Supervised Learning (cont.)

CAI 5107: Machine Learning
Instructor: Anowarul Kabir
Email: akabir@usf.edu
Fall 2025

Learning Multiple Classes

- In the general form, we have K classes denoted as C_i where $i \in \{1, \ldots K\}$
- The dataset has the following form where a sample belongs to one and only one class

$$\mathcal{X} = \{\boldsymbol{x}^t, \boldsymbol{r}^t\}_{t=1}^N$$

where $\boldsymbol{r}$ has K dimensions and

$$r_i^t = \begin{cases} 1 & \text{if } \boldsymbol{x}^t \in C_i \\ 0 & \text{if } \boldsymbol{x}^t \in C_j, j \neq i \end{cases}$$

Learning Multiple Classes

- *K*-class classification can be viewed as *K* two-class problems: *K* decision boundaries
- So, we have K hypothesis to learn such that

$$h_i(\boldsymbol{x}^t) = \begin{cases} 1 & \text{if } \boldsymbol{x}^t \in C_i \\ 0 & \text{if } \boldsymbol{x}^t \in C_j, j \neq i \end{cases}$$

- Empirical error: Total number of misclassification

$$E(\{h_i\}_{i=1}^K | \mathcal{X}) = \sum_{t=1}^{N} \sum_{i=1}^{K} 1(h_i(\boldsymbol{x}^t) \neq r_i^t)$$

- *Multi-label classification: when a sample may belong to multiple classes

Regression

- In the form of data labels associated with real values where $r^t \in \mathbb{R}$

- Two forms of regression: interpolation vs. extrapolation

$$\boldsymbol{r}^t = \boldsymbol{f}(\boldsymbol{x}^t) \qquad \boldsymbol{r}^t = \boldsymbol{f}(\boldsymbol{x}^t) + \boldsymbol{\epsilon}$$

random noise

- Empirical error: Mean-squared error (MSE)

$$E(g|\mathcal{X}) = \frac{1}{N}\sum_{t=1}^{N}[\boldsymbol{r}^t - \boldsymbol{g}(\boldsymbol{x}^t)]^2$$

Model complexities: Linear, 2nd and 6th order polynomials

If the model is linear in the form of

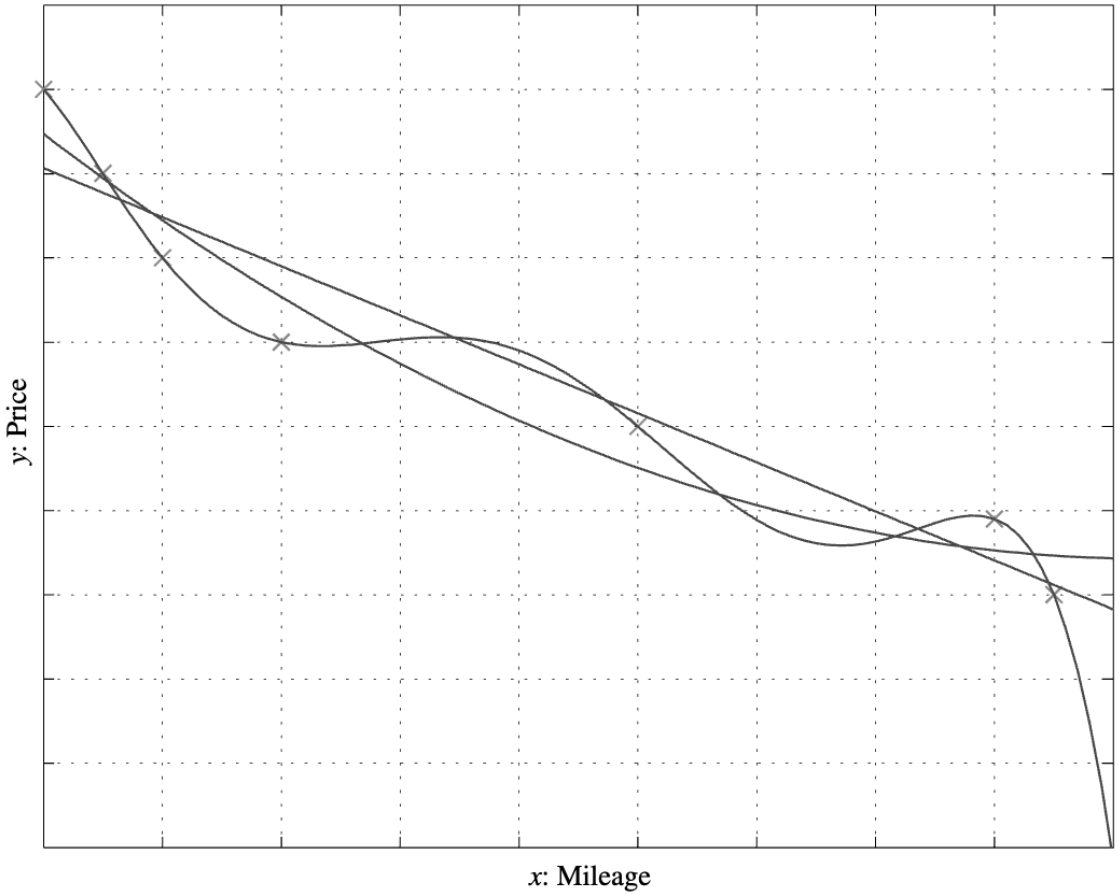
$$g(\boldsymbol{x}) = w_1 x_1 + \cdots + w_d x_d + w_0 = \sum_{j=1}^{d} w_j x_j + w_0$$

