Model Selection Using Cross-Validation

Comparing Linear Models with Increasing Complexity

Eldad Haber

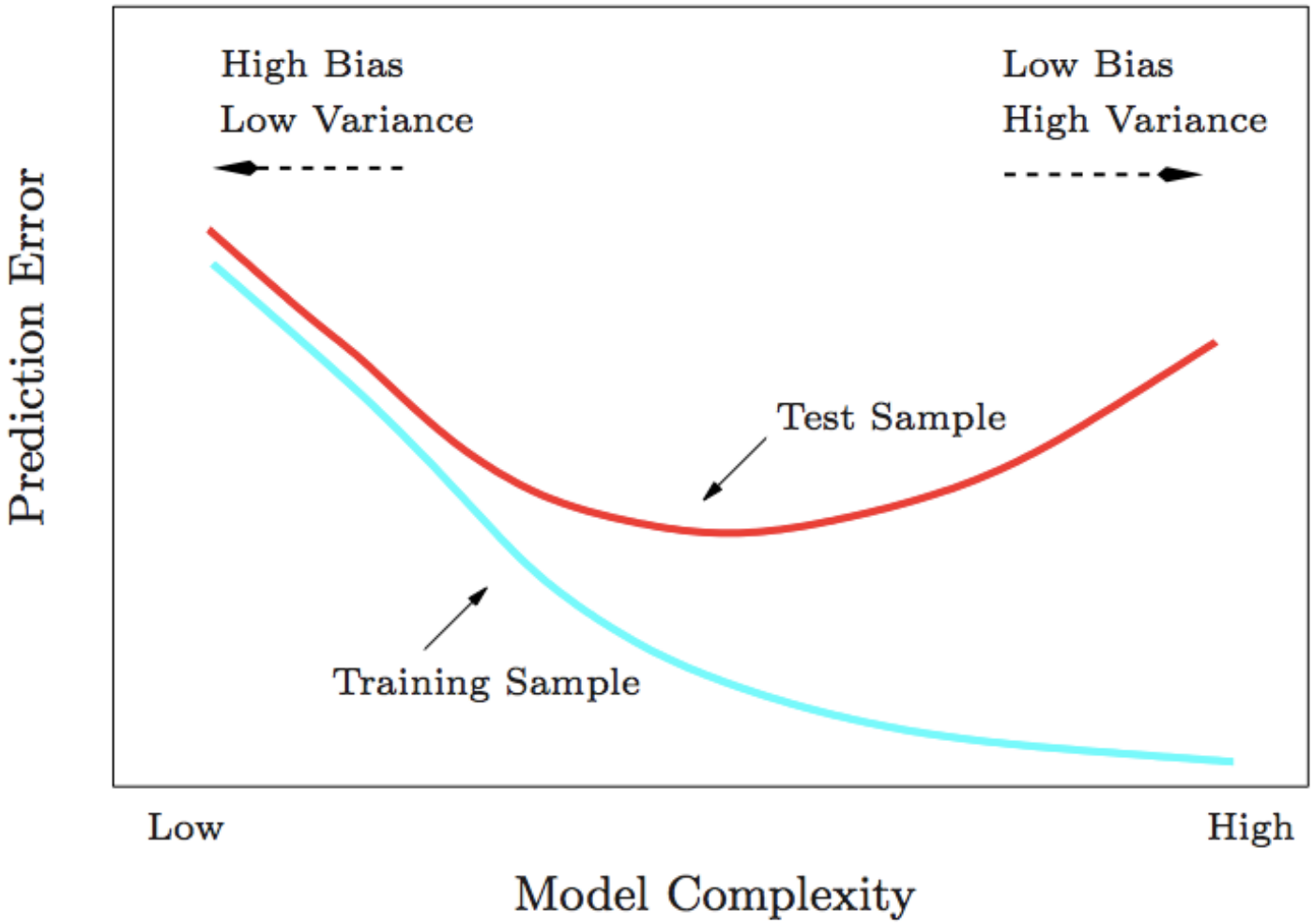
UBC

February 15, 2025

- **Objective:** Explain how to use cross-validation to select the best model among four linear models with increasing complexity.
- **Context:**
  - Trained four linear models with more basis functions (e.g., polynomial features).
  - Need to decide which model generalizes best to unseen data.

What is Model Complexity?

