decisions even when full certainty is not available.

→ Modeling Data Distribution : Many pattern recognition methods, such as Naive Bayes or Gaussian Mixture Models (GMM), assume that the data follows a certain probability distribution. Understanding the likelihood of data belonging to different classes helps in making predictions.

→ Making Inferences : Techniques like bayesian inference use probability to update beliefs based on new evidence, which helps in continuously improving predictions.

Ex : In a CV evaluator, probability model could predict whether a resume belongs to a suitable or unsuitable category based on the likelihood of certain features, matching the desired pattern.

9. Concept of Maximum Likelihood estimation (MLE) in Pattern Recognition :

Maximum likelihood estimate is a statistical method used to estimate the parameters of a probability distribution or model that would make the observation data most likely. In the context of pattern recognition, MLE is used to train models by finding the parameter that maximize the likelihood of the data given the model.

How MLE works ?

1. Likelihood Function : Given a dataset, the likelihood function measures how likely it is to observe that data for different values of the model parameters.

$$L(\theta) = P(D|\theta)$$

where $L(\theta)$ is the likelihood, D is the data and θ are the model parameters.

2. Maximizing the Likelihood : To find the best model parameters, we maximize the likelihood function. This is equivalent to minimizing the negative log-likelihood.

$$\hat{\theta} = \text{argmax}_{\theta} L(\theta)$$

Example in Pattern Recognition :

When we are training a Gaussian Mixture Model (GMM) to classify different types of resumes based on experience levels, MLE helps estimate the mean and variance of each Gaussian distribution in the model, ensuring that the model best fits the observed data.

10. Confusion Matrix :

A confusion matrix is a performance measurement tool for classification models, providing a detailed breakdown of the model's prediction compared to the actual values. It helps evaluate how well the

model classifies different classes and is espically useful where there is an imbalance between the classes. The matrix is typically represented in a 2X2 table of binary classification, but it can also be extended for multi-class problems.

Structure of a Confusion Matrix

	Predicted Positive	Predictive Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

→ True Positive (TP): Correctly predicted positive class.

→ True Negative (TN): Correctly predicted negative class.

→ False Positive (FP): Incorrectly predicted positive class (Type 1 error).

→ False Negative (FN): Incorrectly predicted negative class (Type 2 error).

How Confusion Matrix is Used to evaluate a Classification Model :-

From the evaluation matrix, we can calculate several important matrices,

1. **Accuracy** ÷ The proportion of correctly classified instances,

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2. **Precision** ÷ The proportion of predicted positive instances that are actually positive,

$$\text{Precision} = \frac{TP}{TP + FP}$$

3. **Recall (Sensitivity)** ÷ The proportion of actual positive instances that sense correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

11. Numericals ÷

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{50}{50 + 10} = \frac{50}{60} = 0.8333$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{50}{50 + 5} = \frac{50}{55} = 0.9091$$

$$\text{F1 score} = 2 \times \frac{0.833 \times 0.9091}{0.8333 + 0.9091} = \underline{0.867}$$

12. Log-loss $= -\frac{1}{N}\sum_{i=1}^{N}(y_i \cdot \log(p_i) + (1-y_i) \cdot \log(1-p_i))$

For $y_1 = 1$ and $p_1 = 0.9$

1. $-(1 \cdot \log(0.9) + (1-1) \cdot \log(1-0.9))$

$= -\log(0.9) = 0.1054$

2. For $y_2 = 0$, $p_2 = 0.2$ $= 0.2231$

3. For $y_3 = 1$ & $p_3 = 0.8$ 0.2231

4. For $y_4 = 1$ & $p_4 = 0.8$ 0.5108

$\text{Log loss} = \frac{1}{4}(0.1054 + 0.2231 + 0.2231 + 0.5108)$

$= \underline{0.2656}$ Ans.

13.

	Predicted Yes	Predicted No.
Actual Yes	80	20
Actual No	10	90

Accuracy = $\frac{TP + TN}{FP + FN + TP + TN}$

$= \frac{80 + 90}{80 + 90 + 10 + 20} = 0.85$

$= 85\%$

Precision = $\frac{TP}{TP + FP} = \frac{80}{80 + 10} = \frac{80}{90} = 0.8889$

$= 88.89\%$

Recall = $\frac{TP}{TP + FN} = \frac{80}{80 + 20} = 0.8$

$= 80\%$

14. Polynomial Regression :

$X = [1, 2, 3, 4, 5]$

$Y = [2, 4, 9, 16, 25]$

Where need to fit polynomial of degree 2 quard eq and Predict Y when $X = 6$.