For 1D input linear model

$$g(x) = w_1 x + w_0$$

For 1D input quadratic model

$$g(x) = w_2 x^2 + w_1 x + w_0$$

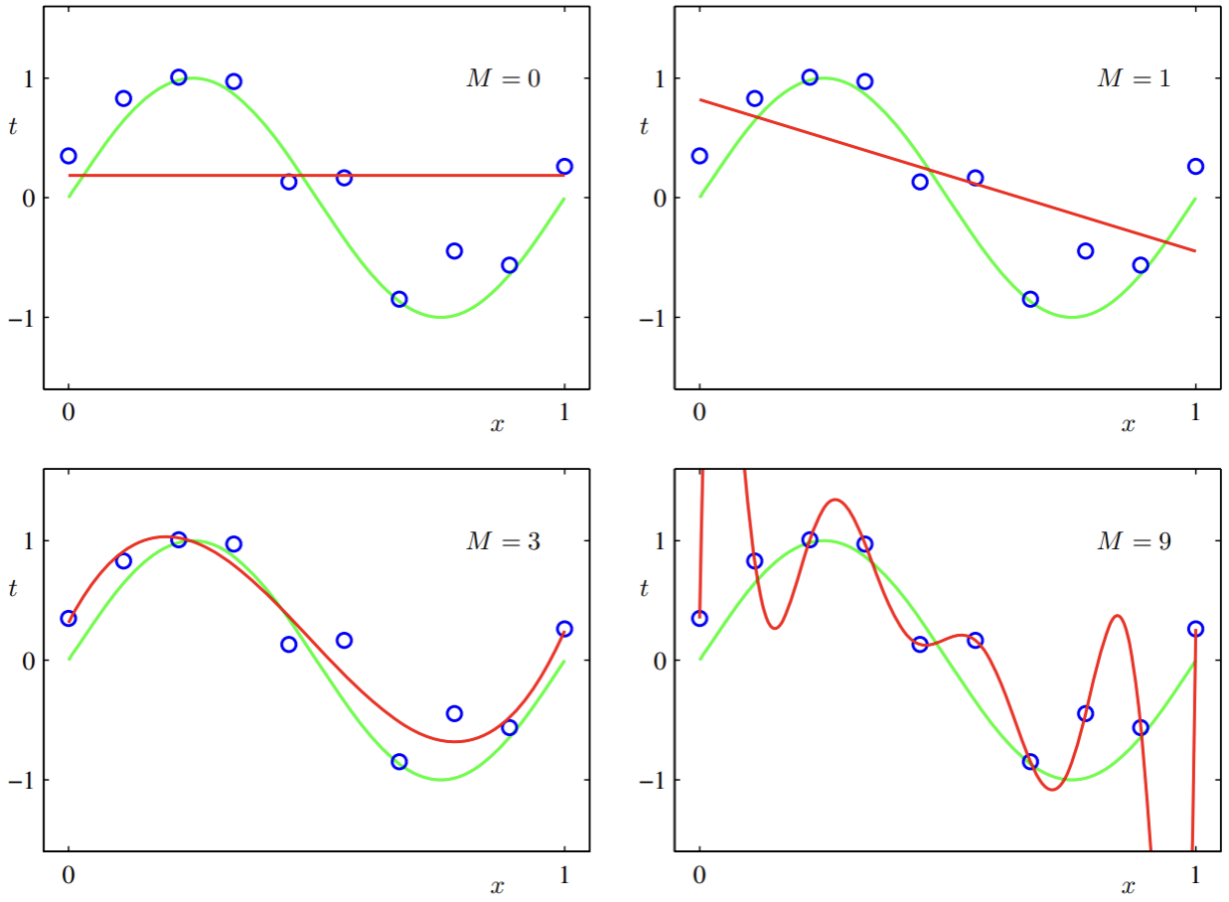


Occam's razor: Simple may be better

Overfitting and underfitting with model complexities

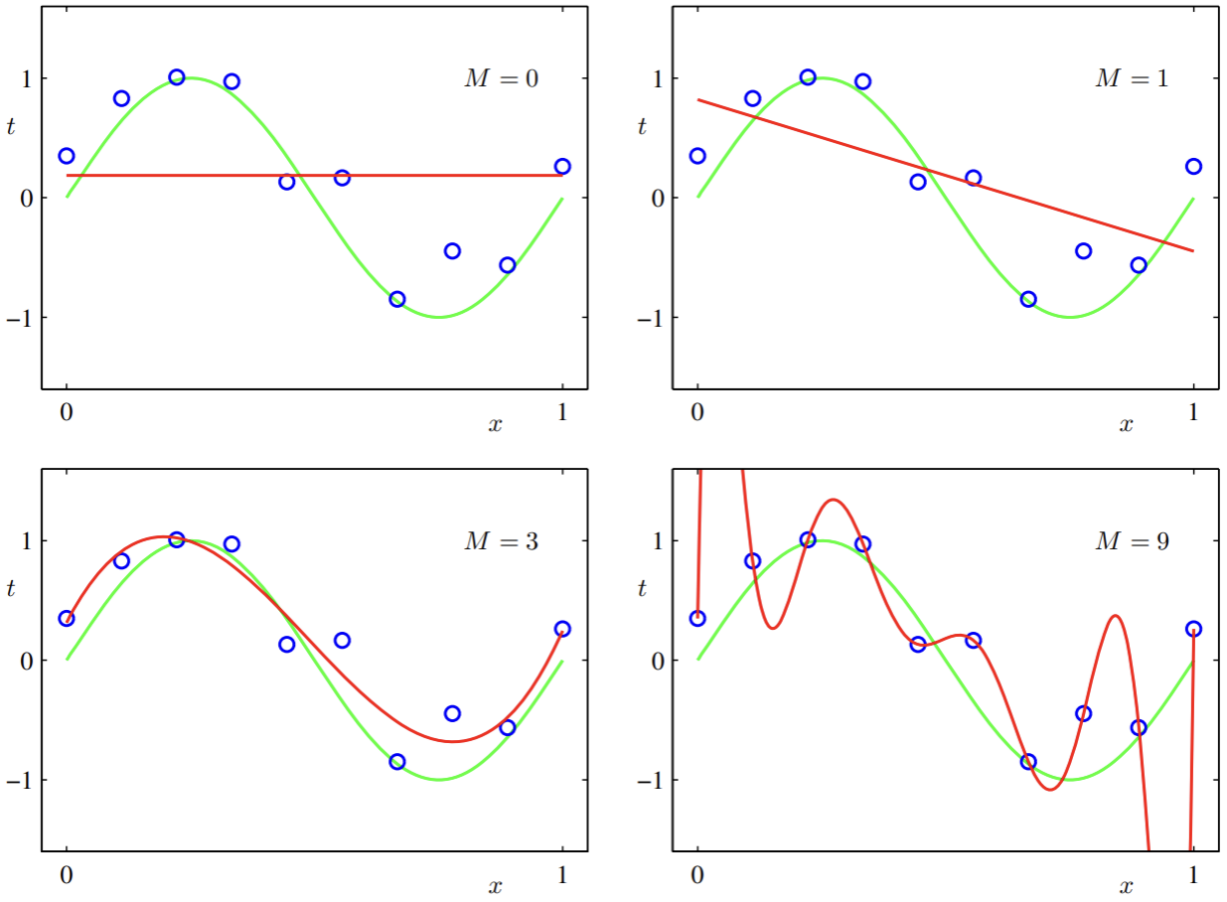
Generic polynomial model

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$



Coefficient values at different degree polynomials

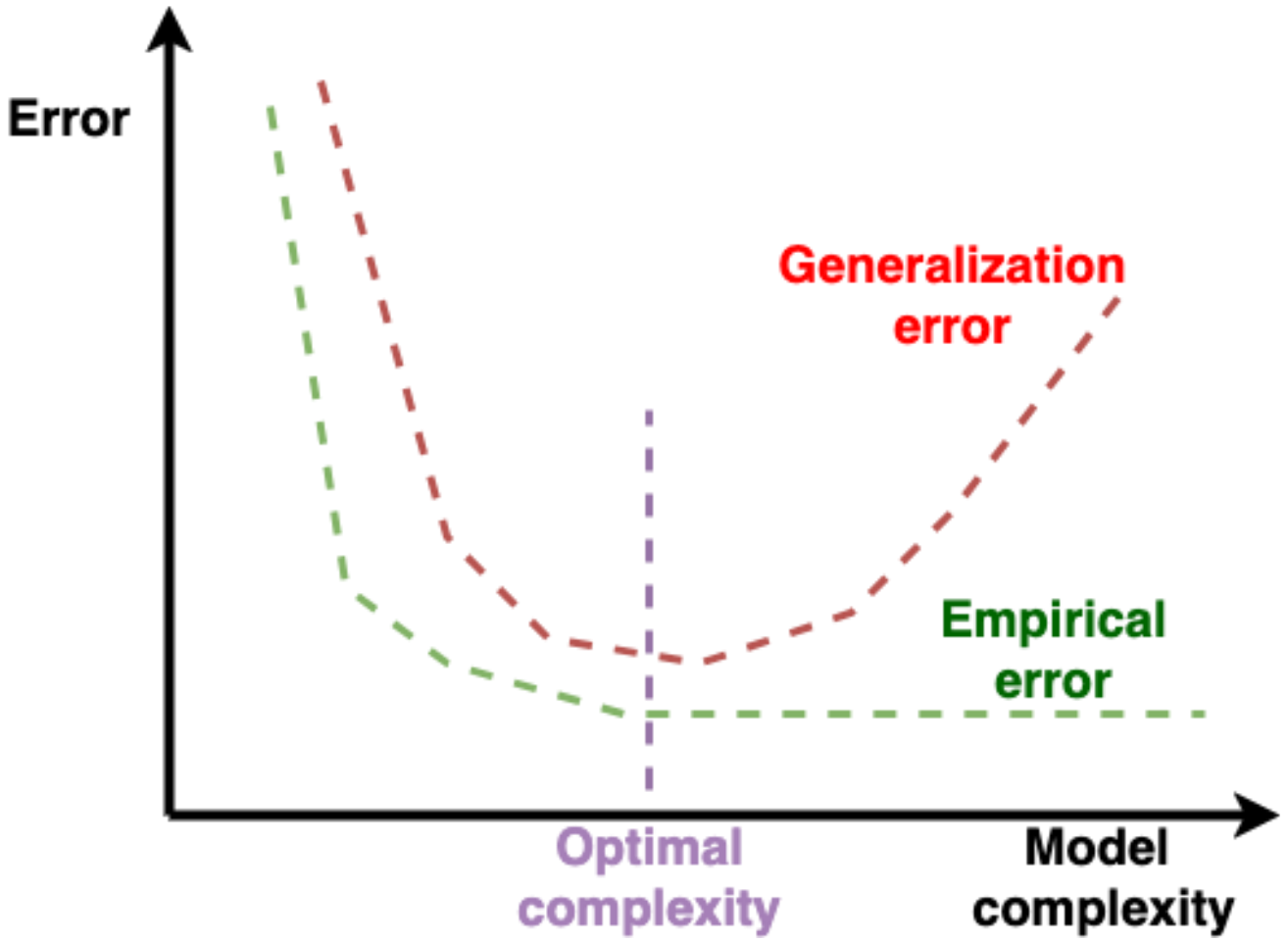
	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43



The magnitude of the coefficients increases dramatically as the order of the polynomial increases.

Model Selection & Generalization

- There is a trade-off between three factors:
  - The complexity or capacity of the hypothesis (model) to fit the data
  - The amount of training data
  - The generalization error on the unseen data



Vapnik-Chervonenkis (VC) Dimension

- VC dimension of a hypothesis class *H* is the largest number of points that can be shattered by *H*

- *H* shatters a set of points if for every possible labeling of the points there exists a hypothesis *h* that correctly classifies the points

- To analyze the complexity of the hypothesis class or model

VC dimension (2 examples)

- The maximum number of data points that can be shattered by a straight line in 2D is what? = 3

- The maximum number of data points that can be shattered by an axis aligned rectangle in 2D is what? = 4

- Discussed in the board with many examples

Let's find out the analytical solution of a linear model

- For 1D input linear model $g(x) = w_1 x + w_0$
- Consider MSE as empirical error:

$$E(g|\mathcal{X}) = \frac{1}{N}\sum_{t=1}^{N}[r^t - g(\mathbf{x}^t)]^2$$

- Take derivatives w.r.t w0 and w1
- And solve for each parameters
- Discussed on the board

- Homework: Find the analytical solution of w0, w1 and w2

$$g(x) = w_2 x^2 + w_1 x + w_0$$