- **Definition:** The flexibility of a model to fit the training data.
- **Example:** Linear model with 1 basis function (simple) vs. 10 basis functions (complex).



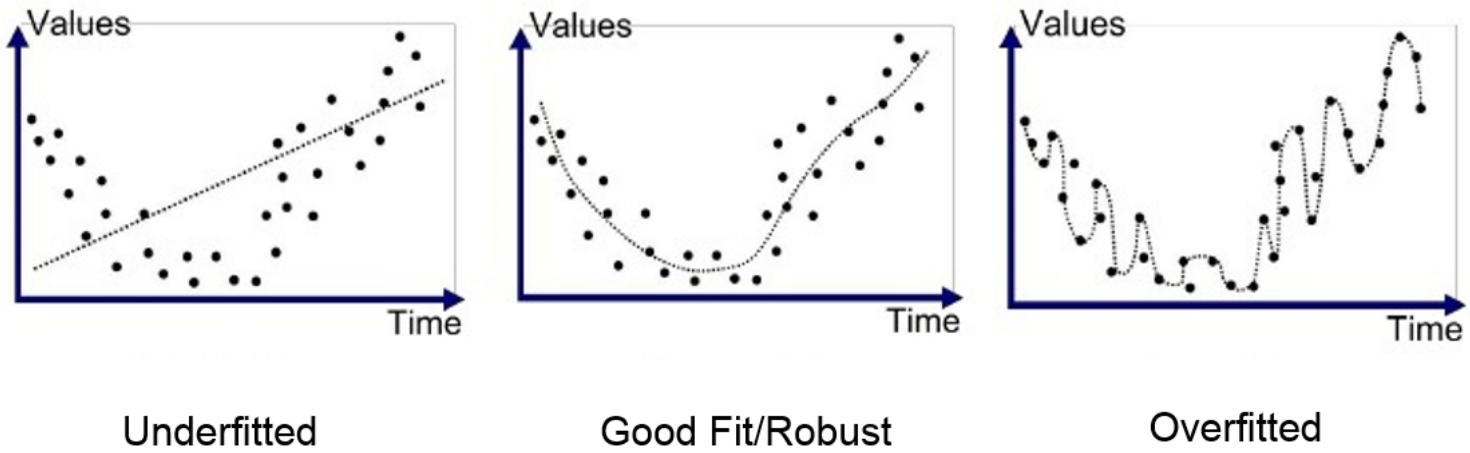
- **Visual:**

The Four Models

- **Model 1:** Linear model with 1 basis function (e.g., $y = w_0 + w_1 x$).
- **Model 2:** Linear model with 2 basis functions (e.g., $y = w_0 + w_1 x + w_2 x^2$).
- **Model 3:** Linear model with 3 basis functions (e.g., $y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$).
- **Model 4:** Linear model with 4 basis functions (e.g., $y = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4$).

The Challenge: Overfitting

- Increasing model complexity can lead to overfitting.
- **Overfitting:** Complex models fit training data well but may fail on unseen data.



