Mock Practice Questions 2

Case Study: PricePredictor — Preventing Overfitting in Real Estate Price Prediction using Regularization Techniques

Background Story:

In the growing city of Urbanville, a startup named "PricePredictor" was building a machine learning model to predict real estate prices.

They collected data on:

- Number of Bedrooms
- Size in Square Feet
- Location Score
- Age of Property
- Nearby Facilities Count
- Distance from City Center

Initially, their model (Linear Regression) performed extremely well on training data but failed badly on test data — a classic case of *Overfitting*.

Solution:

To overcome this, their Data Scientist Ayesha applied *Regularization Techniques*:

1. L1 Regularization (Lasso Regression)

- Removed unnecessary features by shrinking some coefficients to zero.

2. L2 Regularization (Ridge Regression)

- Penalized large coefficients and ensured balanced weight values.

3. Elastic Net

- Combined L1 and L2 for better performance when features were correlated.

The team also tuned the *Regularization Parameter (λ / Alpha)* using Cross-Validation to find the best fit.

MCQs (Single Correct)

1. What is the main objective of Regularization in ML Models?

a) Increase Accuracy
b) Reduce Overfitting
c) Increase Dataset Size
d) Optimize Gradient Descent

---

2. Which Regularization Technique is used for Feature Selection?

a) Ridge
b) Lasso
c) Elastic Net
d) None

---

3. In Ridge Regression, what happens to the feature coefficients?

a) They become zero
b) They are shrunk towards zero but never become zero
c) They increase exponentially
d) They are removed from the model

---

4. Lasso Regression minimizes:

a) Sum of Squared Errors only
b) Sum of Absolute Errors only
c) Sum of Squared Errors + L1 Penalty
d) Sum of Squared Errors + L2 Penalty

---

5. Which metric is often used to tune Regularization strength (λ)?

a) Accuracy
b) Precision
c) Cross-Validation Score
d) Learning Rate

MSQs (Multiple Correct)

6. Overfitting in ML occurs when:

a) Model learns noise from data
b) Model performs poorly on training data
c) Model performs poorly on unseen data
d) Model generalizes well

---

7. Regularization helps in:

a) Reducing variance
b) Increasing model complexity
c) Preventing overfitting
d) Increasing number of features

---

8. Elastic Net is preferred over Lasso when:

a) Features are uncorrelated
b) Features are highly correlated
c) Number of features > Number of data points
d) When Ridge fails

---

9. Higher values of λ in Regularization will:

a) Increase model complexity
b) Shrink coefficients more
c) Lead to underfitting if too large
d) Remove all features

---

10. Select Regularization Techniques used in Machine Learning:

a) L1 Regularization
b) L2 Regularization
c) Dropout
d) Batch Normalization

📘 Case Study: EduStat — Analyzing Student Performance using Pandas & NumPy

🏫 Background:

The Kselis Institute in Learnville runs multiple programs across Math, Science, and Programming. They recently conducted a nationwide online test for over 5,000 students and stored their performance data in a CSV file with columns like:

- StudentID
- Name
- Subject
- Score
- Grade
- Attendance%
- City

To analyze this large dataset, the school's data analyst Tanya used NumPy and Pandas to answer key questions:

🔍 What Tanya Did:

1. Loaded the dataset using pd.read_csv()
2. Filtered students with attendance < 75%

3. Used groupby() to calculate average score per subject
4. Used NumPy's np.percentile() to find top performers
5. Replaced missing grades with 'Pending' using fillna()
6. Converted all City names to uppercase using vectorized operations
7. Calculated number of students per grade using value_counts()
8. Exported a cleaned version of the file using to_csv()

✅ MCQs (Single Correct Answer)

1. Which function is used to read a CSV file in pandas?

a) np.load()
b) pd.read_csv()
c) pd.import_csv()
d) np.read_csv()

---

2. Which NumPy function can you use to compute the 90th percentile of scores?

a) np.average()
b) np.median()
c) np.percentile()
d) np.decile()

---

3. What does df.groupby('Subject')['Score'].mean() return?

a) A single number
b) A new column
c) A Series with mean score for each subject
d) A count of scores

