Centre for Machine Intelligence & Data Science

e-Postgraduate Diploma (ePGD) in Artificial Intelligence and Data Science

Lecture 7
Programming for Machine Learning and Data Science

Lecture Flow

1. Regression Metrics - Overview
2. Mean Absolute Error (MAE)
3. Mean Squared Error (MSE) & Root Mean Squared Error (RMSE)
4. R-squared (R^2) -Coefficient of determination
5. Feature Analysis
   5.1 Feature scaling
   5.2 Feature Selection
   5.3 other feature analysis methods
6. Summary & Key Takeaways

Google Collab link:
https://colab.research.google.com/drive/1XageuGsPZdls6gZexHNiCNvz6U-FYENW#scrollTo=8ZQSRv0xGy_F

Regression Metrics

Q. What is the need of any metrics here?

ANS:

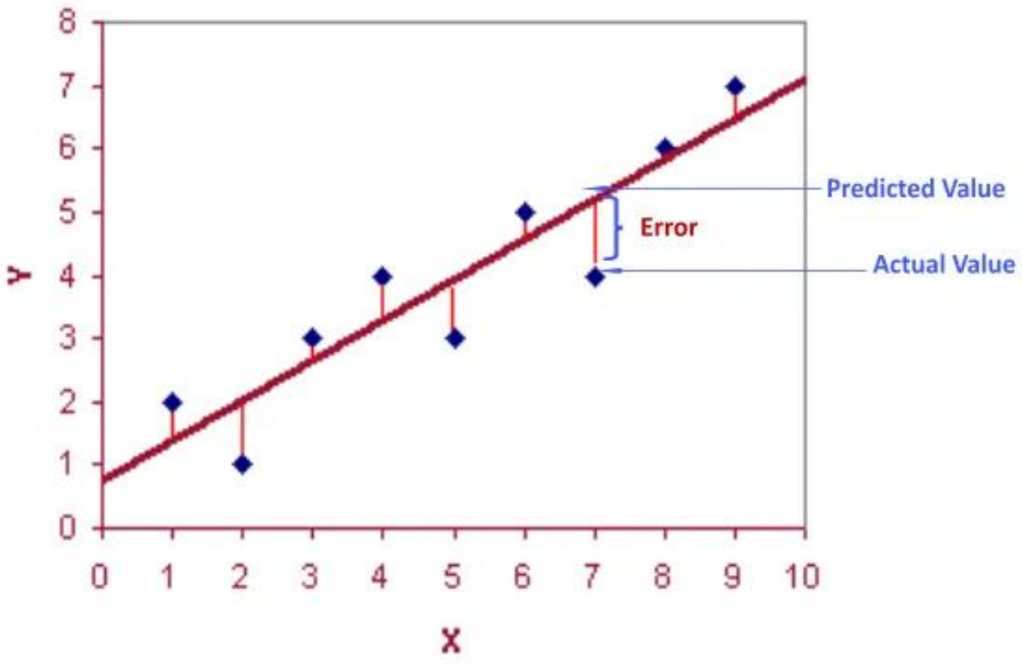
to measure how well its predictions actually match the observed data.

quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation.

Regression Metrics

- Regression performance is evaluated using various error metrics.
- Goal: Measure how well predictions approximate actual values.
- Key metrics:
  - Mean Absolute Error (MAE)
  - Mean Squared Error (MSE)
  - Root Mean Squared Error (RMSE)

Regression Metrics – Overview



https://medium.com/@mygreatlearning/rmse-what-does-it-mean-2d446c0b1d0e

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{Y}_i|$$

- Measures average absolute deviation between actual and predicted values.
- Advantages:
  - Intuitive and interpretable.
  - Less sensitive to large outliers than squared error metrics.

Limitation:
Does not differentiate between under-predictions and over-predictions.