$Y = ax^2 + bx + c$

$y = x^2$

$a = 1, b = 0, c = 0$

$x = 6$

$y = 6^2 = 36$

15. Entropy Calculation :

$$E = -\sum_{i=1}^{N} P(i) \log_2 (P(i))$$

Total Instances = 30 + 20 + 10
= 60

$P(A) = 0.5$
$P(B) = 1/3$
$P(C) = 1/6$

$$E = -\left(0.5 \log_2 0.5 + \frac{1}{3} \log_2 \frac{1}{3} + \frac{1}{6} \log_2 \frac{1}{6}\right)$$

$$= -(0.5 + 0.528 + 0.431) = 1.459$$

Ans = 1.459

Case-Study Based.

16. Anomaly Detection in Healthcare : Identifying Unusual Heart Rates

In Healthcare, an AI system monitors patient's heart rates and uses pattern recognition and anomaly detection to identify unusual values that might signal health issues.

1. Pattern Recognition : The system models normal heart rate patterns based on historical data for each patient, using techniques like time-series analysis to establish a baseline.

2. Anomaly Detection : It depends deviations from the baseline, flagging heart rates that are too high or low.

Statistical methods, machine learning models, and threshold-based approaches help identify these anomalies.

For example if a patient's heart rate excedes unexpectedly to 130 bpm, the system flags it as an anomaly and alerts healthcare providers for further action. This enables early detection of potential health problems

17. In Banking Anomaly detection - Identification of Fraudulent Transactions.

In Banking, anomaly detection is used to identify fraudulent transactions by detecting unusual spending patterns.

1. Data Collection :- STEP-1

- The Bank collects transaction data such as transaction amount, location, time, merchant type, and user behaviour.
- Preprocessing is done to clean and standardize the data, handling missing values, and normalizing amounts.

2. Pattern Recognition :- STEP-2

- The bank builds normal spending profiles for each customer based on historical data, for example, typical transaction amount, and time models using techniques like clustering and time-series analysis.

3. Anomaly Detection :- Step-3.

- Supervised models like Random Forest or Logistic Regression can be trained on labelled data to classify transactions.
- Unsupervised models like Isolation Forests or Autoencoders detect anomalies by identification

transcation that deviates significantly from the customer normal patterns.

- Threshold-based models can be applied where transcation exceeding certain limits are flagged as potential fraud.

4. Real-Time Detection :- STEP-4.

The system flag unusual transcations in real-time, such as a large withdrawal from an unexpected location or multiple rapid purchases that dont fit the customers usual pattern.

- Alert are triggered to bank personnel or to the customer for verification.

18. In Manufacturing Plant For Equipment Failures Prediction and Prevention of Downtime.

In manufacturing plant for equipment failure prediction by analysing time-series data.

STEP-1 → Data Preprocessing

- Time-Series data is collected from sensors installed on equipment, measuring factors like temperature, vibration, pressure, operation speed over time.

- Preprocessing is done to clean the data, handle missing values and remove noise, ensuring high-

quality data for analysis.

STEP-2 → Normal Behavior Modeling :

- The plant uses historical data to model normal operating conditions for each piece of equipment. This can be done using statistical method or machine learning techniques like time-series forecasting.
- The model identifies typical patterns and trends in the equipment's behaviour during normal operations.

STEP-3 → Anomaly Detection :

- Unsupervised models like Isolation Forest, Autoencoders or One-class SVM are used to detect any deviations from normal behaviour in real-time.
- Thresholds are set based on historical data. If the equipment reading exceed these thresholds, it flag as a potential anomaly.

Step-4 → Real-Time Monitoring

The system continuously monitors the equipments sensor data in real-time, comparing it against the establishment normal patterns.

- Anomalies such as sudden spikes in temperature or unusual vibrations are detected early, signaling potential failure.

10. IBM Watson For Anomaly Detection In Network Security :-

IBM watson can be used for Anomaly detection in network security to detect unusual traffic patterns and potential security threats.

Step-1 → Data Collection :-

- Watson would collect data from network traffic such as packet headers, traffic flows volumes, IP addresses, protocols and timestamps.
- Preprocessing steps, like feature extraction are used to convert raw traffic data into useful features, like average packet size or communication frequency.