---

4. Which method replaces NaN values in a DataFrame column?

a) remove()
b) replace()
c) fillna()
d) dropna()

---

5. Which operation is vectorized in both NumPy and Pandas?

a) String concatenation
b) Looping through rows
c) Applying a condition to all elements
d) Creating a new file

✅ MSQs (Multiple Correct Answers)

6. What are benefits of using Pandas over plain Python lists for data analysis?

a) Faster performance on large datasets
b) Built-in functions for analysis
c) Better for images
d) Easy CSV and Excel handling

---

7. Select valid ways to filter students with score > 80 and attendance > 90% in a DataFrame df:

a) df[df.Score > 80 & df.Attendance > 90]
b) df[(df['Score'] > 80) & (df['Attendance'] > 90)]
c) df.query('Score > 80 and Attendance > 90')
d) df.Score > 80 and df.Attendance > 90

---

8. Which of these are NumPy array operations?

a) array.mean()
b) array.append()
c) np.percentile(array, 75)
d) array.upper()

---

9. Which pandas functions help inspect your dataset?

a) df.describe()
b) df.head()
c) df.tail()
d) df.search()

---

10. To write your cleaned DataFrame to a new CSV file, use:

a) df.save_csv()
b) df.write_file()
c) df.to_csv()
d) df.export_csv()

📘 Case Study: EduStat — Analyzing Student Performance using Pandas & NumPy

🏫 Background:

The Kselis Institute in Learnville runs multiple programs across Math, Science, and Programming. They recently conducted a nationwide online test for over 5,000 students and stored their performance data in a CSV file with columns like:

- StudentID
- Name
- Subject
- Score
- Grade
- Attendance%
- City

To analyze this large dataset, the school's data analyst Tanya used NumPy and Pandas to answer key questions:

🔍 What Tanya Did:

1. Loaded the dataset using pd.read_csv()
2. Filtered students with attendance < 75%

3. Used groupby() to calculate average score per subject
4. Used NumPy's np.percentile() to find top performers
5. Replaced missing grades with 'Pending' using fillna()
6. Converted all City names to uppercase using vectorized operations
7. Calculated number of students per grade using value_counts()
8. Exported a cleaned version of the file using to_csv()

✅ MCQs (Single Correct Answer)

1. Which function is used to read a CSV file in pandas?

a) np.load()
b) pd.read_csv()
c) pd.import_csv()
d) np.read_csv()

---

2. Which NumPy function can you use to compute the 90th percentile of scores?

a) np.average()
b) np.median()
c) np.percentile()
d) np.decile()

---

3. What does df.groupby('Subject')['Score'].mean() return?

a) A single number
b) A new column
c) A Series with mean score for each subject
d) A count of scores

---

4. Which method replaces NaN values in a DataFrame column?

a) remove()
b) replace()
c) fillna()
d) dropna()

---

5. Which operation is vectorized in both NumPy and Pandas?

a) String concatenation
b) Looping through rows
c) Applying a condition to all elements
d) Creating a new file

✅ MSQs (Multiple Correct Answers)

6. What are benefits of using Pandas over plain Python lists for data analysis?

a) Faster performance on large datasets
b) Built-in functions for analysis
c) Better for images
d) Easy CSV and Excel handling

---

7. Select valid ways to filter students with score > 80 and attendance > 90% in a DataFrame df:

a) df[df.Score > 80 & df.Attendance > 90]
b) df[(df['Score'] > 80) & (df['Attendance'] > 90)]
c) df.query('Score > 80 and Attendance > 90')
d) df.Score > 80 and df.Attendance > 90

---

8. Which of these are NumPy array operations?

a) array.mean()
b) array.append()
c) np.percentile(array, 75)
d) array.upper()

---

9. Which pandas functions help inspect your dataset?

a) df.describe()
b) df.head()
c) df.tail()
d) df.search()

---

10. To write your cleaned DataFrame to a new CSV file, use:

a) df.save_csv()
b) df.write_file()
c) df.to_csv()
d) df.export_csv()

🩺 Case Study: HeartRisk Predictor — Health Risk Classification using Logistic Regression

📖 Background:

In the health-conscious city of WellnessVille, a government hospital partnered with a data startup to launch HeartRisk Predictor — a system to identify high-risk patients for heart disease based on routine medical checkups.