Mean Squared Error (MSE) & Root Mean Squared Error (RMSE)

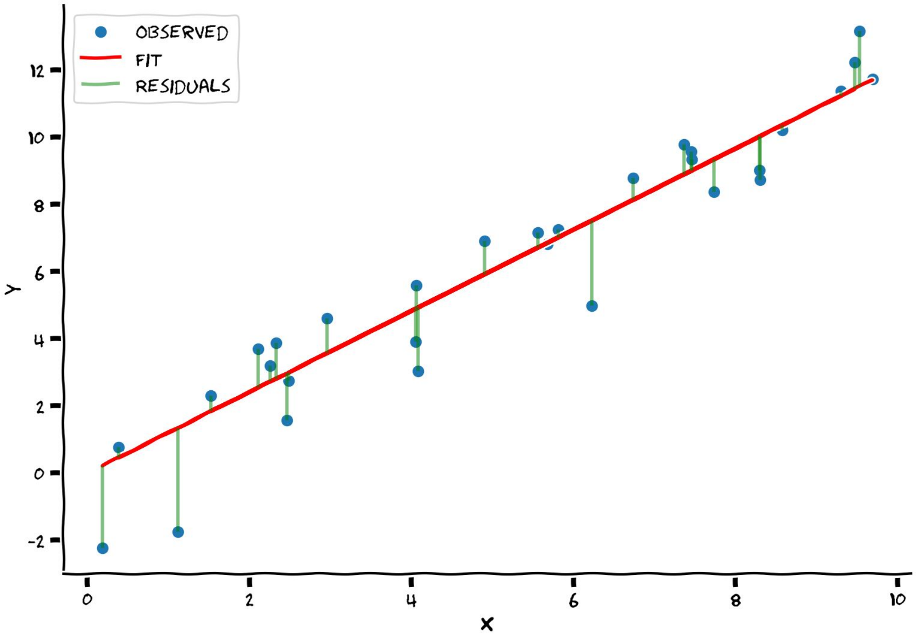
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$RMSE = \sqrt{MSE}$$

- MSE penalizes larger errors more heavily than MAE due to squaring.
- RMSE provides an error metric in the same units as the target variable.
- Trade-off: Sensitive to large errors (outliers), which can disproportionately impact evaluation.

$$\text{MSE} = \overbrace{\boxed{\frac{1}{n}\sum_{i=1}^{n}}}^{\text{Mean}}\Big(\overbrace{\boxed{Y_i - \hat{Y}_i}}^{\text{Error}}\Big)^{\overbrace{\boxed{2}}^{\text{Squared}}}$$

Mean Squared Error (MSE) & Root Mean Squared Error (RMSE)



https://compneuro.neuromatch.io/tutorials/W1D2_ModelFitting/student/W1D2_Tutorial1.html

- RMSE is the standard deviation of the residuals
- RMSE indicates average model prediction error
- The lower values indicate a better fit
- It is measured in same units as the Target variable

R-squared (R^2) - Coefficient of determination

- Formula:

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

where:

  - $SS_{residual} = \sum(Y_i - \hat{Y}_i)^2$ (sum of squared residuals)
  - $SS_{total} = \sum(Y_i - \bar{Y})^2$ (total variance in Y)

- Represents the proportion of variance in explained by the model.
- R^2 Range: [0,1]
  - Higher R^2 indicates a better model fit.
  - R^2=1 means the model perfectly explains variance.
  - R^2=0 means the model does not explain variance at all.

Adjusted R^2

- Adjusted R^2 corrects for overestimation when adding multiple predictors:

$$R^2_{adjusted} = 1 - \left(\frac{(1-R^2)(n-1)}{n-k-1}\right)$$

where n is the number of observations and k is the number of predictors.

Working with Real World Data:

features

Target variable

	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6	target
0	0.038076	0.050680	0.061696	0.021872	-0.044223	-0.034821	-0.043401	-0.002592	0.019907	-0.017646	151.0
1	-0.001882	-0.044642	-0.051474	-0.026328	-0.008449	-0.019163	0.074412	-0.039493	-0.068332	-0.092204	75.0
2	0.085299	0.050680	0.044451	-0.005670	-0.045599	-0.034194	-0.032356	-0.002592	0.002861	-0.025930	141.0
3	-0.089063	-0.044642	-0.011595	-0.036656	0.012191	0.024991	-0.036038	0.034309	0.022688	-0.009362	206.0
4	0.005383	-0.044642	-0.036385	0.021872	0.003935	0.015596	0.008142	-0.002592	-0.031988	-0.046641	135.0

Feature Analysis

Why Feature Analysis Matters?

