AIL303m_Course2

1. Introduction to Supervised Machine Learning and Linear Regression

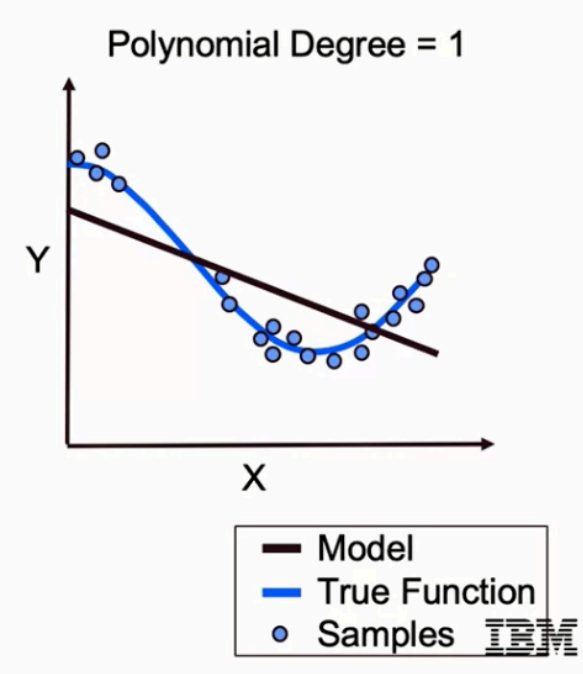
- ML là quá trình máy tính "học" từ dữ liệu để đưa ra dự đoán.
- Có 2 loại chính trong Supervised ML:
  - Regression: dự đoán biến đầu ra dạng số (vd: giá nhà).
  - Classification: dự đoán biến đầu ra dạng phân loại (vd: spam hay không).
- Mục tiêu ML:
  - Interpretation: tìm hiểu mối quan hệ giữa dữ liệu và kết quả.
  - Prediction: dự đoán giá trị chính xác nhất.
- Linear Regression: mô hình quan hệ tuyến tính giữa biến đầu ra (y) và các biến đầu vào (x).
  - Công thức cơ bản:

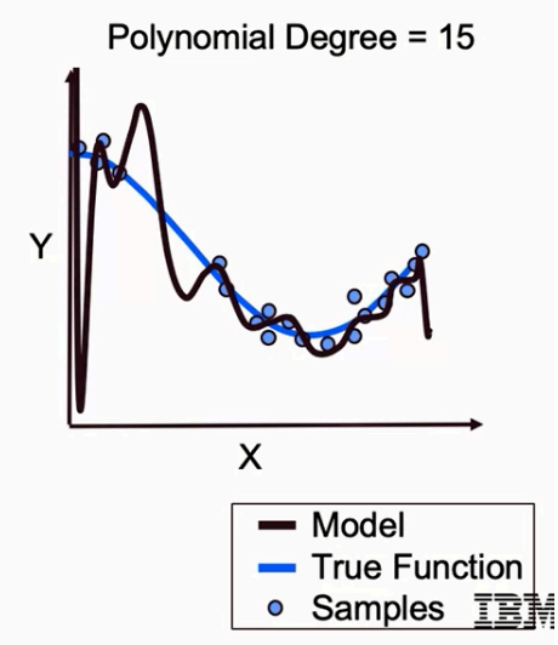
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \epsilon$$

  - Residuals = giá trị thực – giá trị dự đoán.
  - Đánh giá mô hình qua các chỉ số: SSE, TSS, R^2.

2. Data Splits and Polynomial Regression

- Train-Test Split: chia dữ liệu thành tập train để huấn luyện và tập test để kiểm tra khả năng tổng quát.
- Lỗi trên tập train thường nhỏ hơn tập test → giúp phát hiện overfitting.
- Polynomial Regression: mở rộng linear regression bằng cách thêm các đặc trưng bậc cao (x^2, x^3, ...) để mô tả quan hệ phi tuyến.