Each patient's record included:

- Age
- Blood Pressure
- Cholesterol Level
- Gender
- BMI
- Diabetes Status (0 = No, 1 = Yes)

The target was a binary label:

- 1 = High Risk, 0 = Low Risk

Initially, the team tried Linear Regression, but it gave predicted values beyond 0 and 1 — making it unsuitable for classification. So, they switched to Logistic Regression.

🧠 Logistic Regression was used because:

1. It outputs probabilities between 0 and 1 using the Sigmoid function
2. It can be trained using Gradient Descent
3. The output threshold was set at 0.5:
    - Probability ≥ 0.5 → High Risk
    - Probability < 0.5 → Low Risk
4. Regularization was added to prevent overfitting

Tuning was done using metrics like Accuracy, Precision, Recall, F1-Score, and Confusion Matrix.

✅ MCQs (Single Correct Answer)

1. What activation function is used in Logistic Regression?

a) ReLU
b) Tanh
c) Sigmoid
d) Softmax

---

2. What is the output range of the Sigmoid function?

a) -1 to 1
b) 0 to 1
c) 0 to ∞
d) -∞ to ∞

---

3. Logistic Regression is most suitable for:

a) Regression Problems
b) Image Generation
c) Classification Problems
d) Clustering Problems

---

4. If predicted probability = 0.8, and threshold = 0.5, the predicted class is:

a) 0
b) 1
c) 2
d) Cannot say

---

5. Which of the following is NOT an evaluation metric for binary classification?

a) Accuracy
b) Recall
c) R^2 Score
d) F1-Score

✅ MSQs (Multiple Correct Answers)

6. Advantages of Logistic Regression include:

a) Simple and interpretable
b) Works well with linearly separable data
c) Can model multi-class problems (with extensions)
d) Doesn't require labeled data

---

7. Components of Confusion Matrix for binary classification:

a) True Positive (TP)
b) True Negative (TN)
c) Mean Squared Error (MSE)
d) False Positive (FP)

---

8. Which techniques help prevent overfitting in Logistic Regression?

a) L1 Regularization
b) L2 Regularization
c) Dropout
d) Cross-validation

---

9. Which of the following affect the performance of Logistic Regression?

a) Irrelevant features
b) Feature scaling
c) High correlation among features
d) Missing labels in training data

---

10. In which cases might Logistic Regression fail?

a) Highly non-linear data
b) Linearly separable data
c) When classes are imbalanced
d) When data has noise

📚 Case Study: BookFinder — Efficient Search in a Digital Library

📖 Background:

In the digital city of Readopolis, the local university launched an e-library system called BookFinder. The database contained over 100,000 books categorized by:

- BookID
- Title
- Author
- Genre
- Year

Students were allowed to search books using either BookID or Title.

Initially, BookFinder used Linear Search to search titles — scanning one by one. But with a growing database, the search became slow and frustrating.

To improve performance, developers introduced Binary Search — but only for sorted BookID lists.

💡 Key Concepts Used:

- Linear Search:
    - No need for sorted data
    - Scans each element one-by-one
- Binary Search:
    - Works only on sorted data
    - Divides search space into halves

The search engine automatically used Linear Search for Titles (unsorted) and Binary Search for BookID (sorted list).

✅ MCQs (Single Correct Answer)

1. Which condition must be true for Binary Search to work correctly?

a) Data must be unsorted
b) Data must be sorted
c) Data must be in reverse order
d) No condition

---

2. What is the worst-case time complexity of Linear Search?

a) O(log n)
b) O(1)
c) O(n)
d) O(n log n)

---

3. What is the best-case time complexity of Binary Search?

a) O(n)
b) O(log n)
c) O(1)
d) $O(n^2)$

---

4. What is returned if the key is not found in Binary Search?

a) 0
b) -1 or "Not Found"
c) None
d) Index of mid

---

5. Linear Search is more suitable when:

a) List is very long and sorted
b) Only one element exists
c) Data is unsorted
d) Data is numeric only