- It is a part of exploring and visualing data to gain insights for better prediction
- Good feature selection improves **model accuracy & interpretability**.
- Avoids common issues like **multicollinearity, scaling problems, and outliers.**
- Helps select **relevant variables** for regression.

Collab link:
https://colab.research.google.com/drive/1XageuGsPZdls6gZexHNiCNvz6U-FYENW#scrollTo=8ZQSRv0xGy_F

Feature Scaling

Feature scaling is a data preprocessing technique used to standardize the range of independent variables or features in a dataset.

- Features with larger ranges can dominate the learning process, leading to biased models.
- Scaling ensures that each feature contributes equally to the model's performance.

ID	Age (years)	Income ($)	Ratings (1-5)	Has_Children (True/False)	Occupation (Job Type)
1	28	42,000	4	True	Engineer
2	45	85,000	5	False	Teacher
3	38	60,000	3	True	Nurse
4	52	100,000	4	False	Manager
5	33	55,000	2	True	Developer
6	29	48,000	5	False	Salesperson
7	41	70,000	3	True	Technician
8	36	65,000	4	False	Analyst
9	50	90,000	3	True	Consultant
10	27	38,000	4	True	Designer

https://medium.com/@sangeeth.pogula_25515/feature-scaling-explained-techniques-pitfalls-and-how-to-prevent-data-leakage-f7eeb47e7307

Feature Scaling

How to choose the appropriate scaling?

1. Standard scaling:

- Centers data around **zero** with a **standard deviation of 1**.
- Useful when features have **different units and ranges.**

$$X' = \frac{X - \mu}{\sigma}$$

2. Min-Max Scaling

- does not make any assumptions about the data distribution
- Scales features between a **fixed range** (default: **0 to 1**).

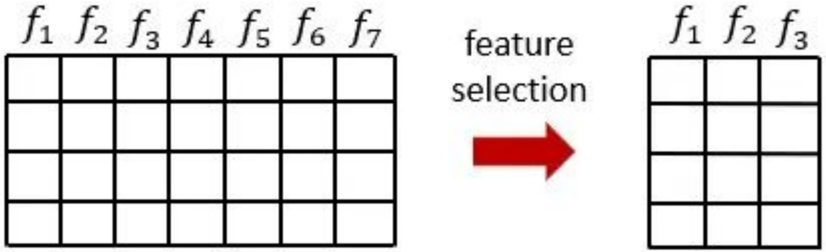
$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Feature selection

Definition: The process of selecting the most relevant features (independent variables) for a regression model.

Why is feature selection important:

- **Reduces noise** → Eliminates irrelevant or redundant features.
- **Prevents multicollinearity** → Avoids highly correlated features causing instability.
- **Improves model performance** → Leads to better generalization on unseen data.
- **Speeds up computation** → Fewer features = Faster model training.



https://medium.com/analytics-vidhya/feature-selection-extended-overview-b58f1d524c1c

Feature Selection

1. Feature Target Correlation:

- Measures **how strongly each feature is related** to the target variable.
- Helps in **feature selection** by identifying which features are most useful for prediction.

Example: scatter matrix plot, Correlation Coefficients

2. Feature Multicollinearity

Multicollinearity is when two or more features in a regression model are highly correlated.

- Detects **redundant features** that are highly correlated with each other.
- Helps avoid **unstable regression coefficients** caused by overlapping information.

Example: scatter matrix plot, Variance Inflation Factor

How to do Feature Selection?

Case 1: When Features are Few (Low Dimensional)

Use a Scatter Matrix Plot

- Helps **visualize pairwise relationships** between features and target.
- Useful when the dataset has **5-6 features**.
- Identifies **strong linear relationships** between independent variables.
- handling both feature target relation and multi colinearlity

Advantage:

- Easy to interpret and spot feature-target relationships.
- Can highlight **non-linear patterns** where linear regression may fail.