Step-2 → Monitoring Normal Traffic Patterns :-

- Supervised or Unsupervised machine learning models like are trained on historical data to identify normal traffic patterns.
- Techniques like clustering or time-series analysis could be used to understand expected traffic behaviours over time.

Date
Page

Step-3 → Anomaly Detection

- Supervised models can detect patterns in labelled data to distinguish between normal and abnormal traffic.
- Unsupervised models are used to detect unlabeled anomalies such as unusual spikes in traffic, sudden changes in IP address behaviour, or unexpected protocol use.
- Thresholds or anomaly scores can be used to flag traffic that deviates significantly from expected patterns.

Step-4 → Real-Time Monitoring

- Watson continuously monitors network traffic in real-time flagging any abnormal spikes, unusual protocols or unfamiliar IP addresses.
- Alerts are generated when suspicious activities, such as DDoS attacks or data infiltration are detected.

2. Document Analysis with supervised learning

In document classification, a company might use supervised learning to train a model to classify document into predefined categories.

Step-1 Data Collection :

- The company collects a labeled dataset of documents where each document is already assigned to a specific category.

Step-2 → Preprocessing :

- The text data undergoes preprocessing steps such as tokenization, stopword removal, stemming and lemmatization to clean and normalize the data.
- Feature extraction techniques like TF-IDF or word embedding convert the text into numerical features the text into numerical features that can be used by machine learning models.

Step-3 → Model Training :

- A supervised machine learning algorithm like Naive Bayes, SVM or Random Forest is used to train the model on the labeled data.
- A model learns to associate patterns in the document's features in with their corresponding categories.

Step-4 → Document Classification :-

- Once the model is trained, it can be used to classify new unseen document into categories based on the patterns it has learned.
- The model assigns a category label to a document based on their feature similarity to the labeled examples in the training data.

Additional Questions :-

21. Curse of Dimensionality :-

The curse of dimensionality refer to the problems that arise when the number of features in a dataset increases, leading to an exponential growth in data volume, which in turn makes analysis and computation more difficult. It :-

→ Increased complexity :- As the dimensionality grows, the distance between data points becomes less informative making it harder for models to distinguish b/w patterns.

→ Overfitting :- Higher dimensions increase the risk of overfitting as models may fit noise in the data instead of the actual underlying patterns.

Computational burden :- More dimensions lead to higher computational costs and longer training

times.

22. Hyperparameter Tuning in ML.

Hyperparameter tuning refers to the process of selecting the optimal set of hyperparameters for a machine learning model. Hyperparameters are parameters that are set before training a model, such as the learning rate, regularization strength or number of hidden layers in a Neural network.

Tuning these parameters helps improving the model's performance.

→ Grid search : Exhaustively searching through a predefined set of hyperparameters.

→ Random Search : Randomly selecting combinations of hyperparameters.

→ Bayesian optimization : Using probabilistic models to guide the search for optimal hyperparameters.

23. Difference between Precision and Recall

Precision : It is the ratio of true positive predictions to all positive predictions made by the model,

Precision = $\frac{TP}{TP + FP}$

→ Recall : It is the ratio of true positive predictions to all actual positive instances.

Recall = $\frac{TP}{TP + FN}$

Example, In email spam detection, precision is more important because we want to ensure that the email flagged as spam are actual spam.

If the model incorrectly marks a legitimate email to spam, it could lead to the loss of important information.

24. Information theory is used in feature selection to measure how much information a feature provides about the target variable. The key concept includes :-

1. Entropy :- Measures the uncertainity or unpredictable of a feature.

2. Mutual information :- Measures the amount of information shared b/w a feature and the target variable. Features with higher mutual information with the target are more informative and relevant

for pattern recognition.

25. **Regularization :-**

Regularization is a technique used to prevent overfitting by adding a penalty term to the model's loss function. This discourages the model from becoming too complex and fitting the noise in the training data.

1. **L2 Regularization (Ridge Regression) :-**

Adds the squared magnitude of the coefficient as a penalty to the loss function. It encourages smaller coefficients.

2. **L1 Regularization (Lasso Regression) :-**

Adds the absolute values of the coefficients as a penalty. It can drive some coefficients to zero, effectively performing feature selection. Regularization helps create a simpler model that generalizes better to unseen data.