---

6. How many comparisons (approx.) will Binary Search need for 1024 elements in worst case?

a) 10
b) 100
c) 512
d) 1024

---

7. Binary Search follows which strategy?

a) Brute Force
b) Divide and Conquer
c) Greedy
d) Dynamic Programming

---

8. Which of the following is not true about Linear Search?

a) Can be used on unsorted data
b) Scans all elements
c) Works only with numbers
d) Time complexity is O(n)

---

9. What is the mid-point index in Binary Search for start = 0, end = 9?

a) 5
b) 4
c) 9
d) 0

---

10. Why might Binary Search perform poorly if data is unsorted?

a) It will still work
b) It may enter an infinite loop
c) It gives wrong output
d) Both b and c

🧪 Case Study: ModelMetrics — Evaluating ML Models for Health & Finance Sectors

📖 Background:

Two different ML teams in the city of DataVille were building models for two sectors:

1. Team A (HealthTech) built a classification model to predict whether a patient is likely to have diabetes (Yes or No) based on features like Glucose Level, Age, BMI, and Blood Pressure.

2. Team B (FinTech) built a regression model to predict house prices based on square footage, number of rooms, and location score.

After model training, both teams had to evaluate their models using different performance metrics.

💡 Team A (Classification) used:

- Accuracy: % of correctly predicted results
- Precision: How many predicted positives were actually positive
- Recall: How many actual positives were detected
- F1-Score: Harmonic mean of Precision and Recall
- Confusion Matrix: TP, TN, FP, FN values

💡 Team B (Regression) used:

- MAE (Mean Absolute Error): Avg. absolute difference between actual and predicted
- MSE (Mean Squared Error): Avg. squared difference
- RMSE: ✓MSE
- R^2 Score: Proportion of variance explained by the model

✅ MCQs (Single Correct Answer Only)

1. Which metric should be used when false positives are more dangerous than false negatives?

a) Accuracy
b) Recall
c) Precision
d) RMSE

---

2. If a classification model has high recall but low precision, it means:

a) It misses many actual positives
b) It makes many false positives
c) It is overfitting
d) It's highly accurate

---

3. The metric that gives equal importance to precision and recall is:

a) Accuracy
b) Specificity
c) F1-Score
d) MAE

---

4. Which metric is most suitable for a regression task?

a) Accuracy
b) Recall
c) RMSE
d) F1-Score

---

5. If R^2 score = 1, what does it indicate?

a) Poor model
b) Model is random
c) Perfect prediction
d) Overfitting

---

6. What does a confusion matrix show?

a) Number of model parameters
b) Accuracy over epochs
c) Count of TP, TN, FP, FN
d) Graph of loss function

---

7. Which metric is calculated as: TP / (TP + FN)?

a) Accuracy
b) Recall
c) Precision
d) F1-Score

---

8. Which error metric penalizes larger errors more than smaller ones?

a) MAE
b) MSE
c) R^2
d) Accuracy

---

9. Which of the following is not a classification metric?

a) Precision
b) Recall
c) Accuracy
d) RMSE

---

10. What is the range of R^2 score?

a) -1 to 1
b) 0 to 100
c) 0 to 1 only
d) $-\infty$ to ∞

📖 Background:

In Constructopolis, a construction analytics firm called BuildPrice AI built a system to predict the cost of building a house based on:

- Plot Area (in sq. ft.)
- Number of Floors
- Material Quality Score
- Age of the Building

👷 Problem:

Initially, they applied Linear Regression to predict the construction cost. However, the predictions were not very accurate for:

- Very large buildings
- Very old buildings

This indicated that the relationship was not linear — cost didn't just increase steadily but accelerated in some cases. So they upgraded the model to use Polynomial Regression (degree = 2), which captured the non-linear relationship better.

💡 What They Used:

- Linear Regression:
    - Simple model
    - Fast and interpretable
    - Used when relationships are linear

- Polynomial Regression:
    - Captured curves and non-linearity
    - Fitted with additional polynomial features like x^2, x^3
    - Trained using the same linear regression algorithm after feature transformation

They also used R^2 Score and RMSE to compare the performance of both models.