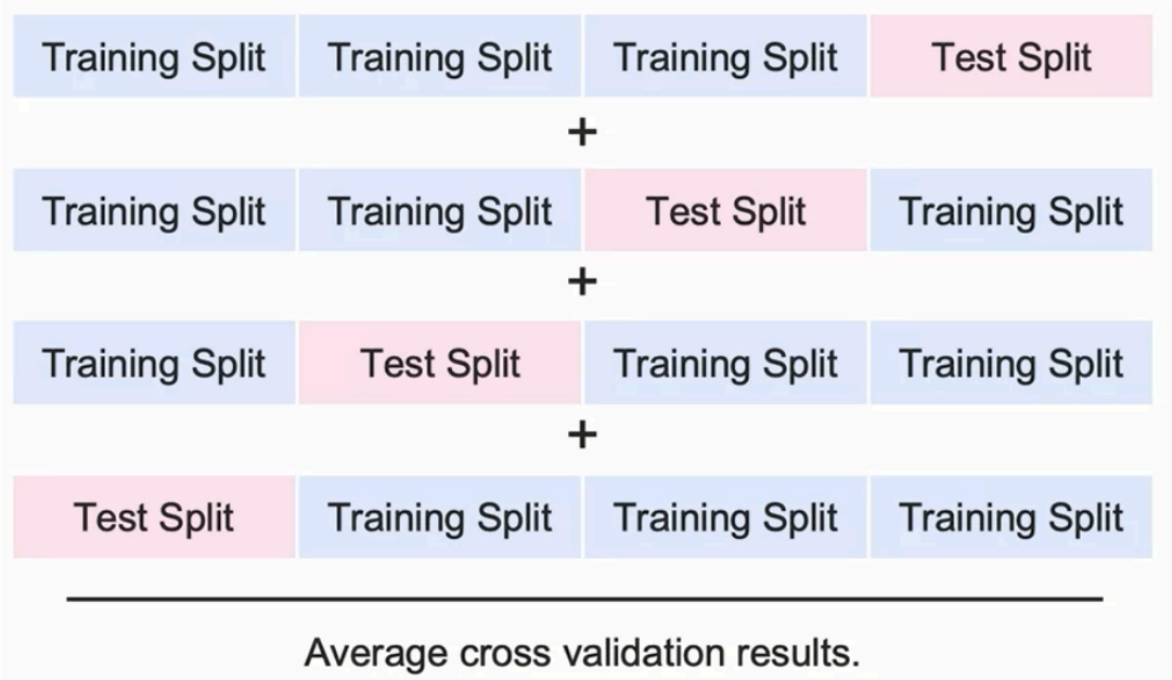




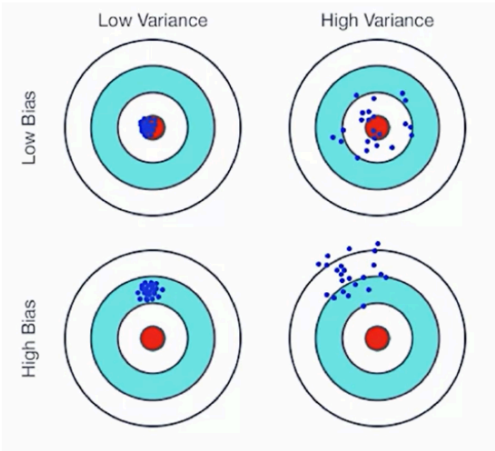
- Ưu điểm: mô hình có thể nắm bắt mối quan hệ phức tạp hơn.
- Nhược điểm: dễ dẫn tới overfitting nếu bậc quá cao.
- Ngoài polynomial, còn nhiều thuật toán khác: Logistic Regression, KNN, Decision Tree, SVM, Random Forest, Deep Learning.

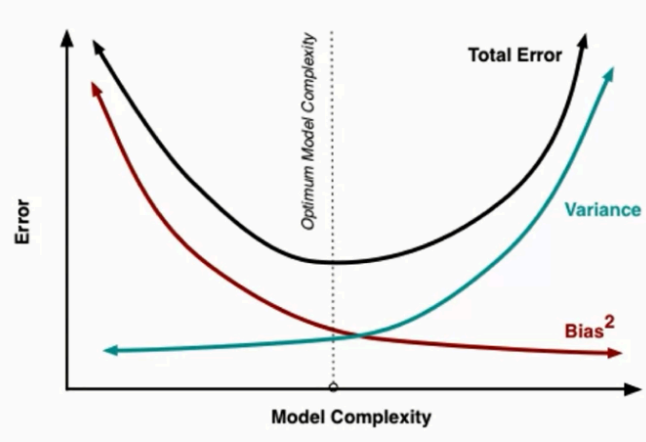
3. Cross Validation

- Vấn đề: một tập test đơn lẻ có thể không đủ tin cậy.
- Cross Validation (CV) giúp đánh giá mô hình chính xác hơn bằng cách chia dữ liệu thành nhiều phần và luân phiên train/test.
- Ba phương pháp CV phổ biến:
  - k-Fold CV: chia dữ liệu thành k phần, lần lượt dùng một phần làm test.
  - Leave-One-Out (LOO): mỗi lần bỏ ra 1 mẫu làm test.
  - Stratified CV: chia dữ liệu nhưng vẫn giữ tỷ lệ nhãn đồng đều.
- CV giúp cân bằng giữa độ phức tạp mô hình – sai số – lượng dữ liệu hạn chế.