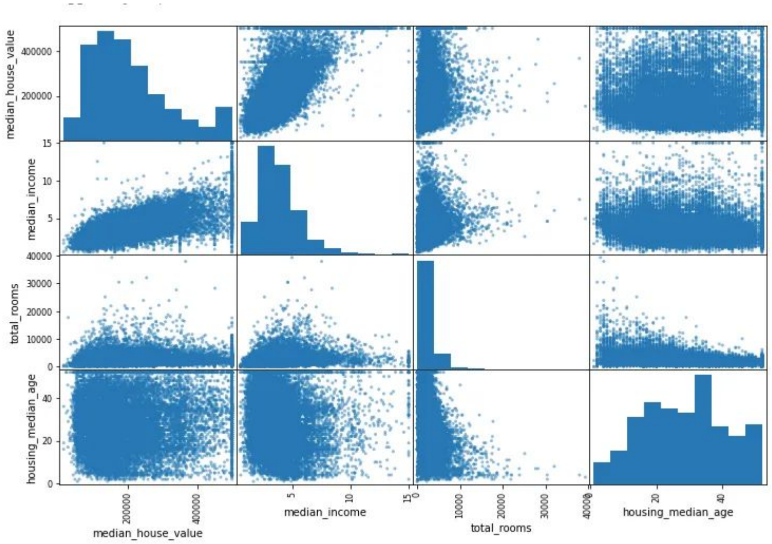


Fig: scatter matrix plot for california housing dataset

Feature Selection

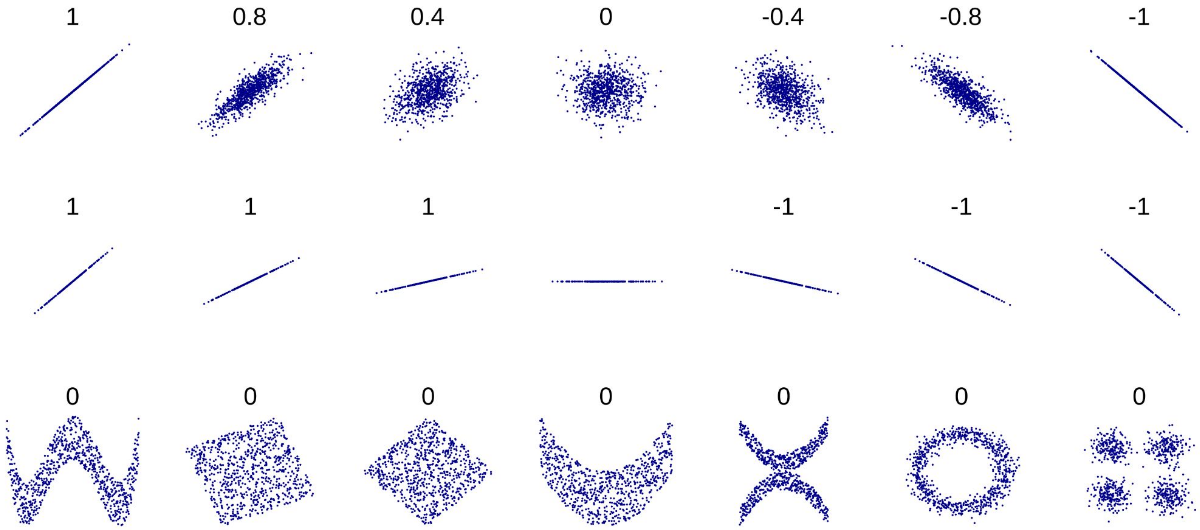


Fig: Standard Corrrelation coefficient of various type of datasets

https://en.wikipedia.org/wiki/File:Correlation_examples2.svg

Feature Selection

Case 2: When Features are Many (High Dimensional)

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} \times \sqrt{\sum(Y_i - \bar{Y})^2}}$$

Xi,Yi → Individual data points
bar{X} , bar{Y} → Mean of X and Y
Σ→ Summation over all data point

Use Correlation Coefficients for Feature Target Correlation

- Computes **Pearson correlation** to measure feature relationships.
- Helps detect **highly correlated (redundant) features**.
- Features with -**0.3**< **correlation <0.3** are usually **dropped**.

Pearson Correlation Coefficient (r) :

- The Pearson correlation coefficient measures the linear relationship between two variables.
- It quantifies how changes in one variable are associated with changes in another.

