

Recap: Generalization — Why Training Error Lies

DS701 Session 8 — Wed Sep 30, 2026

Today's plan

Knowledge-Check Review

KC 1: The contract

State the “supervised contract” — the assumption the lecture says makes it possible for a rule learned on training data to say anything about data it has never seen — and give one concrete situation in which the contract is broken.

KC 2: Why training error only goes down

*In the polynomial curve-fitting example, the training error reaches zero at $k = 9$ while the test error is smallest at $k = 3$. Explain why training error can only go down (never up) as k increases, and why that makes it useless for choosing k . Then say which of **bias** and **variance** the $k = 9$ model suffers from, and which the $k = 1$ model — in one sentence each.*

KC 3: Choosing a depth with 5-fold CV

*You have 200 labelled rows and must choose the maximum depth of a decision tree (a hyperparameter). Describe how you would use 5-fold cross-validation to choose it, and say what role — if any — the **test** set plays in that choice. Why is 5-fold CV preferable to a single random train/validation split when the data is this small?*

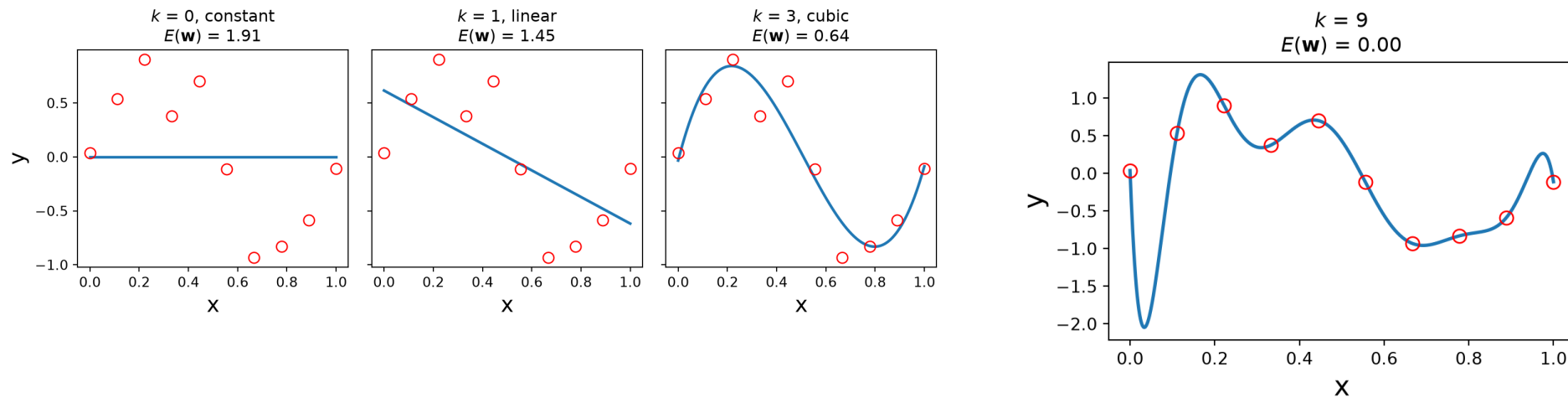
Highlights

Everything supervised rests on one contract

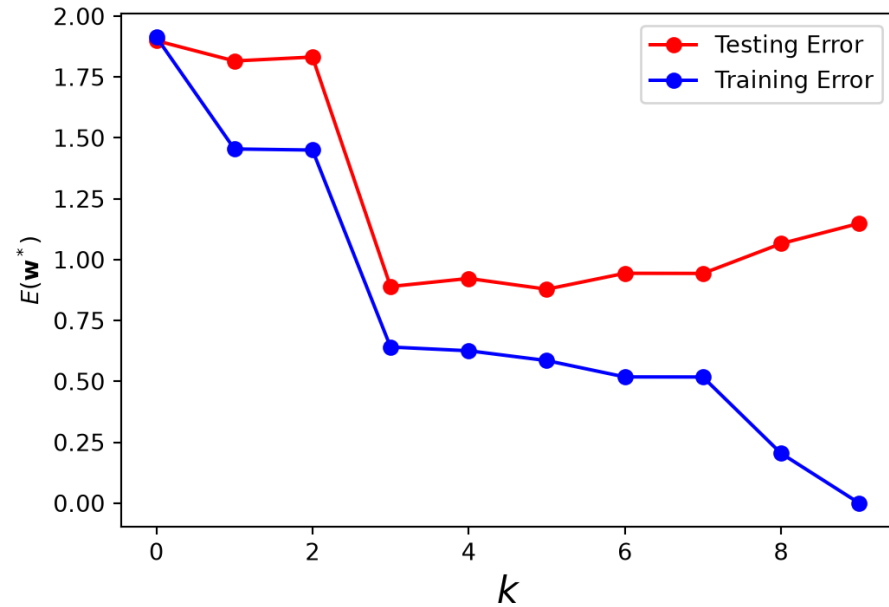
Given training pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, learn a rule that predicts y_j for an $\mathbf{x}_j$ **not in the training data**.