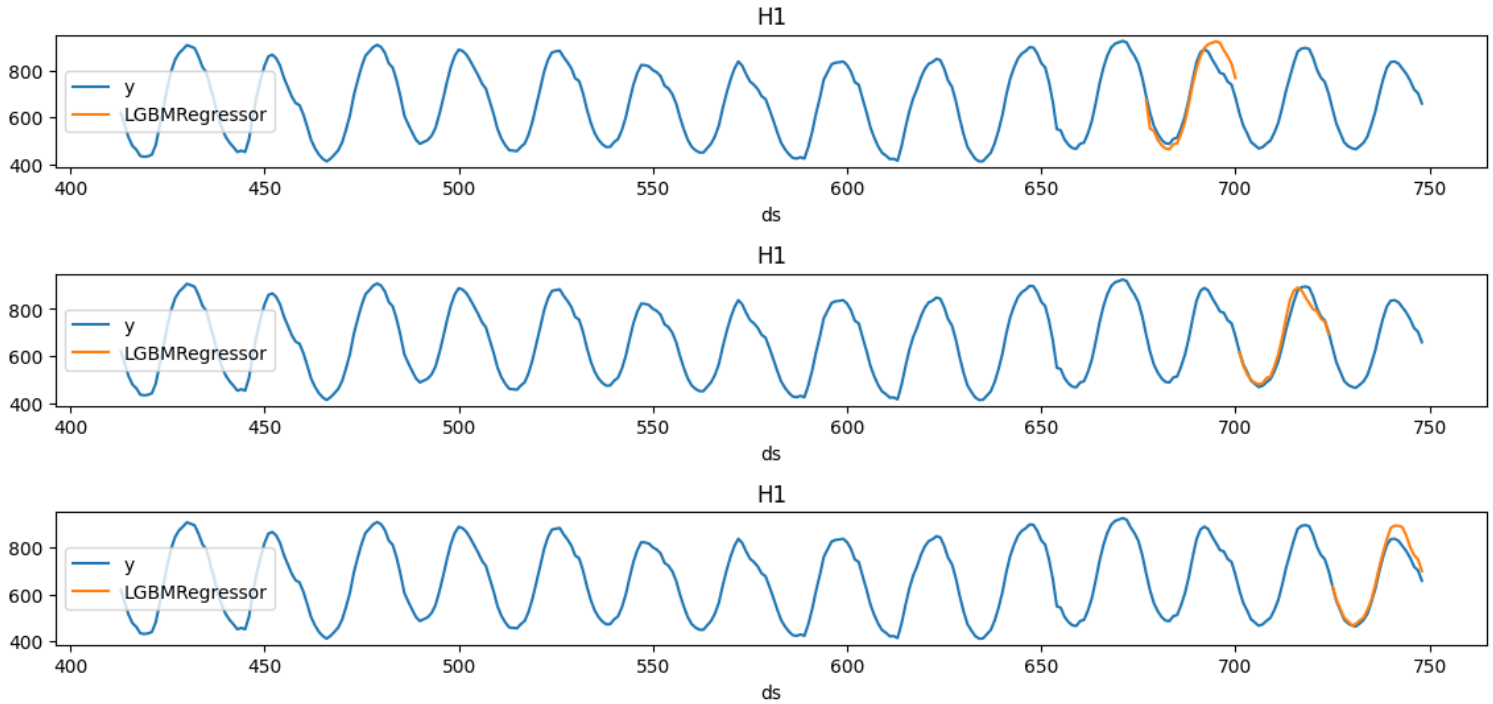
- **Visual:**

What is Cross-Validation?

- **Definition:** A technique to evaluate how well a model generalizes to unseen data.
- **Purpose:** To compare models and select the one that performs best on unseen data.
- **Analogy:** "Testing your knowledge on different questions, not just the ones you've memorized."

K-Fold Cross-Validation

- **Steps:**
  1. Split data into K folds.
  2. Train on K-1 folds, validate on the remaining fold.
  3. Repeat K times and average the results.
- **Visual:**



Applying CV to Compare Models

- **Steps:**
  1. Train each model on the training set.
  2. Use K-Fold CV to evaluate each model's performance.
  3. Compare the average validation scores.
- **Metric:** Use a performance metric like **Mean Squared Error (MSE)** or $\mathbf{R^2}$.

Example: Model 1 (1 Basis Function)

- **Description:** Simple linear model.
- **CV Results:** Average validation score (e.g., MSE =).

Example: Model 2 (2 Basis Functions)

- **Description:** Quadratic model.
- **CV Results:** Average validation score (e.g., MSE =).

Example: Model 3 (3 Basis Functions)

- **Description:** Cubic model.
- **CV Results:** Average validation score (e.g., MSE =).

Example: Model 4 (4 Basis Functions)

- **Description:** Quartic model.
- **CV Results:** Average validation score (e.g., MSE =).

Comparing the Models

Model	Basis Functions	Avg. Validation MSE
Model 1	1	
Model 2	2	
Model 3	3	
Model 4	4	

- **Analysis:**
  - Model ??? has the lowest validation error.
  - Model ??? is slightly worse,

Selecting the Best Model

- **Decision:** Choose **Model ???** because it has the best generalization performance.
- **Why not Model ???** Higher complexity without significant improvement in validation error.

Summary and Questions

- **Recap:**
  - Cross-validation helps compare models and select the one that generalizes best.
  - Increasing complexity improves performance up to a point, after which overfitting occurs.
  - Model 3 (3 basis functions) is the best choice for this dataset.
- **Call to Action:** "Use cross-validation in your projects to make informed model selection decisions!"
- **Questions?**