Feature Selection

Value of r	Interpretation
$r = 1$	Perfect **positive** correlation (X increases, Y increases)
$0.5 \leq r < 1$	Strong **positive** correlation
$0.3 \leq r < 0.5$	Moderate **positive** correlation
$-0.3 < r < 0.3$	Weak or **no correlation**
$-0.5 \leq r < -0.3$	Moderate **negative** correlation
$-1 \leq r < -0.5$	Strong **negative** correlation
$r = -1$	Perfect **negative** correlation (X increases, Y decreases)

Feature selection

Pearson correlation **does not account** for how a feature interacts with all other **independent features.**!

To detect Multicollinearity we use Variance Inflation Factor (VIF)

Variance Inflation Factor (VIF):
Variance inflation factor (VIF) is a statistical metric that measures how much the variance of a regression coefficient increases due to multicollinearity

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where R^2 is the Coefficient of determination for ith feature.

VIF < 5 → No multicollinearity (**Feature is fine**).
VIF > 5 → Moderate multicollinearity (**Consider removing**)
VIF>10 → Very high collinearity **(remove the feature)**

Other feature analysis Methods

3. Handling Outliers (Boxplot & IQR Method)

Why It's Important?

- **Extreme values can distort regression models** (especially MSE & RMSE).
- Outliers affect **coefficients** and increase **prediction errors**.

How to Detect Outliers?

- **Boxplot (Visual)**
- **Interquartile Range (IQR) Method**

Other feature analysis Methods

Interquartile Range (IQR):

The **Interquartile Range (IQR)** is a measure of **statistical dispersion** and is used to **detect outliers** in a dataset. It represents the **middle 50% of the data** by removing extreme values.

he **Interquartile Range (IQR)** is defined as:

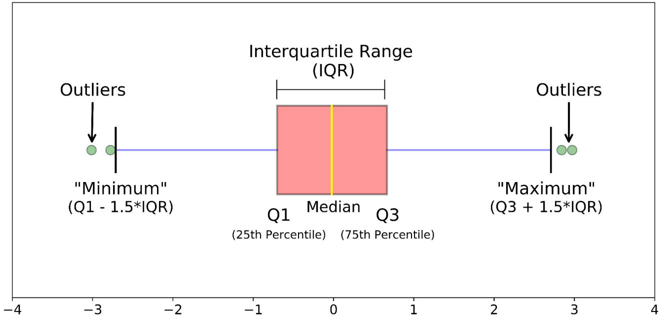
IQR=Q3−Q1

- **Q1 (First Quartile)** → The 25th percentile (25% of data is below this value).
- **Q3(Third Quartile)** → The 75th percentile (75% of data is below this value).
- **IQR**→ The range covering **the middle 50% of the dataset.**

An outlier is any value that falls outside this range:

Lower Bound=Q1−1.5×IQR
Upper Bound=Q3+1.5×IQR

- **If a value is less than Lower Bound, it is considered a low outlier.**
- **If a value is greater than Upper Bound, it is considered a high outlier.**



https://www.kdnuggets.com/2019/11/understanding-boxplots.html

Other feature analysis Methods

4.Feature Encoding – Handling Categorical Variables

- Many datasets contain **categorical variables** (e.g., job titles, house types).
- **Linear Regression cannot process categorical data directly**—it requires numerical input.
- Encoding ensures categorical variables are **interpreted correctly** by models.

Common Encoding Methods:

One-Hot Encoding (OHE)

- Converts categorical variables into **binary columns (0 or 1)**.
- Best for **nominal categories** (e.g., City Names, Colors).

Summary & Key Takeaways

- **Always visualize data before applying regression** to identify trends, outliers, and multicollinearity.
- **Feature scaling ensures fair contribution of all features** and prevents large-scale variables from dominating.
- **Experiment with different scaling methods** like Standard Scaling.
- **Feature selection improves model performance** by removing irrelevant or redundant features.
- **Use correlation matrices, VIF, and scatter plots** to detect multicollinearity and feature importance.
- **Blindly applying linear regression can lead to poor results** if data preprocessing is ignored.

Key takeaways

Experimentation with feature scaling and visualization is crucial before applying linear regression.

Well-prepared data leads to more accurate and interpretable models that generalize better

Questions?

“Good data beats fancy models. If features are wrong, no model can fix it!"

Collab link: https://colab.research.google.com/drive/1XageuGsPZdls6gZexHNiCNvz6U-FYENW#scrollTo=8ZQSRv0xGy_F