Two assumptions make that tractable:

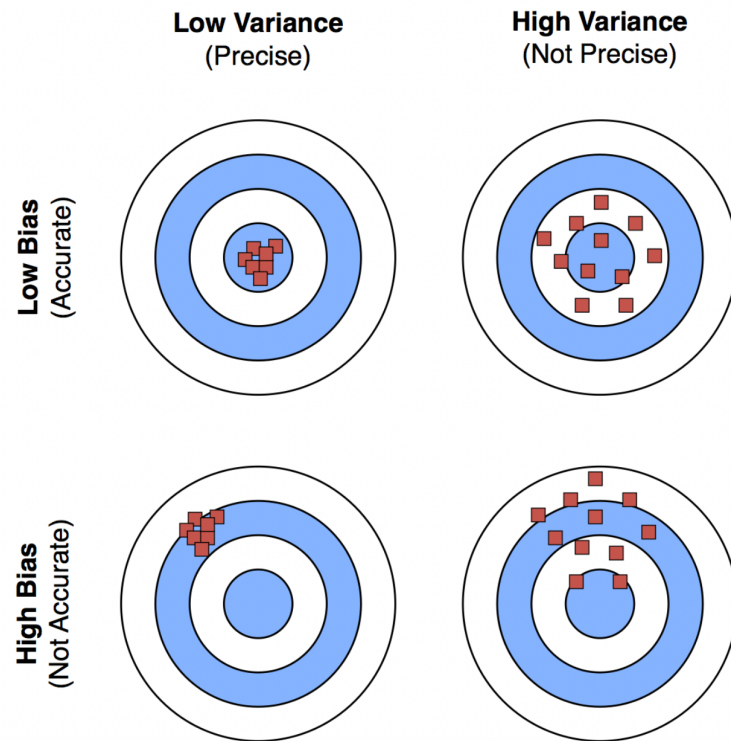
Training error lies



The money plot: train vs test



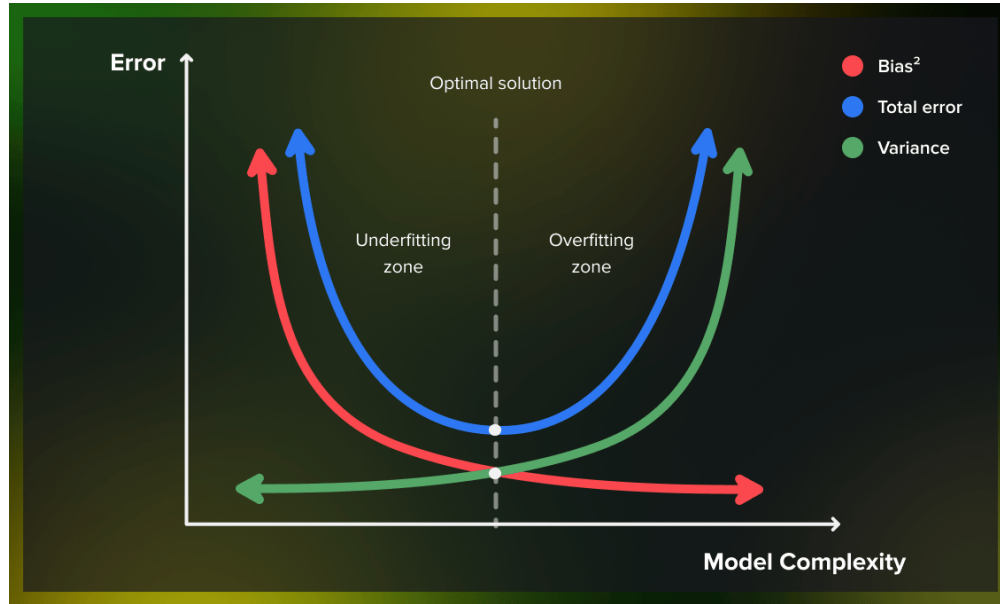
Two ways to be wrong



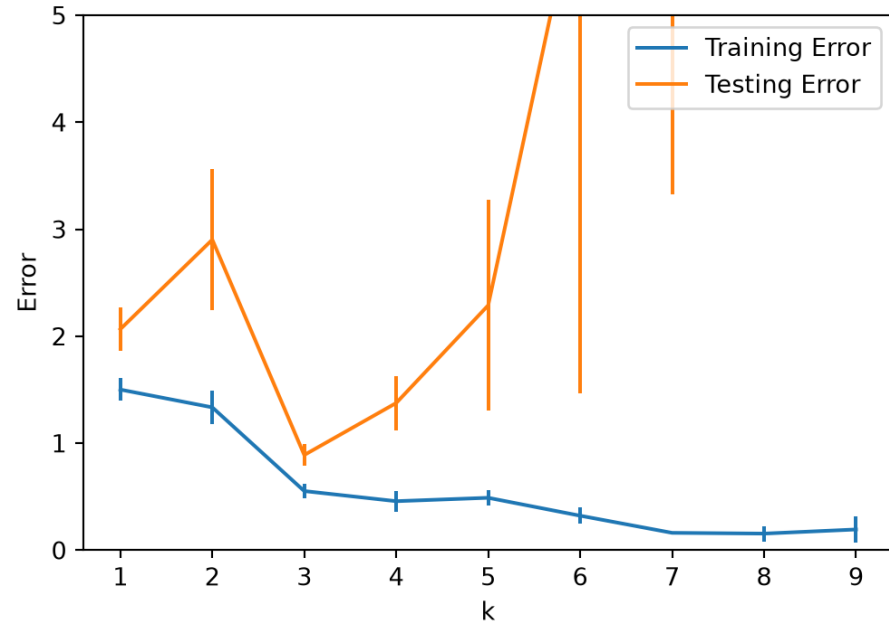
Bias — the model is too simple to capture the pattern; it **underfits** ($k = 0, 1$). Refit it on new data and it is wrong the *same* way every time.

