Lesson 4 — Regression

---

1. Introduction to Regression

Regression is a supervised learning method in machine learning that predicts **numbers** rather than categories. It tries to find the relationship between input features (independent variables) and an output value (dependent variable). In other words, regression answers the question **“how much”** or **“how many.”**

Classification is different because it predicts **labels or groups** instead of numbers. It answers **“which class”** something belongs to.

- **Example of Regression:** Predicting the amount of rainfall tomorrow (in millimeters) using weather data.
- **Example of Classification:** Predicting whether tomorrow will be rainy or sunny.

---

2. Types of Regression

a) Linear Regression

- **Idea:** Fits a straight line between one input and one output.
- **Use Case:** Predicting a student’s test score based on hours studied.
- **Advantage:** Simple and easy to understand.
- **Limitation:** Only works if the relationship is truly a straight line.

b) Multiple Linear Regression

- **Idea:** Uses many inputs at the same time to predict one output.
- **Use Case:** Predicting the price of a mobile phone using brand, storage, and camera quality.
- **Advantage:** Handles real-world problems with several factors.
- **Limitation:** Can become confusing when the inputs are strongly related to each other.

c) Polynomial Regression

- **Idea:** Adds powers of features (like x2x^2x2, x3x^3x3) to model curves.
- **Use Case:** Predicting the speed of a car as it accelerates over time.
- **Advantage:** Can fit curved patterns that linear models cannot.
- **Limitation:** Can easily overfit if the curve is too complex.

---

3. Regression Metrics

To check how good a regression model is, we use error metrics:

- **MAE (Mean Absolute Error):** Average size of mistakes, treating all errors equally.
- **MSE (Mean Squared Error):** Squares mistakes before averaging, so large mistakes count more.
- **RMSE (Root Mean Squared Error):** The square root of MSE, easier to understand because it has the same unit as the data.
- **R^2 (Coefficient of Determination):** Tells how much of the variation in the data is explained by the model (from 0 to 1).

Comparison Table

Metric	Meaning	Large Errors Matter?	Units
MAE	Average mistake	⍰ No	Same as target
MSE	Average squared mistake	⍰ Yes	Squared units
RMSE	Square root of MSE	⍰ Yes	Same as target
R^2	% variation explained	⍰ No	0–1 (or %)

4. Underfitting and Overfitting

- **Underfitting:** The model is too simple, so it misses patterns. Both training and test accuracy are poor.
- **Overfitting:** The model learns the training data too well, including noise, and fails on new data.

Why Overfitting Happens (especially in polynomial regression):

- Using too many features.
- Making the model too complex.
- Having very little training data.

How to Prevent Overfitting:

1. Use simpler models.
2. Apply regularization (L1 or L2).
3. Use cross-validation to test model performance.

5. Real-World Case Study

Case Study: Predicting Student Attendance in Schools

- **Goal:** Estimate the number of students who will attend class on a given day.
- **Data:** Past attendance records, weather conditions, and day of the week.
- **Model Used:** Linear Regression.
- **Results:** The model explained about 70% of the variation in attendance. It showed that rainy days and Mondays had lower attendance.

This helped school managers plan better, such as adjusting resources on low-attendance days.

References

- Alpaydin, E. (2020). *Introduction to Machine Learning*. MIT Press.
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O’Reilly Media.
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
- Khan, R. & Ali, M. (2022). Predicting school attendance using regression models. *International Journal of Educational Data Science*, 5(1), 33–41.