✅ MCQs (Single Correct Answer Only)

1. What kind of relationship does Linear Regression model?

a) Curved
b) Exponential
c) Straight-line
d) Random

---

2. Polynomial Regression is useful when:

a) Data has missing values
b) Features are binary
c) Relationship is non-linear
d) Dataset is very large

---

3. In Polynomial Regression, what is added to the model?

a) Random noise
b) More labels
c) Powers of input features (x^2, x^3, ...)
d) Dropout layers

---

4. Which of the following can result in overfitting in Polynomial Regression?

a) Too few data points
b) Too low learning rate
c) Very high degree of polynomial
d) Low variance features

---

5. Which metric helps compare both models to see which fits the data better?

a) Loss Curve
b) F1-Score
c) R^2 Score
d) Confusion Matrix

---

6. Linear Regression assumes the target variable changes:

a) Randomly with input
b) Linearly with input
c) Logarithmically
d) Periodically

---

7. What transformation is needed before applying Polynomial Regression?

a) Normalization
b) Adding polynomial features
c) Removing outliers
d) Encoding target variable

---

8. Which library in Python is commonly used to create polynomial features?

a) pandas
b) numpy
c) sklearn.preprocessing
d) seaborn

---

9. If the R^2 score is 0.92 for Linear and 0.97 for Polynomial Regression, what can be said?

a) Linear is better
b) Polynomial fits the data better
c) Both are equal
d) Cannot be determined

---

10. Polynomial Regression is still a type of:

a) Decision Tree
b) Linear Model (after feature transformation)
c) Logistic Model
d) Clustering Algorithm

📖 Background:

In the smart tech city of Algoville, a startup named EvalSuite built a dashboard for evaluating ML models across various domains:

- Classification Models for disease prediction
- Regression Models for house pricing
- Clustering Models for customer segmentation
- Autoencoders for anomaly detection

The goal was to show model performance using appropriate metrics depending on the model type.

🚀 Models and Metrics used:

Task	Model Type	Key Metrics
Disease Prediction	Classification	Accuracy, Precision, Recall, F1-Score
Price Prediction	Regression	MAE, MSE
Customer Segmentation	Clustering	Silhouette Score
Anomaly Detection (Autoencoders)	Reconstruction-based	Reconstruction Error

---

✅ MCQs (Single Correct Answer)

1. Which metric is best to evaluate a clustering model like K-Means?

a) Accuracy
b) Silhouette Score
c) F1-Score
d) RMSE

---

2. What does MAE (Mean Absolute Error) measure?

a) Average squared error
b) Average of true labels
c) Average of absolute prediction errors
d) Error rate in classification

---

3. Which metric is calculated as:

(2 × Precision × Recall) / (Precision + Recall)?
a) Accuracy
b) F1-Score
c) MAE
d) MSE

---

4. What is Reconstruction Error mostly used for?

a) Measuring accuracy of KNN
b) Evaluating regression curves
c) Measuring error in unsupervised autoencoders
d) Finding outliers in classification

---

5. When false positives must be minimized (e.g. spam filtering), which metric is most important?

a) Accuracy
b) Precision
c) Recall
d) MSE

✅ MSQs (Multiple Correct Answers)

6. Which of the following are classification metrics?

a) Accuracy
b) Recall
c) MSE
d) F1-Score

---

7. Select all that are regression evaluation metrics:

a) MAE
b) MSE
c) Precision
d) R^2 Score

---

8. When is Recall more important than Precision?

a) Cancer detection
b) Face recognition on phone
c) Email spam filtering
d) Fraud transaction detection

---

9. What are properties of Silhouette Score?

a) Ranges between -1 and 1
b) Measures how well points fit into clusters
c) Used for regression evaluation
d) High score means well-clustered data

---

10. Reconstruction error can be used to:

a) Detect anomalies
b) Evaluate classification models
c) Determine outliers in autoencoders
d) Calculate overfitting in CNNs

For any queries related to this question paper, feel free to contact **Prakhar Gupta (Batch 1)** at **24082700@scale.iitrpr.ac.in**

Thank You and All the best for end semester exams :)