Variance — the model is so flexible it fits the noise; it **overfits** ($k = 9$). Refit it on a different sample from the same distribution and you get a *very different* answer.

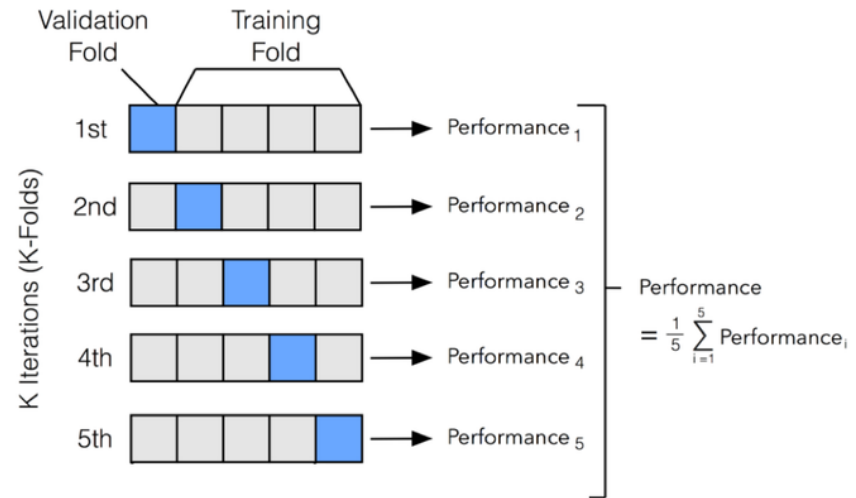
The trade-off *is* the U-shape



Choosing k honestly: hold out, then grid search



K -fold cross-validation



What the lecture did not draw: learning curves

The plots so far vary the **model** at fixed data. Today's stretch holds the model fixed and varies the **amount of data**:

High bias (a line on a cubic)

- train and validation error converge *fast*
- ... to a level well **above** the noise floor
- more data changes nothing

High variance (degree 12 on 10 points)

- training error ≈ 0 , validation error large
- a **gap** that narrows only as n grows
- more data helps — and so would a simpler model

In-Class Activity

Activity: train/test discipline, then learning curves

Goal: on your section's synthetic regression, reproduce the train-vs-test curve for degrees 0–12 and pick the degree; pick it again with 5-fold CV without touching the test set; measure bias and variance for $k = 1$ vs $k = 9$ by refitting on many samples; then draw learning curves for an under-fit and an over-fit model and diagnose a mystery model from its curve.

- Work in groups of 2–3. Open **your section's** notebook.
- Parts marked (**autograded**) are submitted to Gradescope; the rest is participation.
- Staff will circulate — be ready to explain any part of your work.
- Dependencies: [numpy](#), [matplotlib](#), [scikit-learn](#).

Section A1

 [Open in Colab](#)  [Open on GitHub](#)

Section B1

 [Open in Colab](#)  [Open on GitHub](#)

Going deeper

Full lecture notes: [Generalization: Why Training Error Lies](#)

- [The Supervised Contract](#) — the i.i.d. assumption and when it fails
- [A Toy Example: Polynomial Curve Fitting](#) — the $k = 0, 1, 3, 9$ fits
- [Generalization Error](#) — the train-vs-test curve; [Overfitting](#) — why, and three fixes
- [Bias and Variance](#) and the [Bias-Variance Trade-Off](#)
- [Parameters and Hyperparameters](#); [Holding Out Data](#); [Grid Search](#)
- [K-Fold Cross-Validation](#) and [Hold Out Strategies](#)
- [How This Plays Out This Term](#) — the same trade-off, a different knob each week

Before you go: quiz 1 is one week out

Wed Oct 7, in class — closed notes, paper and pencil, covering lectures 01–07 (foundations + clustering + today).

- The **reference sheet** for the unit is published **today**: [quiz1-reference.pdf](#). Study with it — the printed copy handed out on quiz day is identical, and it is the only sheet allowed in the room.
- Today's ideas that live on the sheet: the contract, overfitting, bias vs variance, hold-out / K -fold CV.
- The homework sets and knowledge checks are the practice; the recap decks are the answer keys.