4. Bias Variance Trade off and Regularization Techniques: Ridge, LASSO, and Elastic Net





- Bias: sai lệch do mô hình quá đơn giản → underfitting.
- Variance: sai lệch do mô hình quá phức tạp → overfitting.
- Bias-Variance Tradeoff: phải tìm mức độ phức tạp “vừa đủ”.

- Regularization: thêm tham số phạt λ vào hàm mất mát để kiểm soát độ phức tạp.
    - Ridge Regression (L2): phạt bình phương hệ số → hệ số bị thu nhỏ nhưng không về 0.
    - LASSO Regression (L1): phạt giá trị tuyệt đối → có thể triệt tiêu hẳn hệ số (feature selection).
    - Elastic Net: kết hợp L1 và L2, cân bằng giữa Ridge và LASSO.
- Ứng dụng: giảm overfitting, chọn đặc trưng quan trọng, cải thiện khả năng tổng quát.

5. Regularization Details

- Có nhiều cách hiểu về Regularization:
    - Analytic view: λ làm hẹp khoảng giá trị hệ số → mô hình đơn giản hơn.
    - Geometric view: nghiệm tối ưu bị ràng buộc trong vùng phạt (hình ellipse – Ridge, hình diamond – LASSO).
    - Probabilistic view: Ridge ~ Gaussian prior, LASSO ~ Laplace prior.
- Tóm lại:
    - Regularization giúp tránh overfitting bằng cách giới hạn độ lớn hệ số.
    - Ridge: thu nhỏ hệ số nhưng vẫn giữ tất cả.

$$J(\beta_0, \beta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(\left(\beta_0 + \beta_1 x_{obs}^{(i)}\right) - y_{obs}^{(i)}\right)^2 + \boldsymbol{\lambda}\sum_{j=1}^{k}|\boldsymbol{\beta_j}|$$

    - LASSO: nhiều hệ số = 0 → chọn lọc đặc trưng.

$$J(\beta_0, \beta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(\left(\beta_0 + \beta_1 x_{obs}^{(i)}\right) - y_{obs}^{(i)}\right)^2 + \boldsymbol{\lambda}\sum_{j=1}^{k}|\boldsymbol{\beta_j}|$$

- Elastic Net: cân bằng giữa hai cách.

$$J(\beta_0, \beta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(\left(\beta_0 + \beta_1 x_{obs}^{(i)}\right) - y_{obs}^{(i)}\right)^2 + \lambda\sum_{j=1}^{k}|\beta_j| + \lambda_2\sum_{j=1}^{k}\beta_j^2$$

- Đây là kỹ thuật quan trọng trong hầu hết mô hình ML hiện đại.

6. Note trong lớp

Term

- ablation study
- sensitive analysis
- feature selection

Comparing vs Benchmarking

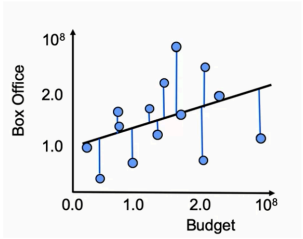
- comparing: so sánh giữa các model với nhau, không dựa vào baseline (quy chuẩn chung)
- benchmarking: so sánh các model với model SOTA, có baseline

Dựa vào các term này để biện luận, không tự nhiên chủ nghĩa

- framework
- methodology
- best practices
- recommendation

Residual vs Error

Calculating the Residuals



$$y_\beta\left(x_{obs}^{(i)}\right) - y_{obs}^{(i)}$$

Minimizing the Error Function

$$J(\beta_0, \beta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(\left(\beta_0 + \beta_1 x_{obs}^{(i)}\right) - y_{obs}^{(i)}\right)^2$$

- Residual: tính khoảng cách giữa ground truth với y_pred → có âm
- Error: tính độ lệch giữa ground truth với y_pred → luôn dương

Tại sao không dùng khoảng cách Euclide mà dùng residual đến đường hồi quy?

- Công thức đơn giản hơn
- Sai số (errors) hay “disturbance term” εi
  - Giả định: các εi được phân phối chuẩn với trung bình bằng 0: εi~N(0,σ2)
  - Tức là phần sai số (phần không giải thích bởi mô hình) được xem là ngẫu nhiên, có phân phối đối xứng, không bị lệch, và độ phân tán (variance) của nó cố định (đều) qua các giá trị của biến độc lập.